С.Н. Мартышенко
Л.С. Мазелис
К.С. Солодухин

АВТОМАТИЗАЦИЯ ПРОЦЕССОВ АНАЛИЗА ДАННЫХ В ИССЛЕДОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

Монография

Министерство науки и высшего образования
Российской Федерации

Владивостокский государственный университет
экономики и сервиса (ВГУЭС)

С.Н. Мартышенко, Л.С. Мазелис, К.С. Солодухин

АВТОМАТИЗАЦИЯ АНАЛИЗА ДАННЫХ В ИССЛЕДОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

Монография

Владивосток
Издательство ВГУЭС
2019

УДК 168.2:316.4
ББК 63.3-2в61
М29

Рецензенты:
Ю.Д. Шмидт, д-р экон. наук, профессор, заведующий кафедрой бизнес-информатики и экономико-математических методов ДВФУ;
П.Г. Рагулин, канд. техн. наук, профессор кафедры компьютерных систем ДВФУ

Мартышенко, Сергей Николаевич

М29 **Автоматизация анализа данных в исследовании социально-экономических процессов** / С.Н. Мартышенко, Л.С. Мазелис, К.С. Солодухин; Владивостокский государственный университет экономики и сервиса. – Владивосток: Изд-во ВГУЭС, 2019. – 164 с.

ISBN 978-5-9736-0583-4

Представлены современные методы и инструментальные средства анализа данных, связанных с исследованием и моделированием социально-экономических процессов. Особое внимание уделено методам анализа, недостаточно представленным в классической научной литературе: обработка качественной информации, восстановление пропусков в первичных данных, анализ и корректировка ошибок. Приведены средства автоматизации разработки когнитивных моделей. Все предлагаемые методы анализа реализованы в виде компьютерных программ, которые доступны для использования в EXCEL. Представленные методы и технологии демонстрируются на примерах, рассчитанных с использованием разработанного программного обеспечения.

Для руководителей, бизнес-аналитиков, преподавателей, аспирантов и студентов экономических и информационных направлений, занимающихся исследованиями в области экономико-математического моделирования социально-экономических процессов.

УДК 168.2:316.4
ББК 63.3-2в61

ISBN 978-5-9736-0583-4

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В последние годы область научного знания, называемая «Анализ данных», получила бурное развитие. В первую очередь это связано с лавинообразным нарастанием объемов информации во всех областях знаний, которое получило специальное название «информационный взрыв». Одновременно с этим возрастают технические возможности сбора, хранения и передачи информации. Уровень компьютерной грамотности пользователей, работающих в различных областях знаний, позволяет перейти на новый уровень обработки информации, основанный на использовании наукоемких технологий. Для рядовых пользователей компьютера использование инструментальных средств EXCEL для обработки данных стало обыденным, и они готовы к использованию более сложного математического аппарата.

Анализ данных – это не просто обработка информации после ее получения и сбора, но средство проверки гипотез. Цель любого анализа данных – понимание исследуемой ситуации целиком (выявление тенденций, в том числе отклонений от плана), прогнозирование развития процессов и разработка рекомендаций для лиц, принимающих решения.

Для достижения этой цели ставятся следующие задачи:

– сбор информации;

– структуризация информации;

– выявление закономерностей, анализ;

– прогнозирование и разработка рекомендаций.

Следует отметить, что перечисленные задачи существовали и до оформления анализа данных в качестве отдельного научного направления. Однако, если раньше рассмотрение этих задач было уделом узких специалистов, то в современных условиях с необходимостью решения таких задач сталкивается более широкий круг исследователей.

Анализ данных – это прикладная наука, в которой нет однозначно предопределённых алгоритмов решения задач. Многие задачи

«индивидуальны», и постоянно появляются все новые классы задач, под которые необходимо разрабатывать математический аппарат. Решение новых задач требует создания новых технологий, техник, приемов, способов и т.п.

При поиске эффективных решений анализа данных важную роль играют эксперименты, в которых изменяются значения параметров, используются различные разбиения, пробуются варианты удаления или оставления шумовых объектов.

Специалисты в области анализа данных используют разного рода эвристические (экспертные) предположения о выборе информативных признаков, классе моделей, параметрах выбранной модели. Эти предположения основываются на опыте аналитика, его интуиции, понимании смысла анализируемого процесса. Иначе говоря, разрешено все, что дает положительный эффект. С другой стороны, анализ данных основан на использовании алгоритмов, заложенных в компьютерных программах, что, в свою очередь, является глубоко формализованным процессом.

При анализе данных необходимо придерживаться следующей стратегии:

– отталкиваться от опыта эксперта;

– рассматривать проблему под разными углами и комбинировать подходы;

– не стремиться сразу к высокой точности, двигаться к решению от более простых и грубых моделей к более сложным и точным.

Анализ данных – это всегда комплексное исследование, то есть предполагается решение множества задач, которое уточняется по ходу работы. Объем и размерность данных, как правило, очень высоки. Часто приходится решать множество повторяющихся (рутинных) задач. Качество результата, в котором заинтересованы исследователи, в значительной степени зависит от качества данных. Поэтому методы повышения качества данных в анализе данных имеют очень важное значение. И, наконец, решать все задачи в комплексе, в том числе, и экспериментировать, невозможно без глубокой автоматизации процессов. В связи с этим в монографии приводятся конкретные алгоритмы автоматизации решения задач анализа данных, ориентированные на практическое применение. Все материалы широко иллюстрируются числовыми примерами.

Глава 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ АНАЛИЗА ДАННЫХ

1.1. Анализ данных как мультидисциплинарная область знаний

Анализ данных – область математики и информатики, занимающаяся построением и исследованием наиболее общих математических методов и вычислительных алгоритмов извлечения знаний из экспериментальных данных; процесс исследования, фильтрации, преобразования и моделирования данных с целью извлечения полезной информации и принятия решений. Анализ данных имеет множество аспектов и подходов, охватывает разные методы в различных областях науки и деятельности.

Анализ данных – мультидисциплинарная область знаний, возникшая на базе таких наук, как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др. (рис. 1.1).

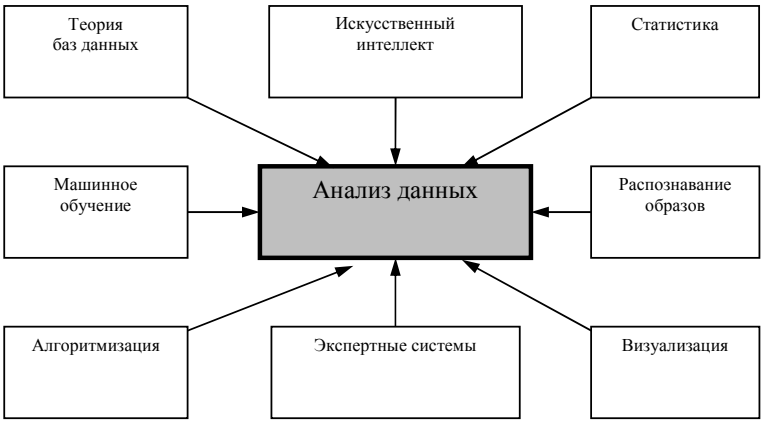


Рис. 1.1. Связь анализа данных с другими дисциплинами

В последние годы анализ данных чаще трактуется как интеллектуальный анализ данных (***data mining***).

Data Mining – процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.

Data Mining – процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе.

Data Mining – процесс, цель которого состоит в обнаружении новых значимых корреляций, образцов и тенденций в результате просеивания большого объема хранимых данных с использованием методик распознавания образцов, а также применение статистических и математических методов (определение Gartner Group).

Основная особенность Data Mining – это сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

С распространением анализа данных на решение различных прикладных задач особое значение приобретают компьютерные технологии решения задач. Компьютерные технологии позволяют автоматизировать работу исследователей из различных областей.

Перспективы технологии Data Mining: выделение типов предметных областей с соответствующими им эвристиками, формализация которых облегчит решение соответствующих задач Data Mining, относящихся к этим областям; создание формальных языков и логических средств, с помощью которых будет формализованы рассуждения и автоматизация которых станет инструментом решения задач Data Mining в конкретных предметных областях; создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать теоретические модели, опирающиеся на эмпирические данные; преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

Фундаментом автоматизации работы исследователей являются структуры и модели данных. В данной работе рассматриваются

методы анализа данных, реализованные программно в виде набора инструментальных средств в EXCEL. Выбор среды EXCEL был сделан в связи с тем, что EXCEL доступен самому широкому кругу пользователей занимающихся анализом данных в различных областях знаний. С другой стороны, EXCEL обладает широким набором собственных инструментальных средств анализа данных.

Структура данных выбиралась в соответствии с логикой исследования социально-экономических явлений путем анкетных опросов. Выбранная структура вполне подходит и для анализа маркетинговой информации.

Многие примеры, рассматриваемые в работе, рассчитаны с помощью собственных программных средств.

1.2. Типы, структуры и модели данных

1.2.1. Многомерные статистические данные. Наблюдения, объекты, признаки, шкалы

В широком понимании данные представляют собой факты, текст, графики, картинки, звуки, аналоговые или цифровые видео-сегменты. Данные могут быть получены в результате измерений, экспериментов, арифметических и логических операций. Данные должны быть представлены в форме, пригодной для хранения, передачи и обработки. Иными словами, данные – это необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на их основе.

Сбор статистических данных имеет своей целью получение новых знаний об объекте или явлении. Под сбором данных часто понимается измерение характеристик объектов или явлений.

Измерение – процесс определения отношения одной величины (измеряемой) к другой, принятой за постоянную единицу измерения.

Шкала измерения – это форма фиксации совокупности признаков изучаемого объекта с упорядочиванием их в определенную числовую систему.

Переменная (*variable*) – свойство или характеристика, общая для всех изучаемых объектов, проявление которой может изменяться от объекта к объекту.

Значение *(value)* переменной является проявлением признака.

При анализе данных, как правило, нет возможности рассмотреть всю совокупность объектов. Изучение очень больших объемов данных является дорогостоящим процессом, требующим больших временных затрат, а также неизбежно приводит к ошибкам, связанным с человеческим фактором.

Вполне достаточно рассмотреть некоторую часть всей совокупности, то есть выборку, и получить интересующую информацию на ее основании.

Однако размер выборки должен зависеть от разнообразия объектов, представленных в генеральной совокупности. В выборке должны быть представлены различные комбинации и элементы генеральной совокупности.

Генеральная совокупность (***population***) – вся совокупность изучаемых объектов, интересующая исследователя.

Выборка (***sample***) – часть генеральной совокупности, определенным способом отобранная с целью исследования и получения выводов о свойствах и характеристиках генеральной совокупности.

С выборкой связано такое понятие, как репрезентативность выборки. «Насколько выборка репрезентативна?» – любимый вопрос непрофессионалов, не имеющих собственного опыта сбора и обработки (анализа) данных. Поэтому на этот вопрос бывает часто трудно ответить. С другой стороны, организация сбора данных (формирования выборок) сомнительной репрезентативности часто является механизмом манипулирования данными с целью подтверждения желаемого результата.

Рассмотрим классическое определение репрезентативности выборки.

Репрезентативность (представительность) – соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом. Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была собрана. Иначе говоря, важно четко определить, что является генеральной совокупностью.

Репрезентативность можно определить как свойство выборочной совокупности представлять параметры генеральной совокупности, значимые с точки зрения задач исследования, то есть некоторая совокупность признаков для одних задач можно считать

репрезентативной, а для других задач не репрезентативной. Абсолютно репрезентативных выборок не бывает. Репрезентативность можно рассматривать только в статистическом смысле.

Например, если при сборе данных производился опрос мужчин и женщин. В результате оказалось, что женщин было опрошено больше, чем мужчин (в смысле присутствия в генеральной совокупности). При механистическом подходе такую выборку нельзя признать репрезентативной. Если, например, рассмотреть отношение опрошенных к системе образования страны, то скорей всего окажется, что отношение не зависит от пола, а следовательно, с точки зрения задачи вполне можно считать выборку репрезентативной. В других вопросах мнение мужчин и женщин может в статистическом смысле различаться, и тогда необходимо рассматривать две выборки по отдельности. Но в этом случае и определение генеральных совокупностей будет разным. Добиться репрезентативности часто можно путем использования операции группировки данных.

Реально изучаемые объекты и явления практически всегда имеют многопризнаковую природу. Поэтому для описания свойств объектов и явлений используются многомерные данные.

Исходная информация об объектах или явлениях (многомерные данные) чаще всего представляется в виде таблиц, состоящих из *n* строк и *m* столбцов. Строки таблицы $a_1, a_2, \ldots, a_i, \ldots, a_n$ отражают информацию об объектах, а столбцы $X_1, X_2, \ldots, X_j, \ldots, X_m$ отражают свойства (признаки, характеристики) этих объектов или явлений. На пересечении *i*-й строки и *j*-го столбца указывается значение (x_{ij}) *j*-го признака у *i*-го объекта. Такая таблица называется таблицей «объект – свойство» или просто таблицей данных.

Информация в таблице данных может быть записана различными способами. Для обеспечения однозначного понимания данных различными исследователями и компьютерными программами эти данные записываются по вполне определенным правилам, разработкой которых занимается специальная научная дисциплина «Теория измерений». В соответствии с этой теорией при планировании процесса сбора и обработки (анализа) данных следует, прежде всего, установить типы шкал, в которых измеряются те или иные признаки объектов. Тип шкалы задает группу допустимых преобразований шкалы.

Рассмотрим основные типы шкал измерения и соответствующие им группы допустимых преобразований.

Все шкалы делят на две группы – **шкалы качественных признаков и шкалы количественных признаков**.

К шкалам качественных признаков относятся номинальная и порядковая шкалы.

Шкала наименований (номинальная шкала). Измерения в этой шкале призваны для того, чтобы различать объекты, то есть фиксируются только два отношения: «равно» «не равно». Единственно допустимой операцией с измерениями в номинальной шкале является счет. Фиксируются такие характеристики, как собственные имена людей, национальность, название населенных пунктов. С этими измерениями недопустимы математические операции сложения или умножения. Не имеет смысла складывать, например, номера телефонов.

Порядковая шкала – это шкала рангов, в которой числа присваиваются объектам для отражения относительной выраженности некоторых характеристик у тех или иных объектов. Простейшим примером служат оценки знаний учащихся. В этой шкале можно задать профессиональный статус. Таблица данных содержит информацию только трех эмпирических отношений: «<, >, =». Допустимыми преобразованиями для данного типа шкал являются все монотонные преобразования, т.е. такие, которые не нарушают порядка следования значений измеренных величин. Эти данные не содержат информации, насколько один ранг отличается от другого.

Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного, например сравнительного, характера, чем количественного. Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах.

К количественным шкалам относятся «шкала интервалов», «шкала отношений» и «абсолютная шкала».

Интервальная шкала – это числовая шкала, в которой количественно равные промежутки отображают промежутки между значениями измеряемых характеристик. Интервальная шкала не только содержит всю информацию, заложенную в порядковой шкале, но и позволяет сравнить различия между ними. Разница между двумя смежными значениями шкалы идентична разнице между двумя

любыми другими смежными значениями интервальной шкалы. Между значениями интервальной шкалы существует постоянный или равный интервал. Интервальная шкала используется, например, при измерении температуры.

В интервальной шкале расположение точки отсчета не фиксируется. Точка начала отсчета и единицы измерения выбираются произвольно. Любое линейное преобразование $y = a + bx$ сохраняет свойства шкалы. Здесь x – первоначальное значение шкалы, y – преобразованное значение шкалы, b – положительная константа.

В шкале отношений по сравнению с интервальной шкалой определена еще и точка начала отсчета. Общеизвестными примерами измерения в этой шкале являются рост, вес, количество денег. Относительные шкалы допускают только преобразование $y = bx$. Один и тот же эмпирический смысл имеют значения: 12 кг, 12 000 г, 0,012 т.

Абсолютная шкала допускает преобразование только в форме тождества $y = x$. Этот тип шкалы удобен для записи количества элементов в некотором конечном множестве. Если пересчитав количество яблок, один исследователь запишет в таблицу данных значение 6, а другой VI, то достаточно знать, что 6 означает то же самое, что и VI, то есть 6=VI.

Относительная информативность измерений в различных шкалах повышается в порядке рассмотрения шкал. Различные шкалы требуют разработки своих методов анализа. При совместном рассмотрении признаков, измеренных в различных шкалах, используются методы преобразования измерительных шкал. Преобразовывать данные из одной шкалы в другую можно только с понижением мощности шкалы.

Мощность шкалы – это ее дифференцирующая способность. Менее мощные шкалы отражают меньше информации о различии объектов по измеряемому свойству. По мере возрастания мощности шкалы располагаются следующим образом: номинальная, порядковая, интервальная, шкала отношений. Определение того, в какой шкале измерен признак, является ключевым моментом анализа данных исследования.

1.2.2. Структуры и модели данных

Структура данных – одно из основополагающих понятий анализа данных с помощью компьютерных программ. Единого определения структуры данных не выработано. Рассмотрим некоторые наиболее общие определения структуры данных.

Структура данных (*data structure*) – множество элементов данных, объединенных и упорядоченных определенным образом.

Структура данных – 1) множество элементов данных и множество связей между ними; 2) систематизированный способ организации данных и доступа к ним; 3) совокупность взаимосвязанных переменных и их значений.

Рис. 1.2. Классификация структур данных

Структуры позволяют хранить их в организованной форме. В зависимости от ситуации данные должны храниться в некотором определенном формате. Их компоновка может быть эффективной в одних операциях и неэффективной в других. Цель разработчика – выбрать из возможных структур компоновки данных оптимальный вариант для возможного списка задач, которые предполагается решать на основании данных. При разработке программного обес-

печения сложность реализации и качество работы программ существенно зависит от правильного выбора структур данных. На рисунке 1.2 представлена классификация структур данных.

Любая структура данных может рассматриваться как физическая и логическая структура.

Логическая структура данных – созданное программными средствами образное, абстрактное представление структуры данных.

Логическая структура описывает способ организации, который выражается совокупностью правил размещения и связей между различными данными.

Физическая структура данных – представление структуры данных в памяти в том виде, как она выглядит «на самом деле».

Оба представления могут сосуществовать при самых разных сочетаниях их свойств: одномерная – двумерная, линейная – нелинейная и т.п.

Обобщенные структуры называют моделями данных (***data model***), поскольку они отражают представление пользователя о данных. С помощью модели данных могут быть представлены объекты предметной области и взаимосвязи между ними.

Модель данных – это совокупность структур данных и операций их обработки.

Модель данных:

– представление данных и их взаимосвязей (отношений), описывающих понятия проблемной среды. Используется для представлений структур данных на концептуальном и внешнем уровнях, но не физическом;

– понятие модели данных связано с их логической структурой;

– совокупность правил порождения структур данных в базах данных и выполнения операций над ними;

– формализованное описание структур данных и операций над ними.

Любая модель данных должна содержать три компоненты:

1. Структура данных – описывает точку зрения пользователя на представление данных.

2. Набор допустимых операций, выполняемых на структуре данных. Модель данных предполагает, как минимум, наличие языка определения данных, описывающего структуру их хранения, и языка манипулирования данными, включающего операции извлечения и модификации данных.

3. Ограничения целостности – механизм поддержания соответствия данных предметной области на основе формально описанных правил.

1.2.3. Преобразование измерительных шкал

Каждый метод многомерного статистического анализа разрабатывается для признаков, измеренных в определенной шкале измерений. Таблицы данных часто содержат признаки, измеренные в различных шкалах. Чтобы применить определенный метод обработки к группе признаков, необходимо привести их к единой шкале измерений. Преобразование признаков можно производить только с понижением мощности шкалы.

Преобразование непрерывного признака к ранговому представлению иначе называется операцией дискретизации. Такую операцию, например, проделывают при расчете частотных рядов непрерывных числовых признаков. Произведем эту операцию с данными, приведенными на рис. 1.3.

	A	B	C	D	E
16					
17	**Таблица данных BOX 7**				
18					
19	№	Признаки		Дискретизация	
20		X	Y	Z1	Z2
21	1	3,94	20,25	4	3
22	2	6,22	13,98	5	2
23	3	-2,04	23,09	2	4
24	4	0,02	20,16	2	3
25	5	=ЕСЛИ(И(B21>$AB$17;B21<$AC$17);1;ЕСЛИ(И(B			
26	6	21>$AC$17;B21<$AD$17);2;ЕСЛИ(И(B21>$AD$17;			
27	7	B21<$AE$17);3;ЕСЛИ(И(B21>$AE$17;B21<$AF$1			
28	8	7);4;ЕСЛИ(И(B21>$AF$17;B21<$AG$17);5;0)))))			
29	9	0,83	17,87	3	3
112	92	0,88	20,85	3	4
113	93	4,14	19,89	4	3
114	94	4,20	14,76	4	2
115	95	2,21	16,51	3	3
116	96	-1,16	20,69	2	4
117	97	-0,17	21,02	2	4
118	98	-3,06	20,84	1	4
119	99	3,60	13,20	4	2
120	100	0,19	22,86	3	4

Рис. 1.3. Пример выполнения операции дискретизации в EXCEL

Для выполнения операции потребуются данные по расчету таблицы параметров выборки, приведенные на рис. 1.4.

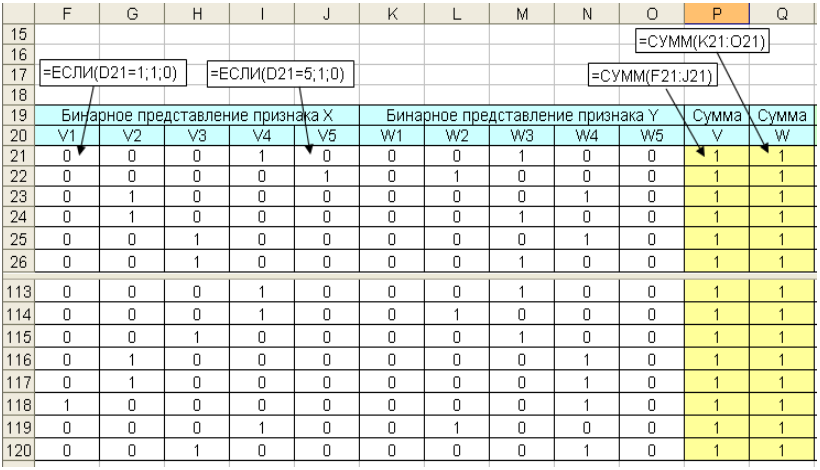
	AA	AB	AC	AD	AE	AF
1	параметры				Признаки	
2					X	Y
3	Среднее значение				0,86	20,30
4	Дисперсия				5,69	17,80
5	Среднеквадратичное отклонение				2,38	4,22
6	Минимальное значение				-4,51	7,41
7	Максимальное значение				7,07	29,29
8	Нижняя граница диапазона				-4,56	7,30
9	Верхняя граница диапазона				7,13	29,40
10	Ширина диапазона				11,70	22,10
11	Шаг по интервалу				2,34	4,42
12	Дельта				0,06	0,11
13	Колчиство интервалов				5	5

Рис. 1.4. Расчет параметров выборки

Самой бедной (маломощной) шкалой считается бинарная шкала. В ней признак может принимать только два значения: ноль или единица (истина или ложь). В некоторых задачах анализа, когда нужно использовать самые разнообразные признаки, оказывается удобным привести измерения к единой бинарной шкале. Переход к бинарной шкале осуществляется при ранговом или порядковом представлении признака. Если какой-то признак представлен в непрерывной шкале, то его всегда можно привести к ранговому представлению с помощью операции дискретизации. Будем использовать ранговое представление признаков из таблицы данных рис. 1.3.

При бинарном представлении каждому ранговому признаку ставится в соответствие k бинарных признаков (k – количество возможных различных значений рангового признака). Бинарное представление признаков X и Y приведено на рис. 1.5.

Это соответственно признаки $v = (v_1, v_2, \ldots, v_k)$ и $w = (w_1, w_2, \ldots, w_k)$. Преобразование производится по следующей схеме. Если исходный ранговый признак принимает значение ранга с номером g, то бинарный признак с номером g принимает значение 1, а все остальные компоненты бинарного вектора принимают значение 0. Сумма значений бинарного вектора будет равна 1. Например, если число ранговых значений g=5, то ранговое значение 3 в бинарном представлении будет иметь вид (0,0,1,0,0).



	F	G	H	I	J	K	L	M	N	O	P	Q
15												
16												
17												
18												
19	Бинарное представление признака X					Бинарное представление признака Y					Сумма	Сумма
20	V1	V2	V3	V4	V5	W1	W2	W3	W4	W5	V	W
21	0	0	0	1	0	0	0	1	0	0	1	1
22	0	0	0	0	1	0	1	0	0	0	1	1
23	0	1	0	0	0	0	0	0	1	0	1	1
24	0	1	0	0	0	0	0	1	0	0	1	1
25	0	0	1	0	0	0	0	0	1	0	1	1
26	0	0	1	0	0	0	0	1	0	0	1	1
113	0	0	0	1	0	0	0	1	0	0	1	1
114	0	0	0	1	0	0	1	0	0	0	1	1
115	0	0	1	0	0	0	0	1	0	0	1	1
116	0	1	0	0	0	0	0	0	1	0	1	1
117	0	1	0	0	0	0	0	0	1	0	1	1
118	1	0	0	0	0	0	0	0	1	0	1	1
119	0	0	0	1	0	0	1	0	0	0	1	1
120	0	0	1	0	0	0	0	0	1	0	1	1

Рис. 1.5. Пример преобразования признаков к бинарному представлению в EXCEL

На рисунке 1.5 отражены результаты преобразования признаков X и Y к бинарному виду. В столбцах таблицы данных V и W рассчитаны суммы значений по двум бинарным векторам. Соответственно суммы по столбцам V и W будут равны количеству наблюдений в таблице данных.

Для автоматизации преобразования измерительных шкал нами было реализовано специальное инструментальное средство.

Некоторые методы обработки многомерных статистических данных требуют предварительной нормировки данных. Нормировка данных состоит в преобразовании данных к новой форме представления. Такие преобразования позволяют исключить влияние на результаты анализа принятых единиц измерения.

Нормировка данных требуется, когда несовместимость единиц измерений переменных может отразиться на результатах, и рекомендуется, когда итоговые отчеты могут быть улучшены, если выразить результаты в определенных понятных/совместимых единицах.

Рассмотрим наиболее распространенные способы нормировки:

– центрирование;

– нормировка по максимальному значению;

– нормировка по минимальному значению;

– нормировка по среднему значению.

Приведем формулы для выполнения нормировок (1.1–1.4):

$$z_{ij} - \frac{x_{ij} - \overline{X_j}}{\sigma_{x_j}} \quad (1.1)$$

$$z_{ij} = \frac{x_{ij}}{x_{выч_o}}, \quad (1.2)$$

$$z_{ij} = \frac{x_{ij}}{x_{\min x_j}} \quad (1.3)$$

$$z_{ij} = \frac{x_{ij}}{\overline{X_j}} \quad (1.4)$$

Рассмотрим пример выполнения нормировки признаков X и Y. Исходные значения признаков приведены на рис. 1.6.

	A	B	C
17	**Таблица данных BOX 7**		
18			
19	№	Признаки	
20		X	Y
21	1	3,94	20,25
22	2	6,22	13,98
23	3	-2,04	23,09
24	4	0,02	20,16
25	5	2,43	20,61
26	6	0,24	16,91
27	7	1,72	20,71
28	8	3.21	15.31
112	92	0,88	20,85
113	93	4,14	19,89
114	94	4,20	14,76
115	95	2,21	16,51
116	96	-1,16	20,69
117	97	-0,17	21,02
118	98	-3,06	20,84
119	99	3,60	13,20
120	100	0,19	22,86

Рис. 1.6. Исходные значения признаков X и Y

Результаты расчета параметров признаков X и Y приведены на рис. 1.7. Коэффициент корреляции равен – 0,62. Результаты выполнения операции нормировки приведены на рис. 1.8.

	AA	AB	AC	AD	AE	AF
1	параметры				Признаки	
2					X	Y
3	Среднее значение				0,86	20,30
4	Дисперсия				5,69	17,80
5	Среднеквадратичное отклонение				2,38	4,22
6	Минимальное значение				-4,51	7,41
7	Максимальное значение				7,07	29,29
8	Нижняя граница диапазона				-4,56	7,30
9	Верхняя граница диапазона				7,13	29,40
10	Ширина диапазона				11,70	22,10
11	Шаг по интервалу				2,34	4,42
12	Дельта				0,06	0,11
13	Колчиство интервалов				5	5

Рис. 1.7. Расчет параметров признаков X и Y

Нормировка центрирование может быть выполнена с помощью функции EXCEL НОРМАЛИЗАЦИЯ (рис. 1.9).

	R	S	T	U	V	W	X	Y
14								
15			=B21/$AE$7					
16								
17	=(B21-$AE$3)/$AE$5				=B21/$AE$6		=B21/$AF$3	
18								
19	Нормировка 1		Нормировка 2		Нормировка 3		Нормировка 4	
20	X1	Y1	X2	Y2	X3	Y3	X4	Y4
21	1,29	-0,01	0,56	0,69	-0,87	2,73	0,19	23,48
22	2,24	-1,50	0,88	0,48	-1,38	1,89	0,31	16,21
23	-1,22	0,66	-0,29	0,79	0,45	3,12	-0,10	26,78
24	-0,35	-0,03	0,00	0,69	-0,01	2,72	0,00	23,37
25	0,66	0,07	0,34	0,70	-0,54	2,78	0,12	23,90
112	0,01	0,13	0,12	0,71	-0,20	2,81	0,04	24,17
113	1,38	-0,10	0,59	0,68	-0,92	2,68	0,20	23,06
114	1,40	-1,31	0,59	0,50	-0,93	1,99	0,21	17,11
115	0,57	-0,90	0,31	0,56	-0,49	2,23	0,11	19,15
116	-0,85	0,09	-0,16	0,71	0,26	2,79	-0,06	23,98
117	-0,43	0,17	-0,02	0,72	0,04	2,84	-0,01	24,37
118	-1,65	0,13	-0,43	0,71	0,68	2,81	-0,15	24,17
119	1,15	-1,68	0,51	0,45	-0,80	1,78	0,18	15,30
120	-0,28	0,61	0,03	0,78	-0,04	3,08	0,01	26,51

Рис. 1.8. Результаты нормировки признаков X и Y

Нормирование данных является необходимым начальным этапом преобразования данных при использовании разных многомерных статистических методов: снижения размерности признакового пространства (факторный, компонентный анализ, классификации объектов (кластерный анализ) и др.), особенно если переменные измерены в шкалах, существенно различающихся в величинах (микроны единиц – миллиарды единиц).

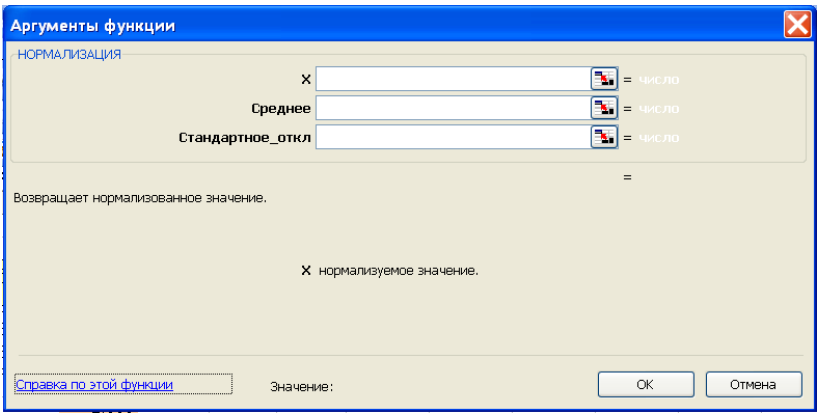


Рис. 1.9. Функция EXCEL НОРМАЛИЗАЦИЯ

Часто выбор способа нормирования осуществляется в процессе компьютерного эксперимента.

1.3. Способы моделирования данных

1.3.1. Простейшие способы моделирования данных

Моделирование данных – это искусственное создание случайных данных, обладающих заданными свойствами. Для моделирования данных используются специальные компьютерные программы, которые называются генераторами данных.

Задача моделирования данных имеет очень важное значение при изучении методов компьютерного анализа данных. Зачем нужно моделирование данных? Моделирование данных необходимо для изучения и тестирования программного обеспечения, предназначенного для решения задач анализа данных. С помощью моделирования дан-

ных можно исследовать возможности вновь создаваемых методов анализа данных.

Чаще всего исследователи при решении задач анализа данных используют стандартное программное обеспечение или специализированные пакеты по обработке данных. Программное обеспечение, как правило, включает целый ряд программ, процедур или функций анализа данных, представляющих собой программную реализацию известных в теории методов обработки данных. Для того чтобы использовать ту или иную программу обработки данных, необходимо очень хорошо знать метод обработки, заложенный в ее основу. Однако знание теории не всегда гарантирует правильное использование программы. Необходимо еще знать устройство программы на уровне входа и выхода. Этому способствуют описания программ. Описания программ далеко не всегда могут быть правильно поняты конкретным исследователем, который впервые сталкивается с новым программным продуктом. Все-таки язык описания программ не так глубоко формализован, как строгий математический аппарат, описывающий метод исследования. С другой стороны, все особенности использования программы их создатели и не в состоянии описать. Они не могут себе даже представить и предусмотреть все варианты неправильного использования программы. Таким образом, перед использованием программы исследователь должен на реальных данных убедиться, все ли правильно он понимает в ее работе. Это он может сделать, решая примеры на данных, обладающих известными свойствами (модельных данных). Проблема изучения программных средств особенно остро проявляется в сложных программах. Усложняется она еще, когда приходится иметь дело с англоязычным интерфейсом или переведенным на русский язык. При переводе специальной терминологии очень часто возникают неточности.

Поскольку теория анализа данных состоит в изучении и разработке методов анализа и программного обеспечения, реализующих эти методы, аппарат моделирования данных является необходимым инструментом для работы. Необходимость в модельных данных сохраняется при изучении любых вновь появляющихся программных продуктов. Поэтому исследователю, занимающемуся анализом данных, необходимы навыки моде-

лирования данных. Более того, даже разработчики новых методов анализа данных и программного обеспечения к ним тоже не могут обойтись без таких данных. Да и любая программа, созданная пользователем для обработки реальных данных с использованием стандартного программного обеспечения, вначале требует тестирования.

Моделирование случайных величин (данных), обладающих заданными свойствами, производится на основе датчика случайных чисел. Случайные числа моделируют равномерное распределение случайной величины на интервале (0,1). В стандартном пакете EXCEL моделирование случайных чисел производится с помощью функции СЛЧИС(), которая входит в состав математических функций. При выборе данной функции выводится диалоговое окно (рис. 1.10), в котором ничего не вводится, а просто нажимается кнопка «ОК». Окно содержит только информацию об особенностях работы программы.

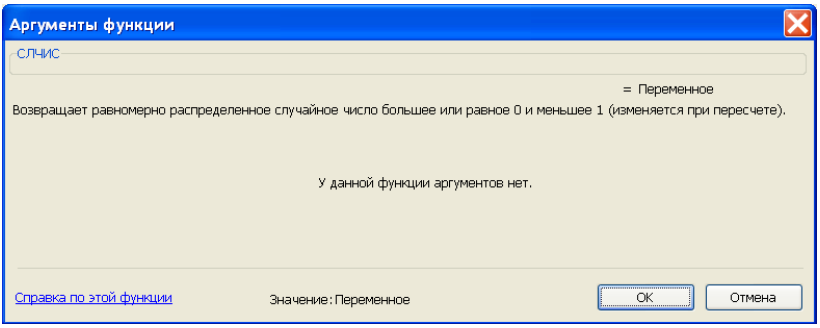


Рис. 1.10. Диалоговое окно функции СЛЧИС()

Для создания последовательности чисел в таблице данных необходимо скопировать содержимое ячейки, содержащей функцию СЛЧИС() по столбцу (протянуть по столбцу). В диалоговом окне функции СЛЧИС() указано, что значение в ячейках, содержащих функцию СЛЧИС(), изменяется при пересчете таблицы данных. Это означает, что при внесении любых изменений на листе EXCEL функция выдаст новое значение. Для некоторых задач моделирования данных такое свойство функции оказывается очень полезным. Однако для решения примеров по апробации программного обеспечения необходимо иметь фиксированные

значения. Для фиксации значений случайных чисел, полученных с помощью функции СЛЧИС(), необходимо выполнить следующие действия:

– после копирования функции методом протягивания, не снимая выделения ячеек в столбце, нажать пиктограмму «Копировать»;

– выбрать в меню «Правка» команду «Специальная вставка»;

– в открывшемся окне выбрать «значения» (рис. 1.11) и нажать кнопку «ОК»;

– завершить всю операцию нажатием клавиши «Enter» на клавиатуре.

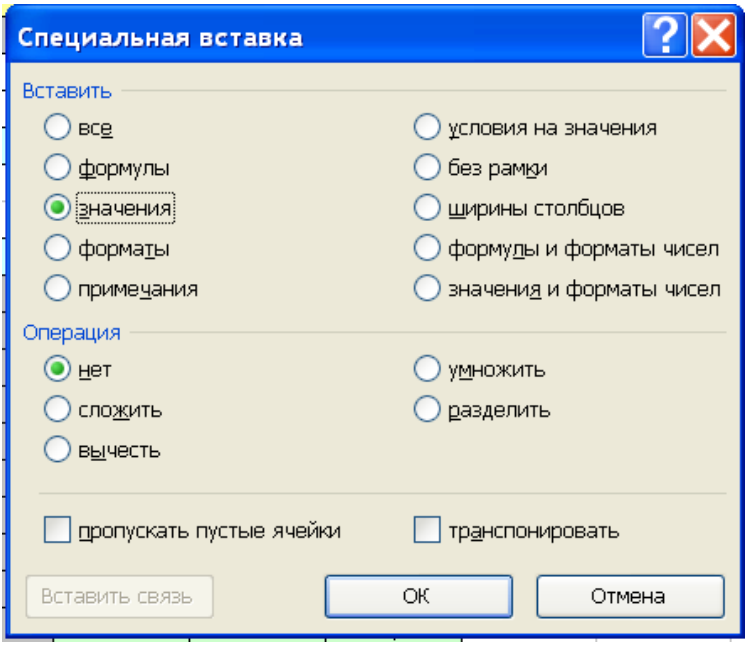


Рис. 1.11. Диалоговое окно функции «Специальная вставка»

Смоделируем две последовательности случайных чисел и разместим их в первых двух столбцах таблицы данных (рис. 1.12). Математическое обозначение таких чисел будет x_{i1} и x_{i2} $(i = \overline{1,n})$. Эти числа нам понадобятся для моделирования других случайных величин.

	A	B	C	D
10	a= 5		b= 15	

=($D$10-$B$10)*P10+$B$10

СЛЧИС()

Таблица данных BOX 2

№	Признаки							
	X1	X2	X3	X4	X5	X6	X7	X8
1	0,45	0,03	9,48	0,27	1,24	5,53	19,96	9,10
2	0,68	0,30	11,80	0,84	-0,25	6,68	14,00	6,17
3	0,49	0,78	9,95	-1,17	0,22	2,67	15,88	10,73
4	0,92	0,17	14,23	0,35	0,19	5,70	15,78	13,83
5	0,41	0,26	9,10	1,33	-0,08	7,67	14,66	13,60
6	0,05	0,12	5,47	1,67	1,83	8,33	22,31	15,20
7	0,56	0,53	10,58	-0,22	-1,06	4,56	10,77	3,45
95	0,59	0,41	10,92	0,57	-0,85	6,15	11,61	13,71
96	0,50	0,54	10,03	-0,29	-1,14	4,42	10,46	9,07
97	0,44	0,35	9,37	1,04	-0,76	7,08	11,98	7,48
98	0,42	0,85	9,23	-1,04	0,80	2,92	18,20	7,54
99	0,27	0,04	7,66	0,44	1,57	5,88	21,27	8,71
100	0,73	0,49	12,29	0,04	-0,79	5,08	11,82	8,64

Рис. 1.12. Таблица данных «объект-свойство»

Рассмотрим простейший случай моделирования данных произвольного равномерного распределения. Случайные числа являются моделью закона равномерной плотности в простейшем случае.

Общий вид функции равномерного распределения имеет вид (1.5):

$$\begin{cases} f(x) = c = \dfrac{1}{b-a} & \text{при } a < x < b \\ f(x) = 0 & \text{при } x < a \text{ или } x > b \end{cases} \qquad (1.5)$$

График функции равномерного распределения приведен на рис. 1.13.

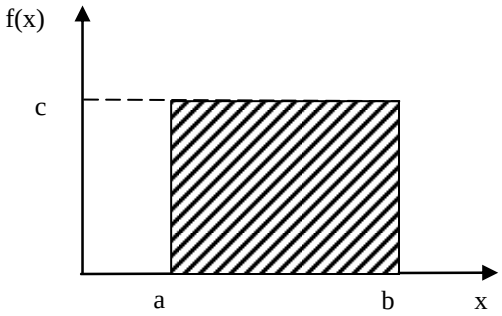


Рис. 1.13. График функции равномерного распределения

Математическое ожидание случайной равномерно распределенной величины рассчитывается по формуле (1.6):

$$m_x = \frac{a+b}{2}. \qquad (1.6)$$

Дисперсия и среднеквадратичное отклонение равномерно распределенной случайной величины рассчитываются по формулам (1.7) и (1.8):

$$D_x = \frac{(b-a)^2}{12}, \qquad (1.7)$$

$$\sigma_x = \sqrt{D_x} = \frac{b-a}{2\sqrt{3}}. \qquad (1.8)$$

Функция СЛЧИС() моделирует равномерное распределение при a=0 и b=1. Математическое ожидание такого закона – распределение $m_x = 1/2$ и дисперсия $D_x = 1/12$.

Последовательность значений произвольного равномерного распределения может быть получена путем преобразования последовательности случайных чисел, полученных с помощью функции СЛЧИС(). Пусть x_{i1} $(i = \overline{1,n})$ случайные числа (первый столбец таблицы данных). Тогда случайные значения, распределенные по равномерному закону распределения с параметрами $a = a^*$ и $b = b^*$, могут быть получены путем преобразования по формуле:

$$x_{i2} = x_{i1}(b^* - a^*) + a^*. \qquad (1.9)$$

Выберем конкретные значения параметров равномерного распределения ($a = 5$ и $b = 15$) и смоделируем равномерную случайную величину с этими параметрами. Полученные данные разместим в третьем столбце таблицы данных (рис. 1.12).

Рассмотрим простейший способ моделирования нормальной случайной величины.

Нормальный закон является наиболее распространенным законом распределения случайных величин. Функция плотности нормального закона распределения имеет вид (1.10):

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{\frac{(x-m_x)^2}{2\sigma_x^2}}. \qquad (1.10)$$

Нормальный закон распределения имеет два параметра: математическое ожидание m_x и среднеквадратичное отклонение $\sigma_x = \sqrt{D_x}$. График функции плотности нормального распределения представлен на рис. 1.14.

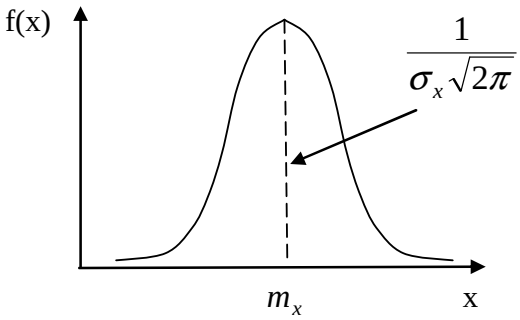


Рис. 1.14. График функции плотности нормального распределения

На рисунках 1.15 и 1.16 приведены графики функций плотности нормального распределения при различных значениях параметров ($m_{x_1} < m_{x_2} < m_{x_3}, \sigma_{x_1} < \sigma_{x_2} < \sigma_{x_3}$).

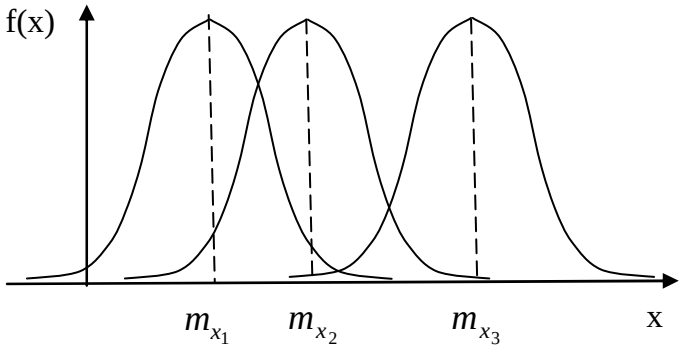


Рис. 1.15. График функции плотности нормального распределения для различных значений математического ожидания

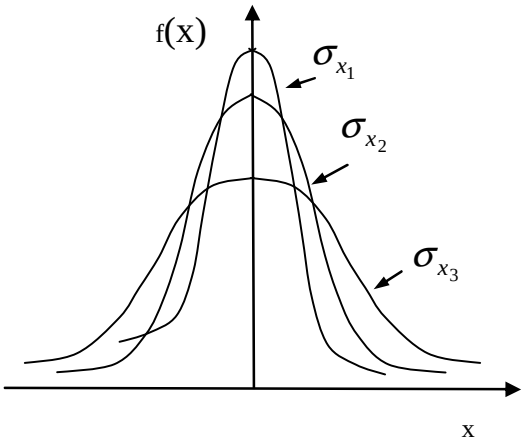


Рис. 1.16. График функции плотности нормального распределения для различных значений среднеквадратичного отклонения

Особое место занимает так называемое стандартное нормальное распределение. Стандартным нормальным законом распределения называется нормальное распределение с параметрами $m_x = 0$, $\sigma_x = 1$. Для обозначения случайной величины, распределенной по нормальному закону, используется обозначение $n(m_x, \sigma_x)$. Соответственно, стандартное нормальное распределение обозначается $n(0,1)$.

Для моделирования стандартной нормальной случайной величины существует множество способов. Самый простой способ моделирования состоит в преобразовании двух случайных чисел α_1 и α_2:

$$\eta_1 = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2,\ \eta_2 = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2, \quad (1.11)$$

где η_1, η_2 – случайные числа, распределенные по стандартному нормальному закону.

Запишем эти формулы для таблицы данных. Смоделируем две последовательности чисел, распределенных по стандартному нормальному закону, и расположим эти последовательности соответственно в четвертом и пятом столбцах таблицы данных. Эти числа получим в результате преобразования случайных чисел, расположенных в первом и втором столбцах таблицы данных.

$$x_{i4} = \sqrt{-2\ln(x_{i1})}\sin(2\pi x_{i2}), \quad (1.12)$$

$$x_{i5} = \sqrt{-2\ln(x_{i1})}\cos(2\pi x_{i2}). \quad (1.13)$$

Моделирование произвольного нормального распределения производится путем преобразования последовательности значений случайной величины, подчиняющейся стандартному нормальному закону. Преобразование состоит в умножении стандартной случайной величины на среднеквадратичное отклонение моделируемого закона распределения и прибавлении математического ожидания моделируемого закона. Смоделируем две последовательности нормально распределенных чисел с различными параметрами и расположим модельные данные соответственно в шестом и седьмом столбцах таблицы данных. Зададим следующие параметры моделирования: $n(m_{x_6} = 5, \sigma_{x_6} = 2)$, $n(m_{x_7} = 15, \sigma_{x_7} = 4)$. Преобразование производится по формулам:

$$x_{i6} = x_{i1}\sigma_{x_6} + m_{x_6}, \quad (1.14)$$

$$x_{i7} = x_{i2}\sigma_{x_7} + m_{x_7}. \quad (1.15)$$

Формулы преобразования 1.11–1.15 в EXCEL будут иметь вид рис. 1.17.

	A	B	C	D	E	F	G
14	среднее признака X6					m1	5
15	стандартное отклонение признака X6					s1	2
16	среднее признака X7					m2	15
17	стандартное отклонение признака X7					s2	4

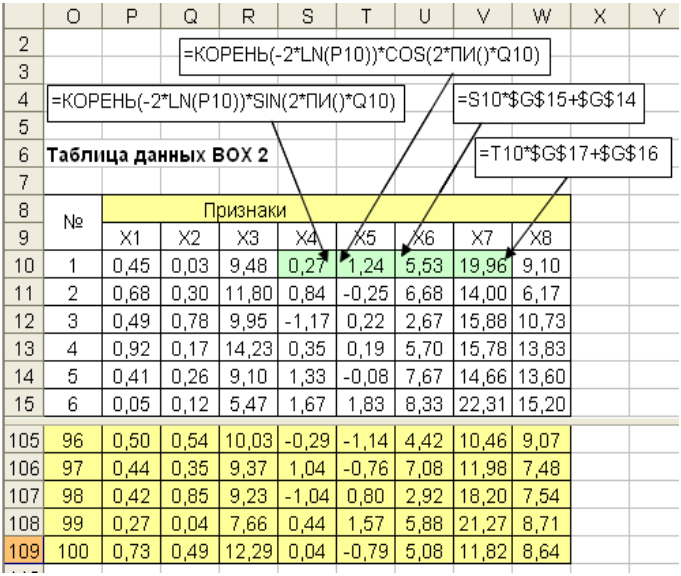


Таблица данных BOX 2

№	Признаки							
	X1	X2	X3	X4	X5	X6	X7	X8
1	0,45	0,03	9,48	0,27	1,24	5,53	19,96	9,10
2	0,68	0,30	11,80	0,84	-0,25	6,68	14,00	6,17
3	0,49	0,78	9,95	-1,17	0,22	2,67	15,88	10,73
4	0,92	0,17	14,23	0,35	0,19	5,70	15,78	13,83
5	0,41	0,26	9,10	1,33	-0,08	7,67	14,66	13,60
6	0,05	0,12	5,47	1,67	1,83	8,33	22,31	15,20
96	0,50	0,54	10,03	-0,29	-1,14	4,42	10,46	9,07
97	0,44	0,35	9,37	1,04	-0,76	7,08	11,98	7,48
98	0,42	0,85	9,23	-1,04	0,80	2,92	18,20	7,54
99	0,27	0,04	7,66	0,44	1,57	5,88	21,27	8,71
100	0,73	0,49	12,29	0,04	-0,79	5,08	11,82	8,64

Рис. 1.17. Формулы преобразования в EXCEL

Рассмотрим моделирование законов распределения случайных величин средствами EXCEL.

Некоторые законы распределения могут быть смоделированы средствами EXCEL. Программа моделирования включена в состав встроенного в EXCEL специального пакета программ («Пакет анализа»). Этот пакет устанавливается в меню «сервис» – «Надстройки» (рис. 1.18).

После установки пакета в меню «Сервис» появляется строчка «Анализ данных» (рис. 1.19). Выберите в меню строчку «Анализ данных». По этой команде раскрывается список программ, входящих в пакет (рис. 1.20).

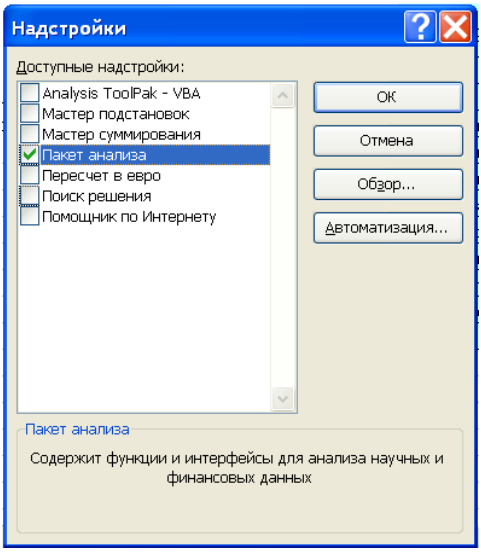


Рис. 1.18. Установка программ «Пакет анализа»

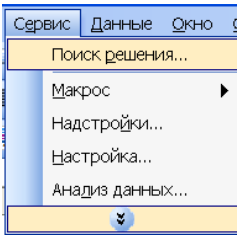


Рис. 1.19. Вызов пакета «Анализ данных»

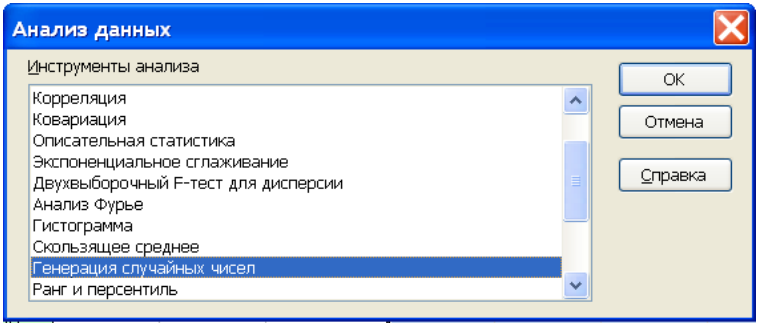


Рис. 1.20. Список программ пакета «Анализ данных»

Программы моделирования случайных закономерностей иначе называются генераторами. Выбор программы моделирования осуществляется в диалоговом окне (рис. 1.20). При запуске программы моделирования выводится диалоговое окно, где определяются параметры моделируемой случайной величины (рис. 1.21). В списке законов распределения выберем нормальное (распределение). Смоделируем нормальное распределение с параметрами $n(m_{x_в} = 10, \sigma_{x_в} = 3)$ и разместим данные в восьмом столбце таблицы данных (рис. 1.17).

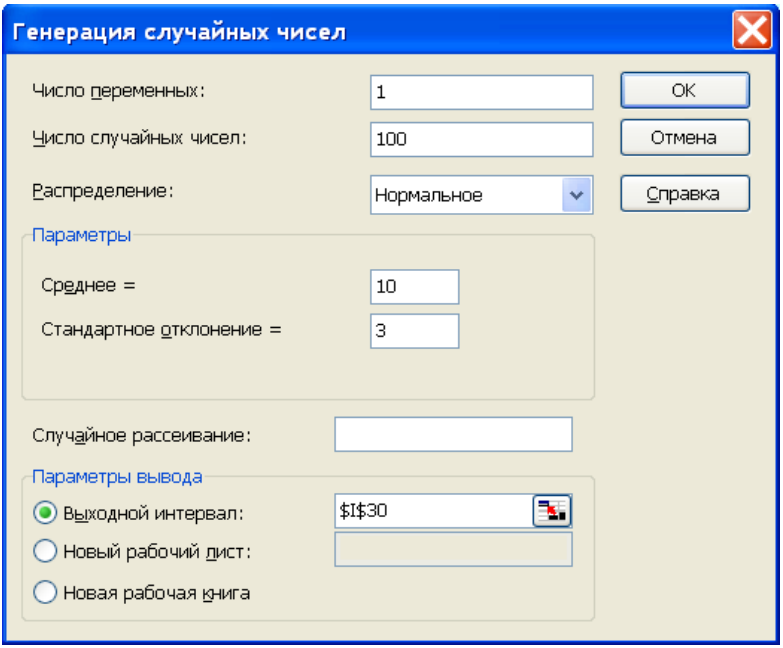


Рис. 1.21. Диалоговое окно моделирования нормального закона распределения

1.3.2. Универсальные способы моделирования данных

Наиболее распространенные датчики генерации случайных величин моделируют данные наиболее известных теоретических распределений. Рассмотрим способ моделирования, который позволяет моделировать эмпирические данные, то есть такие, которые трудно соотнести с каким-либо теоретическим законом распределения.

Предположим, что некоторый исследователь опубликовал свои исследования по изучению случайной величины, которую он наблюдал при изучении реального процесса или явления. Результаты демонстрируются гистограммой. Задача состоит в том, чтобы смоделировать данные, подчиняющиеся представленной гистограмме.

Простейшим методом моделирования эмпирических данных является метод неравномерной рулетки. Этот метод рулетки, его удобно рассмотреть на основе использования данных некоторого гипотетического примера. Пусть данные для гипотетического примера заданы некоторым частотным рядом (рис. 1.22). Гистограмма, построенная по данным гипотетического примера, приведена на рис. 1.23.

	A	B	C	D	E	F	G	H	I
8	Параметры частотного ряда			Номер интервала					
9				0	1	2	3	4	5
10	Границы интервалов			1,5	1,62	1,74	1,86	1,98	2,1
11	Относительные частоты				0,3	0,35	0,27	0,06	0,02
12	Середины интервалов				1,56	1,68	1,8	1,92	2,04
13	Границы интервалов (текст)				1,50÷1,62	1,62÷1,74	1,74÷1,86	1,86÷1,98	1,98÷2,1

Рис. 1.22. Исходные данные гипотетического примера

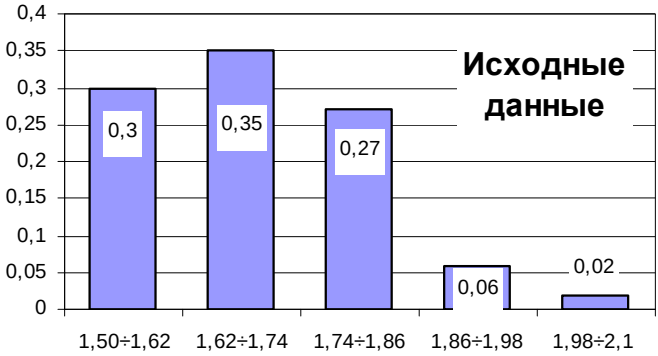


Рис. 1.23. Гистограмма эмпирического частотного ряда

Построим неравномерную рулетку (рис. 1.24). Для этого выложим на отрезке (0,1) частоты заданного частотного ряда (цифры на оси сверху) и рассчитаем накапливаемые частоты (цифры на оси

снизу). Данные по неравномерной рулетке разместим в таблице (рис. 1.25).

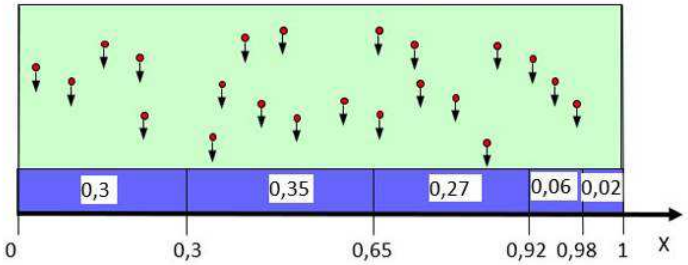


Рис. 1.24. Графическое представление неравномерной рулетки

	L	M	N	O	P	Q
1	Неравномерная рулетка					
2	1	2	3	4	5	6
3	0	0,3	0,65	0,92	0,98	1

Рис. 1.25. Табличное представление неравномерной рулетки

Принцип работы генератора на основе неравномерной рулетки очень прост. Представим себе, что интервалы неравномерной рулетки – это некоторые сосуды различной ширины. Если сверху ссыпать сухой горох сразу во все сосуды, то в результате количество гороха в различных сосудах будет пропорционально ширине сосуда. Теперь, если связать попадание гороха в сосуд с выпадением случайной величины, принадлежащей одному из интервалов диапазона моделируемой величины (границы интервалов в таблице на рис. 25), можно утверждать, что случайная величина будет выпадать пропорционально частоте моделируемого эмпирического частотного ряда. Выпадение определенного интервала мы можем связать с событием выпадения значения моделируемой величины, равного середине соответствующего интервала. Таким образом, количество различных значений, которое будет встречаться в выборке, будет равно количеству интервалов. Эти значения будут встречаться с частотой моделируемого частотного ряда эмпирической случайной величины.

Рассмотрим реализацию метода в EXCEL.

Первый шаг. Создать последовательность случайных чисел и разместить их в таблицу данных. Столбец B на рис. 1.26.

Второй шаг. Записать формулу выбора случайной величины по методу неравномерной рулетки во втором столбце таблицы данных (столбец С – модельные данные 1). В формуле используются данные по неравномерной рулетке из таблицы рис. 1.25 и середины интервалов моделируемой случайной величины из таблицы рис. 1.21. Скопируем формулу по всему столбцу С (рис. 1.26).

	A	B	C	D	E	F	G	H	I
22	**Таблица данных BOX 4**								
23									
24	№	Случайные числа	Модельные двнные 1	Модельные двнные 2					
25	1	0,74	1,80	1,56					
26	2	0,13	1,56	1,56					
27	3	0,77	1,80	1,56					
28	4	0,42	1,68	1,56					
29	5	0,33	1,68	1,68					
30	6	0,06	1,56	1,56					
31	7	0,09	1,56	1,8					
32	8	0,40	1,68	1,56					
33	9	0,30	1,56	1,68					
34	10	0,42	1,68	1,68					
118	94	0,46	1,68	1,56					
119	95	0,23	1,56	1,68					
120	96	0,29	1,56	1,56					
121	97	0,03	1,56	1,56					
122	98	0,16	1,56	1,8					
123	99	0,10	1,56	1,8					
124	100	0,73	1,80	1,8					

=ЕСЛИ(И(B25>$L$3;B25<=$M$3);$E$12;ЕСЛИ(И(B25>$M$3;B25<=$N$3);$F$12;ЕСЛИ(И(B25>$N$3;B25<=$O$3);$G$12;ЕСЛИ(И(B25>$O$3;B25<=$P$3);$H$12;ЕСЛИ(И(B25>$P$3;B25<=$Q$3);$I$12;0)))))

Рис. 1.26. Формула выбора случайной величины по методу неравномерной рулетки

	L	M	N	O	P	Q	R	S	T	U	V	W
7				=СЧЁТЕСЛИ($C$25:$C$124;E12)								
8												
9	Вид частотного ряда						Частоты модельных данных					Сумма
10	Абсолютные частоты (модельные данные 1)						31	31	32	3	3	100
11	Относительные частоты (модельные данные 1						0,31	0,31	0,32	0,03	0,03	1
12	Абсолютные частоты (модельные данные 2)						35	31	28	4	2	100
13	Относительные частоты (модельные данные 2						0,35	0,31	0,28	0,04	0,02	1
14								=СЧЁТЕСЛИ($D$25:$D$124;F12)				
15												

Рис. 1.27. Частотный ряд модельных данных

Проанализируем результаты моделирования. Для этого построим частотный ряд по модельным данным (рис. 1.27) и гистограмму (рис. 1.28).

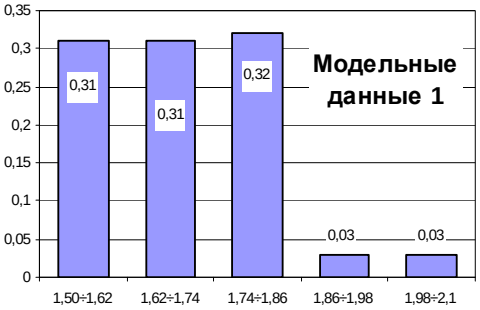


Рис. 1.28. Гистограмма модельных данных

Сравнение формы исходного частотного ряда и модельного свидетельствует об их сходстве. Недостаток метода видится в ограниченном количестве вариантов различных значений в модельном ряду. Несомненное преимущество метода состоит в простоте его реализации.

Аналогичный результат можно получить с помощью программы генерации данных в пакете «Анализ данных». Для этого в списке программ пакета анализа выбрать программу генерации случайных чисел и определить параметры дискретного распределения (рис. 1.29).

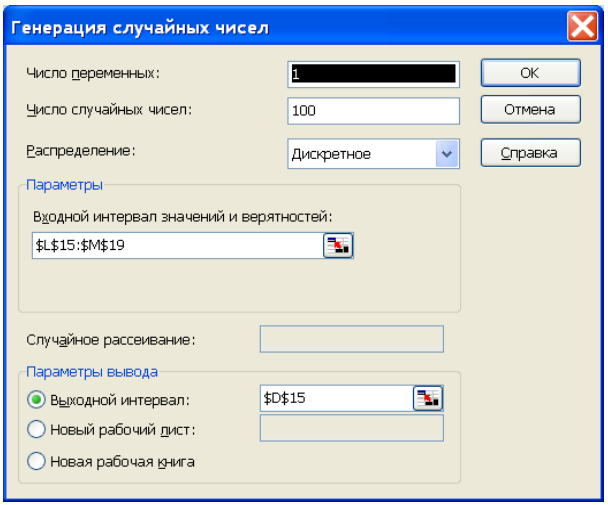


Рис. 1.29. Генерация дискретного распределения

Входной интервал значений и частот должен быть представлен в двух столбцах (рис. 1.30). В качестве выходного интервала укажем первую ячейку таблицы данных в столбце «Модельные данные 2».

	S	T	U
1	Дискретные значения		Частоты
2			
3		1,56	0,3
4		1,68	0,35
5		1,8	0,27
6		1,92	0,06
7		2,04	0,02

Рис. 1.30. Входной интервал значений и вероятностей

По результатам моделирования частотного ряда дискретного распределения («Модельные данные 2») построим частотный ряд (рис. 1.31).

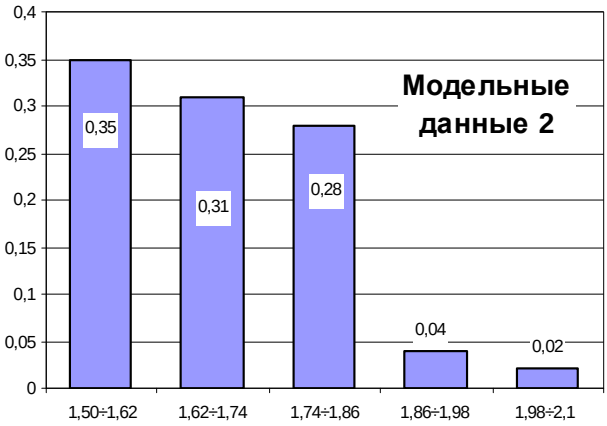


Рис. 1.31. Гистограмма модельных данных 2

Как и в случае моделирования данных с помощью метода неравномерной рулетки, частотный ряд, построенный с помощью программы генерации дискретного распределения, также имеет большое сходство с исходным частотным рядом.

Среди методов моделирования данных можно выделить метод отбраковки.

Для пояснения идеи метода отбраковки будем использовать тот же пример, что и при рассмотрении метода неравномерной рулетки. Для обеспечения согласования ссылок в формулах повторим таблицу исходных данных, используемую при рассмотрении метода неравномерной рулетки (рис. 1.32).

	A	B	C	D	E	F	G	H	I
8	Параметры частотного ряда			Номер интервала					
9				0	1	2	3	4	5
10	Границы интервалов			1,5	1,62	1,74	1,86	1,98	2,1
11	Относительные частоты				0,3	0,35	0,27	0,06	0,02
12	Середины интервалов				1,56	1,68	1,8	1,92	2,04
13	Границы интервалов (текст)				1,50÷1,62	1,62÷1,74	1,74÷1,86	1,86÷1,98	1,98÷2,1

Рис. 1.32. Исходные данные гипотетического примера

Первый шаг. Для моделирования данных потребуются два случайных признака (рис. 1.33). Таблица данных содержит 150 строк.

	A	B	C
27			
28	№	случис1	случис2
29	1	0,5182	0,69374
30	2	0,756773	0,67686
31	3	0,44567	0,72838
32	4	0,61796	0,7335
33	5	0,855891	0,32056
34	6	0,060123	0,28652
35	7	0,64862	0,40331
172	144	0,044458	0,17402
173	145	0,314683	0,84552
174	146	0,229432	0,56332
175	147	0,888999	0,38652
176	148	0,951519	0,90497
177	149	0,294527	0,25195
178	150	0,62623	0,10428

Рис. 1.33. Таблица данных (две последовательности случайных чисел)

Второй шаг. Впишем гистограмму моделируемых данных в прямоугольник (рис. 1.34).

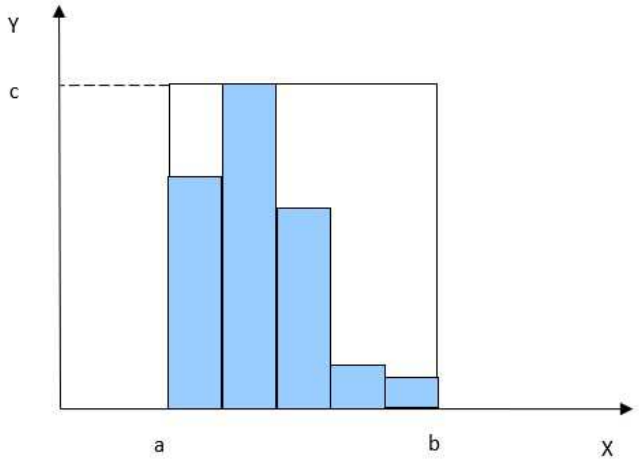


Рис. 1.34. Прямоугольник, описывающий гистограмму моделируемой случайной величины

По оси X расположим моделируемый признак. Из таблицы исходных данных следует, что диапазон значений моделируемого признака (a=1,5; b=2,1). Эти параметры определяют одну сторону прямоугольника. По оси Y будем откладывать частоты по интервалам. Максимальная частота определит другую сторону прямоугольника (c=0,35).

Преобразуем две последовательности случайных в две новых последовательности чисел X и Y так, чтобы точки, определяемые новыми признаками, попадали в прямоугольник, описывающий гистограмму (рис. 1.35).

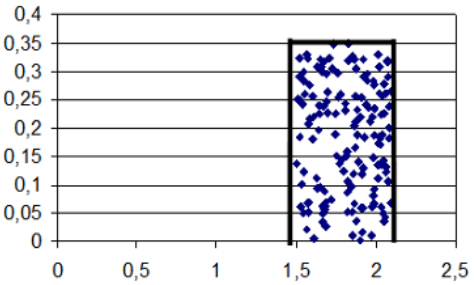


Рис. 1.35. Диаграмма рассеивания случайных признаков X и Y

Преобразование признаков производим по известным формулам:

$$x_i = (\text{случис1})_i \times (2{,}1 - 1{,}5) + 1{,}5 \,, \qquad (1.16)$$

$$y_i = (\text{случис2})_i \times 0{,}35 \,. \qquad (1.17)$$

Преобразованные данные разместим в столбцах X и Y (рис. 1.36).

	A	B	C	D	E
25					
26	=B29*($I$10-$D$10)+$D$10			=C29*$F$1	
27					
28	№	случис1	случис2	X	Y
29	1	0,5182	0,69374	1,81092	0,24281
30	2	0,756773	0,67686	1,95406	0,2369
31	3	0,44567	0,72838	1,7674	0,25493
32	4	0,61796	0,7335	1,87078	0,25672
33	5	0,855891	0,32056	2,01353	0,1122
34	6	0,060123	0,28652	1,53607	0,10028
35	7	0,64862	0,40331	1,88917	0,14116
36	8	0,271885	0,10005	1,66313	0,03502
37	9	0,589807	0,93294	1,85388	0,32653
38	10	0,542677	0,9121	1,82561	0,31924

Рис. 1.36. Расчет значений в столбцах X и Y

Третий шаг. Рассмотрим точки, построенные по столбцам X и Y, на фоне гистограммы моделируемого признака (1.37).

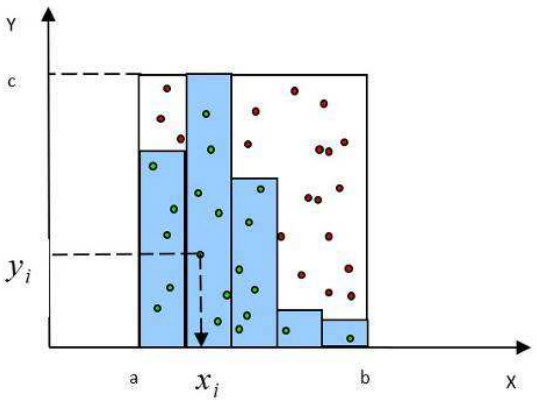


Рис. 1.37. Диаграмма рассеивания случайных признаков X и Y на фоне гистограммы моделируемого признака

Выполним операцию отбраковки, которая состоит в том, что точки, лежащие вне поля гистограммы, отбрасываются и в выборку включаются координаты X точек, попадающих в поле гистограммы. Сформируем новый признак Z, который будет содержать для принятых точек значения моделируемой выборки, а для отброшенных точек значение – 1. Расчетные формулы приведены на рис. 1.38.

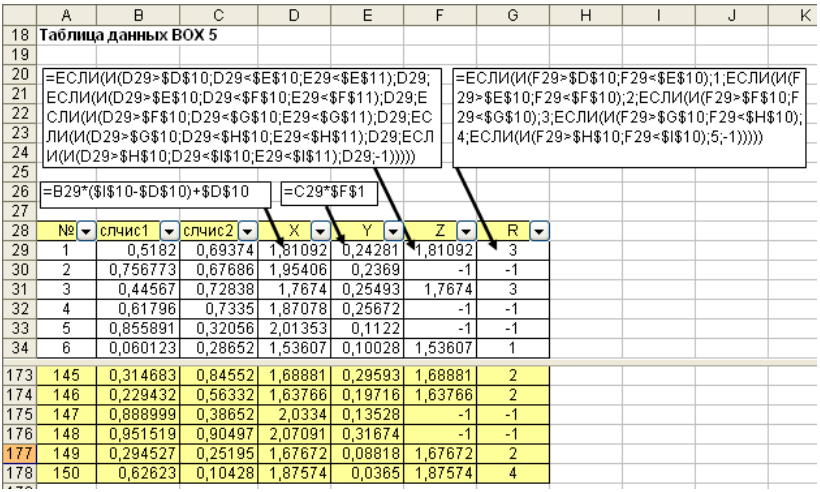


Рис. 1.38. Расчет столбцов Z и R таблицы данных

Четвертый шаг. Определяются номера интервалов для точек, включенных в модельный ряд. Результаты расчетов разместим в столбце R. Эти данные необходимы нам для расчета частотного ряда по модельным данным. Для отбракованных точек значение определим как 1. В принципе отбракованные точки можно отбросить из таблицы данных с помощью автофильтра, а затем произвести перенумерацию строк таблицы данных. Заметим, что результирующая модельная выборка будет содержать меньшее количество точек, чем исходные последовательности случайных чисел. В нашем случае из 150 точек отобрано всего 79 точек. Другими словами, мы заранее не можем точно знать, сколько точек получим в моделируемом ряду.

По модельным данным рассчитаем частотный ряд (рис. 1.39), построим гистограмму (1.40) и сравним с гистограммой заданной в исходных данных задачи. Полученные результаты показывают сходство с моделируемым частотным рядом.

	M	N	O	P	Q	R	S	T	U
3	=СЧЁТЕСЛИ($G$29:$G$178;P6)								
4									
5	Вид частотного ряда			Частоты модельных данных					Сумма
6				1	2	3	4	5	
7	Абсолютные частоты (модель)			24	30	19	6	0	79
8	Относительные частоты (модель)			0,30	0,38	0,24	0,08	0,00	1

Рис. 1.39. Частотный ряд модельных данных

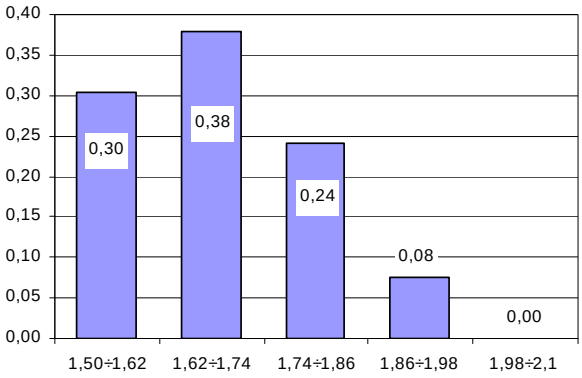


Рис. 1.40. Гистограмма модельных данных

Полученный метод применим не только к исходным данным, заданным гистограммой, но и к любой эмпирической функции, заданной аналитической формулой (рис. 1.41).

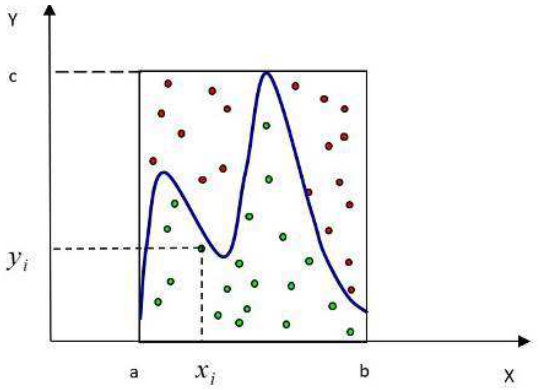


Рис. 1.41. Моделирование непрерывной функции плотности распределения

1.3.3. Моделирование многомерного нормального распределения

Средства моделирования статистических данных служат незаменимым инструментом для исследования возможностей применения того или иного статистического метода или разработки новой компьютерной технологии решения прикладных задач анализа данных.

Однако и в EXCEL, и даже в таких известных пакетах по обработке статистических данных, как STATISTICA или SPSS, представлены программные модули, позволяющие моделировать только одномерные статистические распределения.

Для моделирования многомерных данных, обладающих свойствами реальных многомерных статистических данных, нельзя использовать несколько раз процедуры моделирования одномерных данных, поскольку в этом случае будут созданы независимые признаки и совместный анализ таких данных теряет смысл.

Для более адекватного отражения реальной ситуации при моделировании многомерных данных необходимо иметь возможность воспроизведения зависимости признаков. Среди многомерных законов распределения, учитывающих зависимость признаков, наиболее известен многомерный нормальный закон. Плотность двумерного нормального распределения выражается формулой (1.18):

$$f(x,y) = \frac{1}{\sigma_x \sigma_y \sqrt{2\pi}} e^{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\frac{(x-m_x)^2}{\sigma_x^2} - \frac{(x-m_x)^2}{\sigma_x \sigma_y} + \frac{(x-m_x)^2}{\sigma_y^2}\right]} \quad (1.18)$$

Многомерная совместная нормальная функция плотности является обобщением двумерного случая. Для многомерной случайной величины $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$ многомерная нормальная плотность определяется по формуле (1.19):

$$f(\xi_1, \xi_2, \ldots, \xi_m) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2\left(\vec{\xi} - \vec{\mu}\right)^T \Sigma^{-1} \left(\vec{\xi} - \vec{\mu}\right)}\right], \quad (1.19)$$

где $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$ – многомерная случайная величина, которая математически представляет собой вектор-столбец:

$$\vec{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_m \end{bmatrix} \quad (1.20)$$

$\vec{\mu}$ – вектор математических ожиданий многомерной случайной величины $\vec{\xi}$:

$$\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix} \quad (1.21)$$

$(\vec{\xi} - \vec{\mu})$ – разность двух векторов дает вектор-столбец:

$$\vec{\xi} = \begin{bmatrix} \xi_1 - \mu_1 \\ \xi_2 - \mu_2 \\ \vdots \\ \xi_m - \mu_m \end{bmatrix} \quad (1.22)$$

В результате выполнения операции транспонирования получим вектор-строку:

$$(\vec{\xi} - \vec{\mu})^T = [\xi_1 \mu_1 \ \xi_2 - \mu_2 \dots \ \xi_m - \mu_m]. \quad (1.23)$$

Σ – ковариационная матрица:

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1m}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{1m}^2 \\ \dots & \dots & \dots & \dots \\ \sigma_{m1}^2 & \sigma_{m2}^2 & \dots & \sigma_{mm}^2 \end{bmatrix}, \quad (1.24)$$

где $\sigma_{qt}^2 = M[(\xi_q - \mu_q)(\xi_t - \mu_t)]$, q=1, 2, …, m, t=1, 2, …, m,

$(\sigma_{11}^2, \sigma_{22}^2, \dots, \sigma_{mm}^2)$ – диагональные элементы, представляющие собой дисперсии признаков;

m – размерность многомерной случайной величины (количество признаков);

$|\Sigma|$ – определитель ковариационной матрицы;

Σ^{-1} – матрица, обратная к ковариационной.

Мы разработали компьютерную программу моделирования многомерной нормальной выборки. В основу программы был положен алгоритм моделирования нормального распределения, представленный в работе [1]. Рассмотрим формальное описание алгоритма.

Вектор $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_m)$ произвольной нормально распределенной случайной величины можно получить специальным линейным преобразованием вектора $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$, компоненты которого есть независимые нормально распределенные случайные величины с параметрами $\mu = 0$, $\sigma = 1$. Для моделирования одномерной нормальной случайной величины существует множество способов. Самый простой способ моделирования состоит в преобразовании двух случайных чисел α_1 и α_2:

$$\eta_1 = \sqrt{-2\ln\alpha_1}\sin 2\pi\alpha_2,\ \eta_2 = \sqrt{-2\ln\alpha_1}\cos 2\pi\alpha_2. \qquad (1.25)$$

Преобразование η в ξ производится по формуле:

$$\xi = A\eta + \mu\ . \qquad (1.26)$$

В преобразовании участвует некоторая треугольная матрица *A*:

$$A = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n1} & \ldots & a_{mm} \end{bmatrix} \qquad (1.27)$$

Коэффициенты a_{ij} могут быть определены с помощью рекуррентной процедуры. Общая рекуррентная формула имеет вид:

$$a_{qt} = \frac{\sigma_{qt}^2 - \sum_{r=1}^{t-1} a_{qr}a_{tr}}{\sqrt{\sigma_{tt}^2 - \sum_{r=1}^{t-1} a_{tr}^2}}\ , \qquad (1.28)$$

где индексы изменяются в диапазоне $1 \le t \le q \le m$, а суммы с верхним нулевым приделом равны нулю ($\sum_{r=1}^{0} a_{qr}a_{tr} = 0$, $\sum_{r=1}^{0} a_{tr}^2 = 0$).

Ключевым элементом алгоритма является вычисление матрицы преобразований *A*. Хотя алгоритм имеет строгое обоснование, общий вид расчетной формулы для исследователей-практиков носит характер справочной информации.

В простейших случаях моделирования двухмерных или даже трехмерных выборок выполнить расчеты матрицы преобразований может любой исследователь, занимающийся анализом многомерных данных. В двухмерном пространстве формулы расчета элементов матрицы преобразования принимают достаточно простой вид:

$$a_{11} = \frac{\sigma_{11}^2}{\sqrt{\sigma_{11}^2}} = \sqrt{\sigma_{11}^2} \qquad (1.29)$$

$$a_{21} = \frac{\sigma_{21}^2}{\sqrt{\sigma_{11}^2}} = \frac{\sigma_{21}^2}{a_{11}} \qquad (1.30)$$

$$a_{22} = \frac{\sigma_{22}^2 - \frac{\left(\sigma_{22}^2\right)^2}{\sigma_{11}^2}}{\sqrt{\sigma_{22}^2 - \frac{\left(\sigma_{22}^2\right)^2}{\sigma_{11}^2}}} = \sqrt{\sigma_{22}^2 - \frac{\left(\sigma_{22}^2\right)^2}{\sigma_{11}^2}} \qquad (1.31)$$

Использование приведенных формул для моделирования двухмерного нормального закона мы можем продемонстрировать на примере решения задачи моделирования в EXCEL. Пусть задача состоит в моделировании трех классов нормальных выборок с заданными параметрами. Исходные данные задачи приведены на рис. 1.42.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
13	Таблица 1							Таблица 2							Таблица 3					
14																				
15	Средние признаков по классам							Дисперсии признаков по классам							Коэффициенты корреляции					
16	КЛАСС 1		КЛАСС 2		КЛАСС 3			КЛАСС 1		КЛАСС 2		КЛАСС 3			КЛАСС 1		КЛАСС 2		КЛАСС 3	
17	$\mu_x^{(1)}$	$\mu_y^{(1)}$	$\mu_x^{(2)}$	$\mu_y^{(2)}$	$\mu_x^{(3)}$	$\mu_y^{(3)}$		$S_x^{(1)}$	$S_y^{(1)}$	$S_x^{(2)}$	$S_y^{(2)}$	$S_x^{(3)}$	$S_y^{(3)}$		$\rho_{xy}^{(1)}$		$\rho_{xy}^{(2)}$		$\rho_{xy}^{(3)}$	
18																				
19	25	5	10	15	1	20		5	8	9	5	5	15		0,4		0,5		-0,5	

Рис. 1.42. Исходные данные задачи

Первый шаг. На основании исходных данных рассчитаем корреляционные и ковариационные матрицы, а также матрицы преобразования по всем трем классам (рис. 1.43).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
22																				
23																				
24																				
25	Корреляционные матрицы по классам							Ковариационные матрицы по классам							Матрица преобразований по классам					
26	КЛАСС 1		КЛАСС 2		КЛАСС 3			КЛАСС 1		КЛАСС 2		КЛАСС 3			КЛАСС 1		КЛАСС 2		КЛАСС 3	
27	1	0,4	1	0,5	1	-0,5		5	2,53	9	3,35	5	-4,33		2,24	0,00	3,00	0,00	2,24	0,00
28	0,4	1	0,5	1	-0,5	1		2,53	8	3,35	5	-4,33	15		1,13	2,59	1,12	1,94	-1,94	3,35
29																				
30																				
31																				

=A28*КОРЕНЬ(H19*I19)

=КОРЕНЬ(H27)

=H28/O27

=КОРЕНЬ(I28-H28^2/H27)

Рис. 1.43. Первый шаг алгоритма

Второй шаг. Для моделирования шести признаков (два признака – три класса) необходимо смоделировать шесть признаков стандартного нормального распределения $\boldsymbol{n}(0,1)$. Эти признаки обозначим $\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6$ (рис. 1.44). Для моделирования воспользуемся генератором случайных чисел «Пакет анализа». Для расчета зависимых нормальных признаков X и Y по классам будем использовать матрицы преобразований (рис. 1.44).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
32														
33	**Таблица данных BOX 6**					=$O$27*B37+$A$19		=$O$28*B37+$P$28*C37+$B$19						
34														
35		Нормальный стандартный вектор						КЛАСС 1		КЛАСС 2		КЛАСС 3		
36	№	η_1	η_2	η_3	η_4	η_5	η_6	X	Y	X	Y	X	Y	
37	1	-0,3	0,66	-0,06	0,78	1,31	0,83	24,3	6,36	9,83	16,5	3,94	20,3	
38	2	-1,28	-0,34	1,22	0,02	2,33	-0,45	22,1	2,68	13,7	16,4	6,22	14	
39	3	0,24	-0,15	-1,63	-0,51	-1,36	0,14	25,5	4,9	5,11	12,2	-2,04	23,1	
40	4	1,28	-0,6	1,58	-0,75	-0,44	-0,2	27,9	4,89	14,7	15,3	0,02	20,2	
41	5	1,2	-0,15	1,47	-0,46	0,64	0,55	27,7	5,96	14,4	15,7	2,43	20,6	
42	6	1,73	-2,23	-1,41	0,13	-0,34	-1,12	28,9	1,17	5,76	13,7	0,24	16,9	
43	7	-2,18	-0,13	0,07	0,12	0,32	0,4	20,1	2,18	10,2	15,3	1,72	20,7	
44	8	-0,23	-1,33	-0,42	-0,55	0,99	-0,83	24,5	1,29	8,73	13,5	3,21	15,3	
131	95	1,24	-0,01	-1,52	-0,91	0,54	-0,73	27,8	6,38	5,43	11,5	2,21	16,5	
132	96	-0,31	0,16	1,08	-0,61	-0,97	-0,35	24,3	5,07	13,2	15	-1,16	20,7	
133	97	-0,84	-1,88	0,2	3,12	-0,52	0	23,1	-0,83	10,6	21,3	-0,17	21	
134	98	-0,82	-0,26	-0,96	0,27	-1,82	-0,8	23,2	3,41	7,12	14,4	-3,06	20,8	
135	99	-0,43	-0,94	-0,14	0,17	1,16	-1,36	24	2,08	9,57	15,2	3,6	13,2	
136	100	-0,45	-1	-1,17	-0,3	-0,36	0,64	24	1,9	6,49	13,1	0,19	22,9	
137	Средние	-0,04	-0,03	0,05	0,00	-0,06	0,05	24,9	4,86	10,1	15	0,86	20,3	
138	Дисперс	1,17	1,00	1,12	0,98	1,14	1,01	5,83	8,61	10,1	4,94	5,69	17,8	
139	Корреляция							0,47		0,51		-0,62		

Рис. 1.44. Таблица модельных данных

По данным классифицированной модельной выборки, представленной в таблице данных, построим диаграмму рассеивания (рис. 1.45). Изменяя исходные данные, задающие параметры классов на диаграмме рассеивания, можно наблюдать изменения образов классов.

Для моделирования нормальных смесей с произвольным количеством классов мы разработали специальный программный модуль. Программный модуль был реализован в виде дополнительной надстройки к EXCEL как самому распространенному среди практиков пакету по обработке данных. Тем более, что данные из пакета EXCEL без труда могут быть экспортированы в любой специализированный пакет по обработке статистических данных.

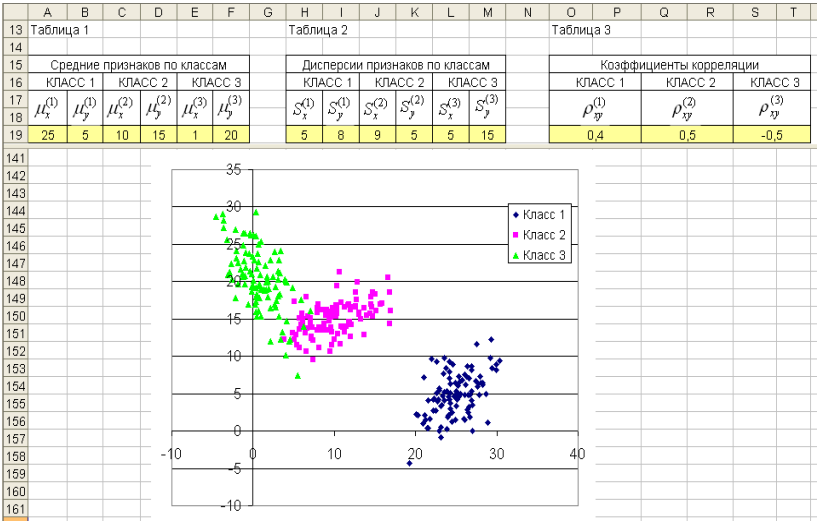


Таблица 1						Таблица 2						Таблица 3		
Средние признаков по классам						Дисперсии признаков по классам						Коэффициенты корреляции		
КЛАСС 1		КЛАСС 2		КЛАСС 3		КЛАСС 1		КЛАСС 2		КЛАСС 3		КЛАСС 1	КЛАСС 2	КЛАСС 3
$\mu_x^{(1)}$	$\mu_y^{(1)}$	$\mu_x^{(2)}$	$\mu_y^{(2)}$	$\mu_x^{(3)}$	$\mu_y^{(3)}$	$S_x^{(1)}$	$S_y^{(1)}$	$S_x^{(2)}$	$S_y^{(2)}$	$S_x^{(3)}$	$S_y^{(3)}$	$\rho_{xy}^{(1)}$	$\rho_{xy}^{(2)}$	$\rho_{xy}^{(3)}$
25	5	10	15	1	20	5	8	9	5	5	15	0,4	0,5	-0,5

Рис. 1.45. Диаграмма рассеивания модельных данных

Таким образом, мы реализовали в EXCEL программу, моделирующую смеси двухмерных нормальных распределений. Такие программы называются генераторами случайных чисел или симуляторами.

Программный модуль устанавливается в EXCEL после запуска специальной установочной программы. При установке программы необходимо установить низкий уровень безопасности (меню Сервис-Макрос-Безопасность). После установки программы на нижней панели EXCEL появится значок вызова программы моделирования (рис. 1.46)

Рис. 1.46. Кнопка вызова программы моделирования нормальных смесей

Рассмотрим принцип работы программы. Для демонстрации программы будем использовать те же данные (три класса – два признака).

	A	B	C	D	E	F	G	H	I
2	КЛАСС 1								
3									
4	Средние признаков			Дисперсии признаков			Корреляционная матрица		
5	25			5			1	0,4	
6	5			8			0,4	1	
7									
8	КЛАСС 2								
9									
10	Средние признаков			Дисперсии признаков			Корреляционная матрица		
11	10			9			1	0,5	
12	15			5			0,5	1	
13									
14	КЛАСС 3								
15									
16	Средние признаков			Дисперсии признаков			Корреляционная матрица		
17	1			5			1	-0,5	
18	20			15			-0,5	1	

Рис. 1.47. Форма представления данных для моделирования трех классов

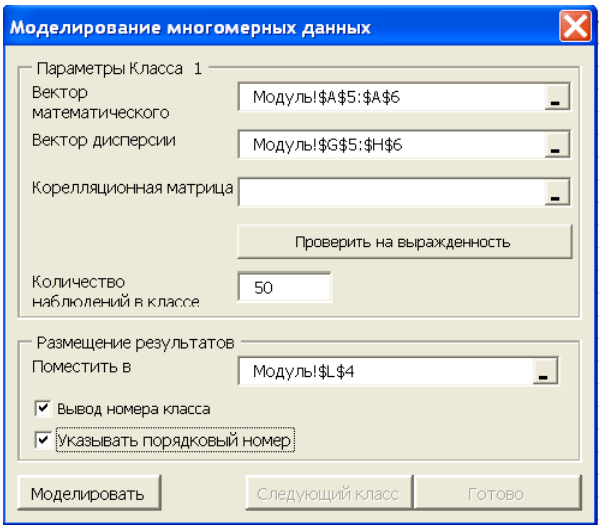


Рис. 1.48. Параметры программы моделирования нормальных смесей

При моделировании данных, размерность которых свыше трех, необходимо задавать такие корреляционные матрицы, у которых ковариационная матрица положительно определена. В противном случае такие данные не могут быть смоделированы. Поэтому в процессе работы программы определяется это условие и выдается

сообщение о его выполнении. Это условие следует из того, что зависимость двух признаков отражается и на зависимости их с третьим признаком.

Рассмотрим пример моделирования трех многомерных нормальных выборок с тремя зависимыми признаками. Некоторые классы составляют совокупность объектов выборок. При запуске программы для трех классов были установлены следующие объемы выборок $N_1 = 200, N_2 = 100, N_3 = 300$. Остальные параметры многомерной выборки представлены в табл. 1.1. Различия в степени зависимости признаков задаются коэффициентами корреляционной матрицы. Как видно из табл. 1.1, степень зависимости между признаками первого класса гораздо ниже, чем между признаками во втором и третьем классах. В силу того, что при моделировании используется датчик случайных чисел, параметры выборок, рассчитанные по модельным данным (табл. 1.2), будут несколько отличаться от заданных параметров (табл. 1.1).

Приведем расчетные данные по матрице преобразований A для 1-, 2- и 3-го классов.

$$A_1 = \begin{Vmatrix} 1{,}000000 & 0 & 0 \\ 0{,}300000 & 0{,}953939 & 0 \\ 0{,}200000 & 0{,}041931 & 0{,}978898 \end{Vmatrix} \quad (1.32)$$

$$A_2 = \begin{Vmatrix} 1{,}414214 & 0 & 0 \\ 0{,}707107 & 1{,}224745 & 0 \\ 0{,}707107 & 0{,}408248 & 1{,}154701 \end{Vmatrix} \quad (1.33)$$

$$A_3 = \begin{Vmatrix} 1{,}732051 & 0 & 0 \\ 1{,}558846 & 0{,}754983 & 0 \\ 1{,}385641 & 0{,}715247 & 0{,}753937 \end{Vmatrix} \quad (1.34)$$

Таблица 1.1

Параметры, заданные при моделировании многомерной нормальной выборки

№ класса	Признак	Среднее	Дисперсия	Корреляционная матрица		
Класс 1	Признак 1	5	1	1,0	0,3	0,2
	Признак 2	5	1	0,3	1,0	0,1
	Признак 3	4	1	0,2	0,1	1,0

Окончание табл. 1.1

№ класса	Признак	Среднее	Дисперсия	Корреляционная матрица		
Класс 2	Признак 1	10	2	1,0	0,5	0,5
	Признак 2	10	2	0,5	1,0	0,5
	Признак 3	10	2	0,5	0,5	1,0
Класс 3	Признак 1	8	3	1,0	0,9	0,8
	Признак 2	8	3	0,9	1,0	0,9
	Признак 3	2	3	0,8	0,9	1,0

Таблица 1.2

Параметры, рассчитанные по данным многомерной модельной выборки

№ класса	Признак	Среднее	Дисперсия	Корреляционная матрица		
Класс 1	Признак 1	5,02	1,16	1,00	0,25	0,27
	Признак 2	5,00	1,12	0,25	1,00	0,14
	Признак 3	3,95	0,83	0,27	0,14	1,00
Класс 2	Признак 1	9,89	1,53	1,00	0,37	0,47
	Признак 2	9,85	1,76	0,37	1,00	0,53
	Признак 3	10,08	1,83	0,47	0,53	1,00
Класс 3	Признак 1	8,02	3,08	1,00	0,92	0,80
	Признак 2	8,05	2,99	0,92	1,00	0,90
	Признак 3	2,10	3,06	0,80	0,90	1,00

Графический образ, сгенерированных данных можно проанализировать на трехмерном графике (рис. 1.49) и на трех двухмерных графиках (рис. 1.50).

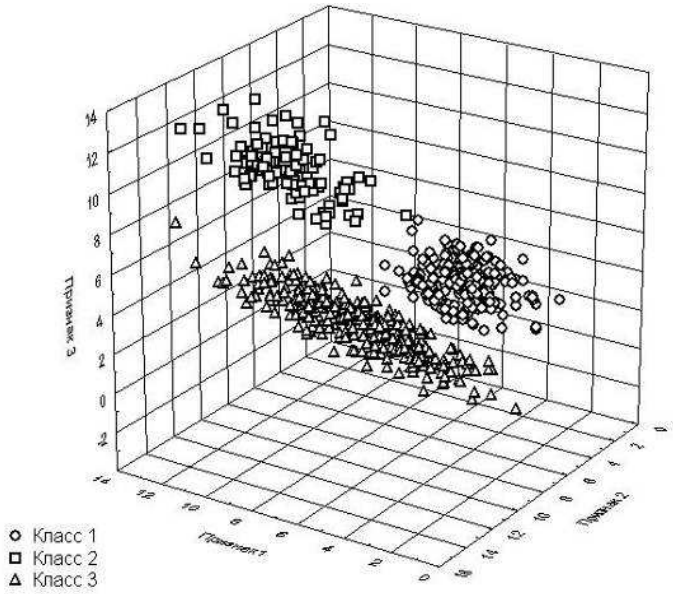


Рис. 1.49. Трехмерная диаграмма рассеивания смоделированной многомерной выборки

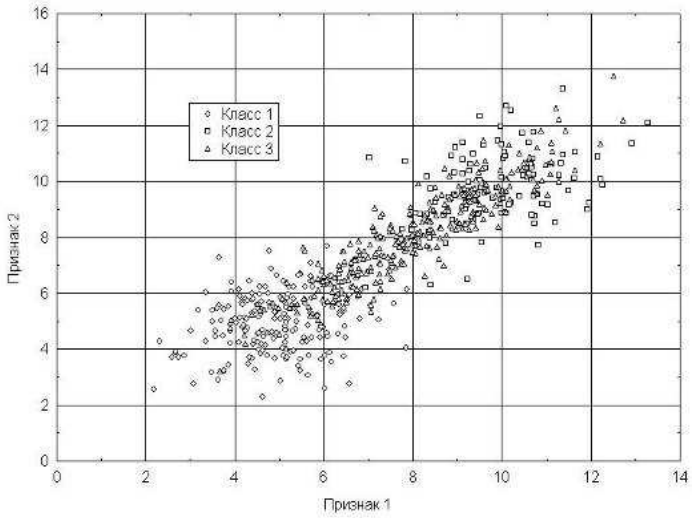


Рис. 1.50. Проекция диаграммы рассеивания многомерной выборки на плоскость с осями «Признак 1» – «Признак 2»

Из диаграмм рассеивания многомерной выборки следует, что классы визуально хорошо различимы в пространстве трех признаков. Различия классов проявляются и в проекции выборок на плоскость с осями «Признак 1» – «Признак 3» и в проекции на плоскость с осями «Признак 2» – «Признак 3». Однако классы достаточно плохо визуально различимы на плоскости в координатах «Признак 1» – «Признак 2» (рис. 1.50).

1.4. Предварительные этапы процесса анализа данных

Можно выделить следующие начальные этапы процесса анализа данных:

– анализ предметной области;
– постановка задачи;
– подготовка данных;
– группировка данных;
– проверка статистических гипотез;
– поиск и анализ зависимостей признаков.

Этап 1. Анализ предметной области

Процесс исследования определенной предметной области, объекта или явления заключается в наблюдении свойств объектов с целью выявления и оценки важных, с точки зрения субъекта-исследователя, закономерных отношений между показателями данных свойств.

Предметная область – это мысленно ограниченная область реальной действительности, подлежащая описанию или моделированию и исследованию; состоит из объектов, различаемых по свойствам и находящихся в определенных отношениях между собой или взаимодействующих каким-либо образом.

Предметная область – часть реального мира, она бесконечна и содержит с точки зрения проводимого исследования как существенные, так и незначащие данные. Исследователю необходимо уметь выделить существенную их часть. В процессе изучения предметной области должна быть создана ее модель. Знания из различных источников должны быть формализированы при помощи каких-либо средств.

Этап 2. Постановка задачи

Постановка задачи включает следующие шаги:

– формулировка задачи;

– формализация задачи.

Постановка задачи также предполагает описание статического и динамического поведения исследуемых объектов.

Описание статики подразумевает описание объектов и их свойств. При описании динамики описывается поведение объектов и те причины, которые влияют на их поведение.

На этом этапе необходимо в общих чертах сформулировать список задач, которые предполагается решать по данным, описывающим предметную область. По мере решения задач список задач может уточняться.

Этап 3. Подготовка данных

Цель этапа: разработка базы данных, которая будет использоваться для решения спектра возможных задач.

Подготовка данных является важнейшим этапом, от качества выполнения которого зависит возможность получения качественных результатов всего процесса анализа данных. Кроме того, следует помнить, что на этап подготовки данных, по некоторым оценкам, может быть потрачено до 80% всего времени, отведенного на проект.

На этом этапе решаются следующие задачи:

1. Определение и анализ требований к данным.

Здесь осуществляется так называемое моделирование данных, т.е. определение и анализ требований к данным, которые необходимы для осуществления анализа данных.

2. Сбор данных.

На этом этапе определяется процедура сбора и хранения данных, необходимое количество данных (по количеству исследуемых признаков и объектов); осуществляется кодирование некоторых данных. Набор данных должен быть репрезентативным и представлять как можно больше возможных ситуаций.

3. Предварительная обработка данных.

Производится оценивание качества данных и по возможности повышение качества данных. Такие данные обеспечивают полу-

чение качественного результата – знаний, которые смогут поддерживать процесс принятия решений.

Качество данных (*data quality*) – критерий, определяющий полноту, точность, своевременность и возможность интерпретации данных.

Данные низкого качества иначе называют грязными.

Рассмотрим более детально понятия качества данных.

Грязные данные могут появиться по разным причинам, таким, как ошибка при вводе данных, использование иных форматов представления или единиц измерения, несоответствие стандартам, отсутствие своевременного обновления, неудачное обновление всех копий данных, неудачное удаление записей-дубликатов и т.д.

Можно выделить следующие группы грязных данных:

– грязные данные, которые могут быть автоматически обнаружены и очищены;

– данные, появление которых может быть предотвращено;

– непригодные для автоматического обнаружения и очистки;

– данные, появление которых невозможно предотвратить.

Специальные средства очистки могут справиться не со всеми видами грязных данных.

Рассмотрим наиболее распространенные виды грязных данных:

– пропущенные значения;

– дубликаты данных;

– шумы и выбросы.

Распространенным источником данных о состоянии социально-экономических систем являются анкетные опросы. Классификация источников ошибок, возникающих при сборе анкетных данных, представлена на рис. 1.51.

Визуализация данных позволяет представить их, в том числе и выбросы, в графическом виде.

Этап 4. Группировка данных

Группировка – это разбиение совокупности на группы, однородные по какому-либо признаку. С точки зрения отдельных единиц совокупности группировка – это объединение отдельных единиц совокупности в группы, однородные по каким-либо признакам.

Устойчивое разграничение объектов выражается классификацией, которая основывается на самых существенных признаках

(например, классификация отраслей народного хозяйства, основных фондов и т.д.). Таким образом, классификация – общепринятая, нормативная группировка.

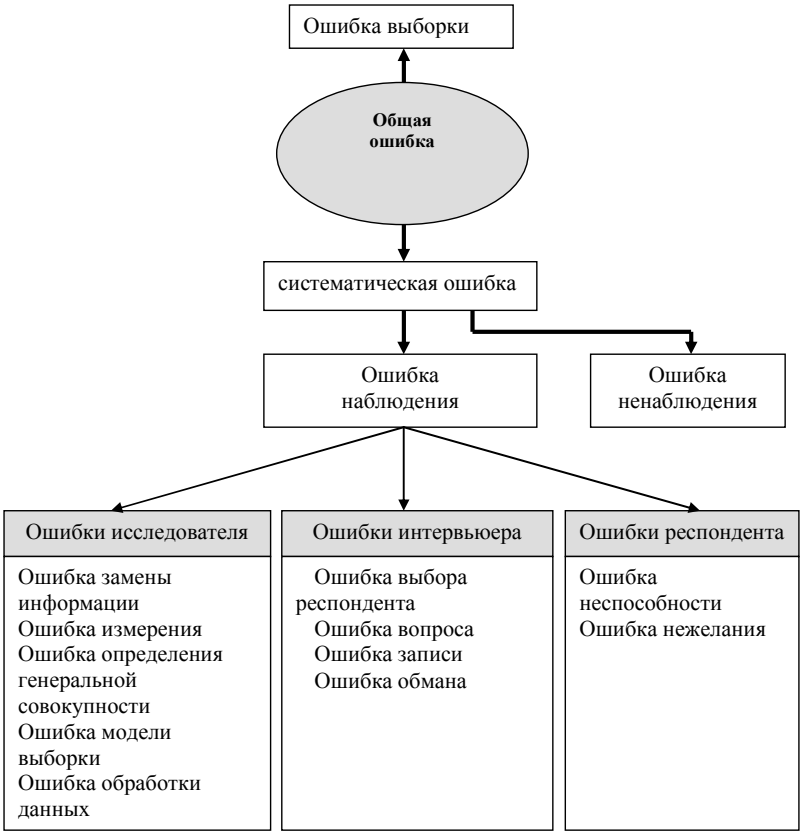


Рис. 1.51. Источники ошибок при проведении анкетного опроса

Метод группировки основывается на следующих категориях: группировочный признак, интервал группировки и число групп.

Группировочный признак – признак, по которому происходит объединение отдельных единиц совокупности в однородные группы.

Интервал очерчивает количественные границы групп. Как правило, он представляет собой промежуток между максимальными и минимальными значениями признака в группе. Интервалы бывают:

– *равные*, когда разность между максимальным и минимальным значениями в каждом из интервалов одинакова;

– *неравные*, когда, например, ширина интервала постепенно увеличивается, а верхний интервал часто не закрывается вовсе;

– *открытые*, когда имеется только либо верхняя, либо нижняя граница;

– *закрытые*, когда имеются и нижняя, и верхняя границы.

При проведении группировки приходится решать ряд задач:

1) выделение группировочного признака;

2) определение числа групп и величины интервалов;

3) при наличии нескольких группировочных признаков описание того, как они комбинируются между собой;

4) установление показателей, которыми должны характеризоваться группы, т.е. сказуемого группировки.

Определение числа групп. Здесь необходимо учитывать несколько условий:

а) число групп детерминируется уровнем изменчивости группировочного признака. Чем значительнее вариация признака, тем больше при прочих равных условиях должно быть групп;

б) число групп должно отражать реальную структуру изучаемой совокупности;

в) не допускается выделение пустых групп. Если проблема пустых групп все же возникает, при проведении структурных группировок используют неравные интервалы.

Статистические группировки и классификации преследуют цели выделения качественно однородных совокупностей, изучения структуры совокупности, исследования существующих зависимостей. Каждой из этих целей соответствует особый вид группировки: типологическая, структурная, аналитическая (факторная).

Типологическая группировка решает задачу выявления и характеристики социально-экономических типов (частных подсовокупностей).

Структурная дает возможность описать составные части совокупности или строение типов, а также проанализировать структурные сдвиги.

Аналитическая (факторная) группировка позволяет оценивать связи между взаимодействующими признаками.

В зависимости от числа положенных в их основание признаков различают простые и многомерные группировки.

Группировка, выполненная по одному признаку, называется простой.

Многомерная группировка производится по двум и более признакам. Частным случаем многомерной группировки является *комбинационная группировка*, базирующаяся на двух и более признаках, взятых во взаимосвязи, в комбинации.

Структурная группировка применяется для характеристики структуры совокупности и структуры сдвигов.

Структурной называется группировка, в которой происходит разделение выделенных с помощью технологической группировки типов явлений, однородных совокупностей на группы, характеризующие их структуру по какому-либо варьирующему признаку. Например, группировка населения по размеру среднедушевого дохода. Анализ структурных группировок, взятых за ряд периодов или моментов времени, показывает изменения структуры изучаемых явлений, то есть структурные сдвиги. В изменении структуры общественных явлений отражаются важнейшие закономерности их развития.

Показатель численности групп представлен либо частотой (количеством единиц в каждой группе), либо частотностью (удельным весом каждой группы).

Среди простых группировок особо отметим ряды распределения, т.е. группировки, в которых для характеристики групп (упорядоченно расположенных по значению признака) применяется один показатель – численность группы. Другими словами, это ряд чисел, показывающий, как распределяются единицы некоторой совокупности по изучаемому признаку.

Этап 5. Проверка статистических гипотез

Статистическая гипотеза – это предположение (суждение) о генеральной совокупности – ее распределении или параметрах, подвергаемое проверке по выборке, в результате которой она прини-

мается или отвергается. Формулировка гипотезы должна исходить из возможности использования известных законов распределения.

Сущность проверки статистических гипотез заключается в том, чтобы установить, согласуются или нет данные наблюдений и выдвинутая гипотеза. Можно ли расхождение между гипотезой и результатами выборочных наблюдений отнести за счет случайной погрешности, обусловленной механизмом случайного отбора?

На этом этапе важно исследовать однородность выборки.

Критерии однородности предназначены для проверки нулевой гипотезы о том, что две выборки (или несколько) взяты из одного распределения либо их распределения имеют одинаковые значения математического ожидания, дисперсии или других параметров.

Параметрические критерии предполагают, что выборка порождена распределением из заданного параметрического семейства. В частности, существует много критериев, предназначенных для анализа выборок из нормального распределения. Преимущество этих критериев в том, что они более мощные. Если выборка действительно удовлетворяет дополнительным предположениям, то параметрические критерии дают более точные результаты. Однако если выборка им не удовлетворяет, то вероятность ошибок (как I, так и II рода) может резко возрасти. Прежде чем применять такие критерии, необходимо убедиться, что выборка удовлетворяет дополнительным предположениям. Гипотезы о виде распределения проверяются с помощью критериев согласия.

Непараметрические критерии не опираются на дополнительные предположения о распределении. В частности, к этому типу критериев относится большинство ранговых критериев.

Этап 6. Поиск и анализ зависимостей признаков

Для анализа зависимостей используются следующие группы статистических методов:

– корреляционный анализ;
– регрессионный анализ;
– дисперсионный анализ;
– факторный анализ.

Корреляционный анализ – группа статистических методов, направленная на выявление и математическое представление структурных зависимостей между выборками.

Регрессионный анализ – набор статистических методов исследования влияния одной или нескольких независимых переменных $X_1, X_2, \ldots, X_j, \ldots, X_m$ на зависимую переменную Y.

Дисперсионный анализ применяется для исследования влияния одной или нескольких качественных переменных (факторов) на одну зависимую количественную переменную (отклик).

Факторный анализ позволяет находить в многомерном пространстве первичные переменные (значения которых регистрируются в эксперименте), сокращенную систему вторичных переменных (факторов).

Анализ данных – необходимый этап разработки управленческих решений.

Анализ – это метод исследования, который заключается в изучении отдельных свойств, сторон и составных частей некоего целого.

Для разработки управленческих решений управления социально-экономическими системами необходимо представлять себе ситуацию в целом. Для этого служат модели принятия решений. Чем более точно решается задача анализа, тем больше возможностей для построения эффективных моделей принятия решений.

Глава 2. ПРАКТИЧЕСКИЕ ЗАДАЧИ АНАЛИЗА ДАННЫХ

2.1. Методики и особенности сбора информации в сети Интернет

2.1.1. Инструментальные средства сбора анкетных данных в сети Интернет

При исследовании социально-экономических процессов все шире начинают применяться анкетные опросы. Со временем анкетные формы совершенствуются, включают большое разнообразие типов вопросов. Включение разнообразных типов вопросов расширяет возможности исследователя по анализу ситуации и выработке управленческих решений. На этапе составления анкеты исследователь должен учитывать множество факторов: от возможности убедить респондентов в представлении достоверной информации до оценки своих возможностей по обработке собранных данных – спектром методов и технологий, которыми владеет исследователь.

Стремление к всестороннему анализу ситуации приводит к тому, что часто анкеты включают достаточно большое количество вопросов. Использование многомерных методов анализа данных предъявляет особые требования к объему выборок. Все чаще ученые в своих исследованиях проводят не разовые акции по сбору данных, а повторяют сбор с определенной периодичностью (мониторинг ситуации). Поэтому исследователи сталкиваются с большой проблемой сбора первичного материала, а затем и переноса его на машинный носитель для дальнейшей обработки с использованием разнообразных программных средств анализа данных. Этап сбора информации во многом определяет качество результатов, которых возможно добиться после обработки данных. Свои возможности по сбору данных исследователь всегда должен соизмерять с доступными ему временными и финансовыми ресурсами.

Использование сложных статистических методов обработки данных требует переноса данных в среду, в которой возможно использование инструментов, реализующих такие методы в виде специальных программных модулей. Расширить возможности исследователей может сочетание различных методик сбора, хранения и передачи данных. Иначе говоря, актуальной проблемой является разработка программных средств, обеспечивающих согласование различных технологий сбора и обработки данных. В настоящей работе рассматривается программный модуль, позволяющий согласовать различные методики сбора данных и объединять их в единую базу данных.

В последнее время все большее распространение среди российских исследователей начинают получать сервисы онлайн-опросов. Поэтому в работе, посвященной автоматизации сбора анкетных данных, целесообразно специально выделить эту технологию.

Рассмотрим современные средства поддержки интернет-опросов.

Понимание значимости и очень высокой трудоемкости этапа сбора данных способствовало появлению множества программных средств по автоматизации сбора анкетных данных в сети Интернет (сервисов для проведения онлайн-опросов) [2–4]. Технология сбора данных посредством самостоятельного заполнения интервьюером анкет в Интернете в мировой практике известна как технология-CAWI (***computer-assisted web interviewing***). Различные программные средства отличаются набором инструментальных средств, доступных исследователю для составления анкетных форм, распространения анкет в сети, представления и обработки данных. В России широкое использование онлайн-опросов началось гораздо позже, чем в западных странах. Особенно возросло количество анкетных опросов в последние несколько лет. В настоящее время существует уже немало отечественных публикаций, в которых рассматриваются современные тенденции проведения онлайн-исследований [5, 6]. Онлайн-опросы с большим успехом применяются в исследованиях общественного мнения. Большие перспективы использования технологии онлайн-опросов населения имеют для установления обратной связи органов управления и населения. С их помощью можно оперативно оценивать качество обслуживания населения в различных сферах.

Известные программные средства автоматизации сбора данных в Интернете во многих случаях облегчают работу исследователей. По отношению к традиционной методике сбора анкетных данных на бумажном носителе использование интернет-сервисов обеспечивает исследователю ряд преимуществ. CAWI-технология не требует привлечения специальных сотрудников по сбору данных (интервьюеров) и технических работников по переносу данных из бумажного носителя в компьютерное представление. Однако CAWI-технология имеет и свои недостатки. В последние годы появились публикации, посвященные исследованию угроз качеству и надежности данных, полученных с помощью онлайн-опросов [7–9].

В качестве примера наиболее распространенных интернет-сервисов можно привести следующие: Google формы, Survio.com, SurveyMonkey.ru, Testograf.ru, Simpoll.ru, Webanketa, LimeSurvey. Тем не менее, в онлайн-сервисах представлен ограниченный набор средств анализа данных, который больше пригоден для предварительного анализа. В случае решения сложных задач чаще всего используется многомерный анализ данных, который не входит в число инструментов интернет-сервисов. Поэтому данные, собранные в сети, чаще всего приходится экспортировать в среды, обладающие мощными инструментами обработки. Наиболее распространенной средой обработки многомерных данных служит EXCEL.

С увеличением количества исследований, основанных на онлайн-опросах, потребность в программных средствах, позволяющих осуществлять согласование различных технологий, будет только возрастать. Онлайн-сервисы сделали анкетные опросы более доступными для исследователей, обладающих различным уровнем компьютерной грамотности. Большинство из них не являются специалистами в области разработки программных средств, но по мере накопления опыта работы по сбору данных онлайн у них, естественно, появится потребность применения различных технологий обработки данных, согласования данных, собранных всеми доступными им способами.

В настоящей работе предлагаются к рассмотрению программные средства, позволяющие осуществлять объединение данных, собранных различными способами, в базу данных. Другими словами, целью работы является повышение эффективности науч-

ных исследований, основанных на использовании опросов населения.

В качестве среды объединения данных выбрана среда EXCEL, которая в настоящее время является одним из наиболее распространенных приложений в мире. Кроме того, что в среде EXCEL разработано достаточно много собственных средств анализа данных, данные из EXCEL легко экспортируются в другие системы анализа данных.

Предлагаем рассмотреть программное средство, относящееся к классу систем автоматизации научных исследований. Данная программа разработана на основании обобщения опыта проведения анкетных опросов по исследованию социально-экономических процессов во Владивостокском государственном университете экономики и сервиса.

Использование анкетных интернет-форм характеризуется рядом недостатков, среди которых ограниченные возможности влияния на формирование аудитории респондентов, сложность стимулирования респондентов к заполнению анкетных форм. Отсюда наряду с использованием интернет-анкетирования часто возникает необходимость в применении других форм организации опросов.

Вообще говоря, для исследователя наибольший интерес представляет этап обработки данных, на котором и формируются результаты исследования, вырабатываются выводы и предложения по принятию управленческих решений. Для обработки данных используются инструментальные средства в виде компьютерных программ, реализующих те или иные методы обработки данных. Другими словами, работа по анализу данных начинается тогда, когда в распоряжении исследователя появляется компьютерное представление данных анкетного опроса. Для того чтобы получить данные, отвечающие требованиям исследователя, он, как правило, сам участвует в организации системы сбора данных и подготовки данных. Вместе этап сбора и подготовки данных можно определить как подготовительный этап работы (рис. 2.1). Для определенности будем считать, что подготовительный этап заканчивается моментом, с которого исследователь может приступить к обработке и анализу данных на компьютере.

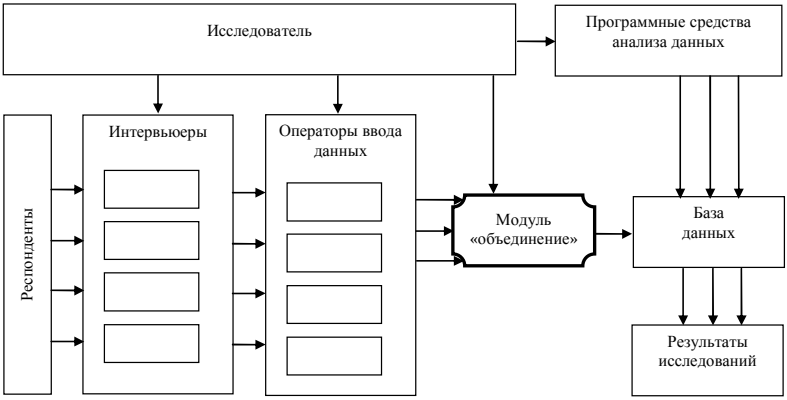


Рис. 2.1. Обобщенная схема обработки анкетных данных

Различные формы организации опроса сопряжены со своими сложностями (проблемами), которые исследователь должен учитывать при выборе той или иной формы организации работы на подготовительном этапе. Можно выделить следующие основные факторы, определяющие эффективность подготовительного этапа работы: длительность подготовительного этапа работы, качество информации и стоимость организации работы. Эти факторы взаимозависимы. Принимая решение об организации работ на подготовительном этапе, исследователь вынужден искать компромисс между желаемым и возможным.

Он, как правило, не в состоянии самостоятельно выполнить все работы на этапе сбора информации, поэтому вынужден для выполнения отдельных видов работы привлекать исполнителей. Такая работа носит эпизодический характер и не требует очень высокой квалификации, часто ее выполняют студенты. Таким образом, исследователь, как правило, на этапе сбора данных налаживает взаимодействие с группой исполнителей, которая может быть достаточно большой. При этом взаимодействие должно быть четко оговорено процедурой и осуществляться в оперативном порядке. Программное средство, предлагаемое к рассмотрению в настоящей работе, позволяет автоматизировать труд исследователя (руководителя проекта) на этапе сбора данных при взаимодействии с коллективом исполнителей, участвующих в работе. Конкретные функции

программы были выработаны, исходя из обобщения практического опыта работы с использованием различных способов организации сбора данных.

Чтобы дать представление о возможностях разработанного программного средства, рассмотрим различные сценарии организации системы сбора анкетных данных, в которых может быть использовано данное программное средство.

Сценарий 1

В этом случае респондент самостоятельно вводит данные в компьютерную форму, заранее подготовленную в определенном программном продукте. Как частный случай, в качестве формы может выступать и форма, подготовленная с помощью внешних интернет-сервисов создания анкет. Формы ввода анкетных данных могут быть созданы исследователем с использованием инструментов Access или Excel. При разработке анкетных форм в этих программных продуктах можно создать более совершенные формы, чем с помощью типовых инструментов конструкторов форм интернет-опросов. Отличие в возможностях будет примерно такое же, как при строительстве зданий из типовых блоков и по индивидуальному проекту.

Файл с разработанной формой в формате Access или Excel может быть передан респондентам самим исследователем. В роли респондентов могут выступать аудитория студентов или группа экспертов. При этом исследователь сам участвует в формировании группы респондентов и инструктирует их по правилам заполнения формы. В такой схеме роль интервьюера выполняет сам исследователь, а в роли оператора выступает респондент. В результате такого опроса исследователь получает множество файлов определенного формата. Такие файлы всегда можно собрать в одну папку на компьютере, используя при этом корпоративную сеть учреждения (например, университета). Файлы можно пересылать и по электронной почте. При большом количестве собираемых анкет работа по объединению файлов в единую базу данных может потребовать от исследователя много непроизводительных затрат времени.

Подобная схема может быть реализована множеством различных способов, в зависимости от особенностей исследуемого явления или процесса и возможностей самого исследователя.

Во всех случаях в результате будем иметь множество единообразных файлов, которые требуется интегрировать в единую базу.

Сценарий 2

Отличие этого сценария состоит в том, что исследователь частично делегирует свои функции группе интервьюеров, которые организуют сбор данных в электронном виде. В своей работе интервьюеры могут использовать анкетирование на бумажном носителе. Иногда необходимо иметь дубликат всех анкет на бумажном носителе. В результате использования такого сценария в распоряжении исследователя оказывается группа файлов, содержащих несколько записей определенного формата, которые исследователю необходимо объединить в общую базу данных.

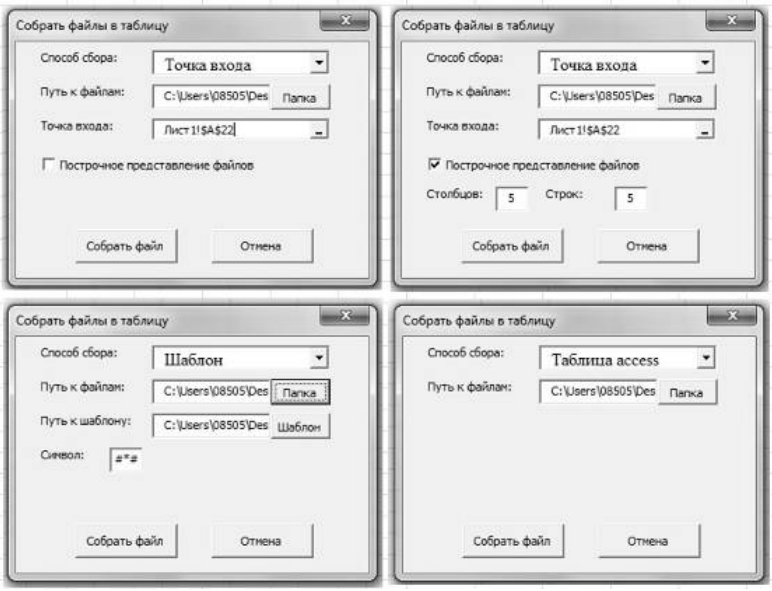


Рис. 2.2. Интерфейс программного модуля «Сбор файлов»

В результате анализа различных форм представления анкетных данных был разработан программный модуль в среде Excel, который позволяет осуществлять «сборку» различных вариантов оформления первичных данных в единую базу данных. Для объе-

динения данных могут быть использованы четыре режима работы программы:

– точка входа в таблице Excel;

– точка входа в таблице Excel с построчным представлением данных в результирующей базе данных;

– сборка данных таблиц Excel по заданному шаблону;

– сборка данных, полученных в результате заполнения форм Access.

На рисунке 2.2 представлен интерфейс обращения к программе в различных режимах. На рисунке 2.3 отражены примеры файлов, которые необходимо объединить, и результат работы программы в различных режимах.

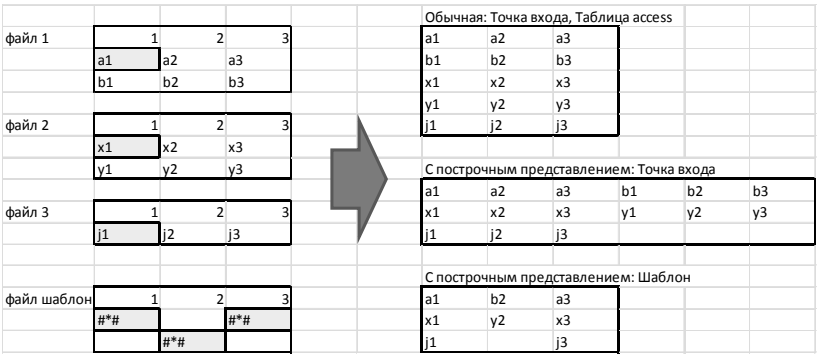


Рис. 2.3. Примеры вариантов сбора данных в единую базу в различных режимах работы программы

Разработанное программное обеспечение обладает двумя важными качествами: простота использования и универсальность.

Программное обеспечение прошло апробацию при организации подготовки данных в ряде анкетных опросов. Оно позволяет исследователю одновременно использовать несколько схем сбора информации. Совмещение этапа сбора информации и ввода данных в компьютер существенно сокращает длительность подготовительного этапа и позволяет исследователю в сжатые сроки приступить к содержательному анализу данных. Оперативность получения информации несет еще ряд преимуществ. Получение

информации с минимальной задержкой позволяет вовремя внести изменения в анкетную форму уже в начале сбора информации, что способствует улучшению качества собираемой информации.

Оперативный анализ данных, поступающих от различных интервьюеров, позволяет своевременно отреагировать на некачественную работу отдельных интервьюеров и принять меры по улучшению их работы. Сочетание различных организационных форм сбора анкетных данных дает возможность применять технологию «конструирования» выборки, добиваясь более высокой репрезентативности выборки.

Автоматизация сбора информации способствует развитию компьютерных технологий, которые могут быть использованы для выполнения гражданских онлайн-экспертиз по привлечению населения к выработке управленческих решений.

2.1.2. Инструментальные средства сбора информации по интернет-сайтам

Весь Интернет, по сути, состоит из информации: библиотеки и архивы, телеконференции и электронные газеты, новейшие разработки в области науки и советы народной медицины и т.п. Со временем в сети накапливается все больше мусора, что затрудняет поиск действительно полезной информации. Для автоматизации сбора информации был разработан специальный инструмент **«парсинг».**

Инструменты web scraping (парсинг) разработаны для извлечения, сбора любой открытой информации с веб-сайтов. Эти ресурсы нужны тогда, когда необходимо быстро получить и сохранить в структурированном виде любые данные из Интернета.

С помощью парсинга можно собирать следующую информацию:

– для исследования рынка: товары из каталогов, объемы продаж, цены и динамику цен;

– профили пользователей, зарегистрированных в социальных сетях;

– контактную информацию;

– данные пользовательской активности на сайте (комментарии, лайки, репосты);

– сбор аудитории для рекламной кампании.

Этапы парсинга:

– поиск данных. В программу-парсер загружается исходный HTML-код страницы сайта. С кодом начинает работать скрипт, который разбивает весь текст на лексемы, выделяя необходимую информацию;

– извлечение информации. Поиск данных происходит благодаря определенному набору знаков, описывающих цель поиска. Этот набор также называется регулярными выражениями. Они позволяют выделить из всего массива только интересующие фрагменты;

– сохранение данных. После получения информация сохраняется в виде таблиц или вносится в базу данных.

Для сбора информации в сети используются специальные инструментальные средства (программы). Среди наиболее распространенных программ можно назвать: Import.io, Webhose.io, Dexi.io, Scrapinghub, ParseHub, VisualScraper, Scraper.

2.2. Методы повышения достоверности данных

2.2.1. Методы восстановления пропусков в данных, представленных в различных измерительных шкалах

Большинство ученых, исследующих социально-экономические процессы, сталкиваются с проблемой пропуска данных или неответа в таблицах объект-свойство [10]. Иначе эту проблему называют проблемой неполноты данных [11]. Часто выбросы тоже можно рассматривать как пропущенные данные. К выбросам относят данные, которые явно противоречат данным всей выборки. Причем противоречие может возникать не только со значениями одного признака, но и со значениям прочих признаков одного наблюдения. В обоих случаях перед исследователем стоит дилемма: либо отбросить всю строку таблицы данных, либо каким-то образом исправить ошибку (восстановить данные). Часть противоречий (ошибок) может быть выявлена и исправлена на предварительных этапах анализа данных путем логического анализа противоречий в многомерных данных. Для этого можно использовать специальные средства [12].

При большом количестве исследуемых признаков количество пропусков может быть значительным. Часто отбрасывать данные нежелательно по той причине, что на основании многомерных данных решается множество задач, в которых используются либо од-

номерные признаки (частотные ряды), либо часть признаков многомерных наблюдений. В одной задаче все признаки задействуются крайне редко. Если говорить об анкетных данных, то анкеты могут включать много вопросов, которые служат для классификации данных (например, данные по социально-демографическому портрету респондентов), а, следовательно, задействованы при решении определенного круга задач.

Многообразие ситуаций и причин возникновения пропусков в данных породило обилие исследований в этой области. Особенно много работ, посвященных исследованию данной проблемы, в зарубежных источниках. Обширный список таких работ можно найти в отдельных работах отечественных ученых [13, 14]. Большое количество методов потребовало систематизации подходов и разработки классификации методов [15, 16]. Большинство авторов за основу принимают схему классификации, представленную в работах [17, 18] (рис. 2.4), в которых приводятся основные принципы распространенных методов восстановления данных. Можно отметить, что новые разрабатываемые методы, как правило, вписываются в представленную схему классификации.

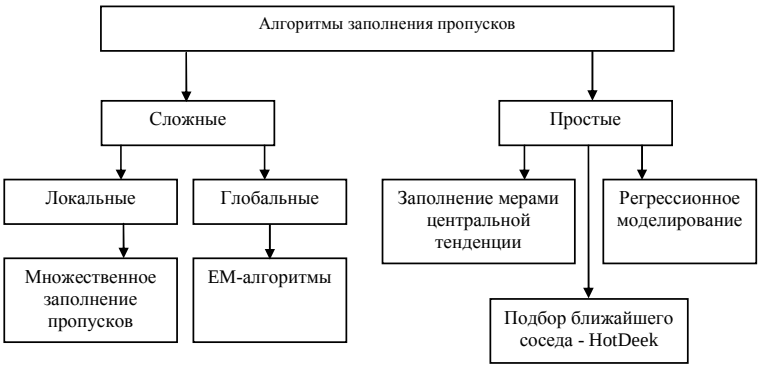


Рис. 2.4. Классификация методов заполнения пропусков

Можно утверждать, что теория восстановления пропущенных данных постоянно развивается, и соответственно появляются новые алгоритмы и модернизируются известные. Это связано с тем, что не может быть разработано универсального алгоритма, который был

бы применим и давал наилучшие результаты во всех ситуациях. Исследователи, доказывая преимущество того или иного подхода или метода, демонстрируют достоинства метода на конкретном примере. Но примеры тоже являются частными случаями и не доказывают полного превосходства одного метода над другим. Несмотря на существование большого количества методов восстановления данных, в широко известных пакетах по обработке статистических данных представлены лишь простейшие алгоритмы, которые в большинстве случаев не дают требуемой точности. Иначе говоря, задача восстановления данных сейчас в основном носит исследовательский характер и используется специалистами, более или менее представляющими механизм работы используемых алгоритмов, сохраняется как теоретическая проблема оценки точности результатов, полученных в результате применения алгоритмов восстановления.

Мы предлагаем рассмотреть метод восстановления данных, который может использоваться в ситуации, когда большинство известных методов не применимо. В большинстве методов восстановления данных используются признаки, измеренные в шкале отношений. При исследовании социально-экономических процессов часто получают данные, представленные в различных шкалах. Предлагаемый алгоритм позволяет работать с различными признаками. Конечно, он тоже не всегда гарантирует получение требуемой точности. Возможности алгоритма всегда ограничены имеющимися данными и их латентной структурой. В дополнение к алгоритму предлагаются несколько процедур оценки точности результатов, что позволяет исследователю самому принять решение о приемлемости полученного результата.

Рассмотрим алгоритм, основанный на разработке эталонов классифицированных данных.

Этот алгоритм строится на предположении случайности возникновения пропусков данных в таблице «объект-свойство». Для такого предположения часто используется аббревиатура MCAR (***missing completely at random***). Это предположение принимается в большинстве известных алгоритмов. Чаще всего предположение выполняется, и его можно проверить с помощью известных статистических методов. В таблице данных допускается присутствие данных, измеренных в различных шкалах. Таблицу данных представим в форме отсортированной таблицы (рис. 2.5), содержащей m+1 столбец. Первые m столбцов (X_1 X_2 ... X_i ... X_m) со-

держат значения признаков, не имеющих пропусков. Эти признаки будем называть восстанавливающими. Столбец Y содержит признак, в котором допущены пропуски. Этот признак будем называть восстанавливаемым. Первые n_0 содержат наблюдения без пропусков. Следующие n_1 строк имеют пропуски в признаке Y, то есть необходимо восстановить n_1 значений признака Y.

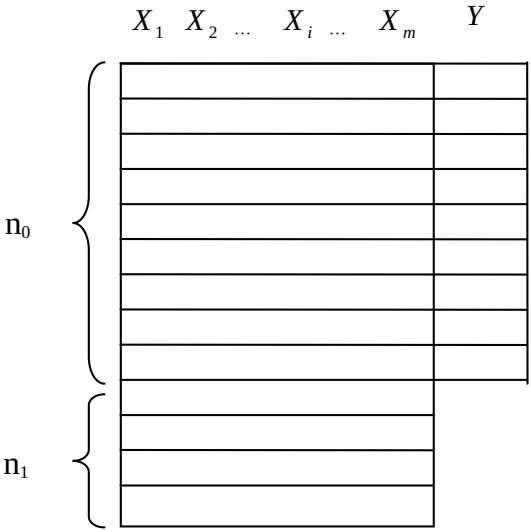


Рис. 2.5. Представление таблицы данных

Процедура более эффективно работает при восстановлении числовых признаков, но при достаточно большом количестве данных (не менее тысячи) можно пытаться восстанавливать и данные других типов. Для простоты будем считать, что данные числовые. Рассмотрим работу алгоритма по этапам.

Первый этап. Осуществляется преобразование всех числовых значений признаков к ранговым значениям (операция ранжирования). Признаки номинальные и ранговые не преобразуются. При этом номинальные признаки должны иметь небольшое количество значений (желательно меньше 10). Иначе номинальные признаки нужно подвергнуть предварительной обработке, приводя их к структурированному виду. Для этого применяются процедуры обработки качественных данных, описанные в работе [19].

Процедура ранжирования заключается в разбиении значений признака на равные интервалы и замене исходных значений ранговыми (номерами интервалов). Количество интервалов r должно быть не очень большим (рекомендуется 5), иначе могут появиться интервалы без значений, что нежелательно (но ситуация допустимая). Ранговые признаки ранжировать нет необходимости, можно использовать имеющуюся систему рангов. Ранжированные значения обозначим той же буквой только со штрихом.

Далее выборка (таблица данных) разбивается на две части, которые далее рассматриваются по отдельности. Первую выборку назовем «обучающей выборкой», вторую «контрольной выборкой».

Второй этап. Производится сортировка «обучающей выборки» по рангам признака Y'. Пусть признак Y имеет k рангов (классов). Отсортированная выборка представлена на рис. 2.6. Сумма количества наблюдений по классам равна объему «обучающей выборки» n_0 (2.1).

$$n_0 = s_1 + s_2 + s_3 + \cdots + s_k \qquad (2.1)$$

Заметим, что в таблице на рис. 2.6 столбец Y' содержит k групп повторяющихся значений.

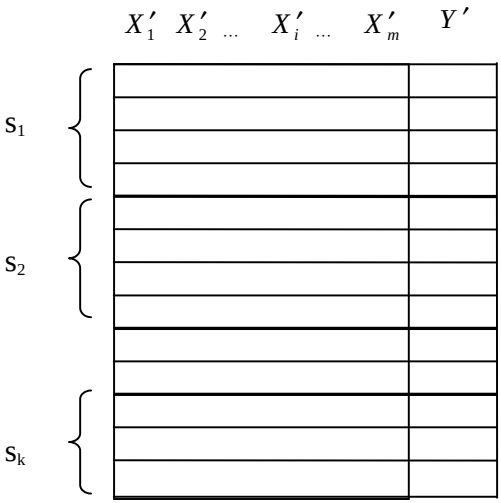


Рис. 2.6. Классифицированная таблица ранговых значений признаков

Третий этап. По данным каждого столбца X_j' рассчитываются таблицы абсолютных условных частотных рядов признаков по классам. Каждая такая таблица будет содержать *k* строк (по количеству классов) и *r* столбцов (по количеству градаций признаков X_j'). После этого таблицы нормируются по строкам путем деления на соответствующее количество элементов класса s_t. Тогда сумма элементов строк каждой таблицы будет равна единице. Нормированные таблицы представлены на рис. 2.7. Эти частотные ряды есть не что иное, как выборочные условные распределения переменных *X* при заданных значениях *Y*.

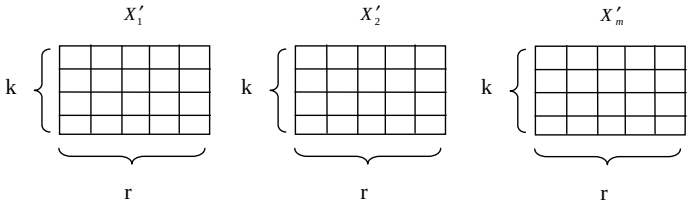


Рис. 2.7. Частотные ряды признаков по классам значений признаков

Четвертый этап. На этом этапе рассчитывается вектор-строка «эталонов» классов выборки. Эталон состоит из *m* частей по количеству признаков *X*. Каждая часть эталона состоит из *r* разрядов по количеству дискретных значений признаков X. Макет эталона представлен на рис. 2.8. Рассмотрим правило расчета элементов эталона. Каждая часть эталона рассчитывается по соответствующей таблице (рис. 2.7). Общее количество столбцов во всех таблицах также равно $r \times m$. Соответственно, размерность эталона тоже $r \times m$. Для расчета каждого элемента используются данные одного столбца таблицы. По данным каждого столбца определяется максимальное значение, и номер строки (номер класса) присваивается соответствующему элементу эталона.

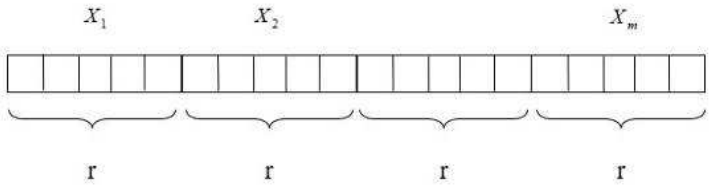


Рис. 2.8. Макет эталона классов

Процедуре расчета элементов эталона можно дать геометрическую интерпретацию. На рисунке 2.9 представлена графическая интерпретация расчета одной части эталона. Все остальные части рассчитываются аналогично. При расчете пятого элемента эталона для примера, приведенного на рис. 6, возникает неопределенная ситуация, которая состоит в том, что максимум достигается сразу в двух строках – второй и третьей. В этом случае предпочтение отдается тому классу (строке таблицы условных распределений), в котором количество элементов класса s_t больше. Предположим, что в нашем случае $s_3 > s_2$.

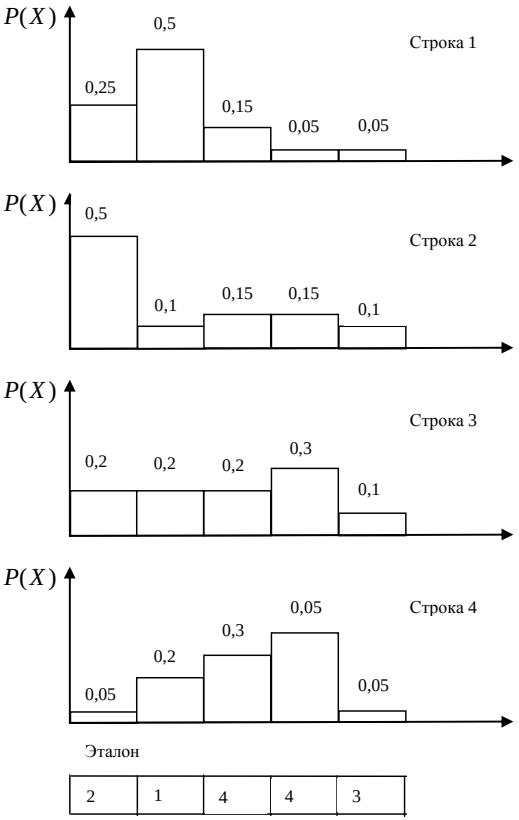


Рис. 2.9. Графическая интерпретация расчета одной части эталона классов

Пятый этап. На этом этапе производится сравнение многомерных данных «контрольной выборки» с эталоном и прогнозирование номера класса восстанавливаемого признака Y для наблюдений «контрольной выборки». Процедуру сравнения продемонстрируем на числовом примере (рис. 2.10).

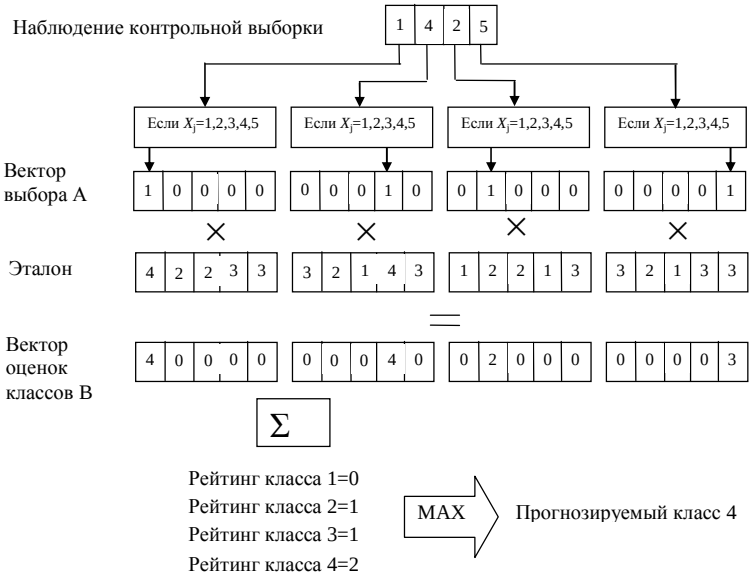


Рис. 2.10. Процедура сравнения наблюдения «контрольной выборки» с эталоном

Расчет производится в несколько шагов:

1) формируется вспомогательный вектор выбора А. Значение «1» указывает номер интервала, в котором лежит значение признака X_J;

2) формируется вспомогательный вектор В как произведение элементов вектора А и вектора эталона;

3) подсчитываются рейтинги классов как количество оценок по каждому классу в векторе оценок классов В;

4) определяется прогноз класса для наблюдения «контрольной выборки» по максимальному рейтингу класса.

На четвертом шаге опять может возникнуть неопределенность. Она возникает, когда максимум рейтинга класса достигается сразу

для нескольких классов. В этом случае предпочтение тоже отдается классу с наибольшим объемом выборки $s_t\left(t=\overline{1,k}\right)$.

Шестой этап. На этом заключительном этапе спрогнозированные значения номеров классов для элементов «обучающей выборки» заменяются средними значениями признака *Y*, рассчитанными по данным «обучающей выборки».

Рассмотренный алгоритм в соответствии с классификацией (рис. 2.4) можно отнести к классу сложных, глобальных не в силу сложности расчетов и множества этапов расчета, а в силу того, что при использовании алгоритма для решения конкретной задачи перед исследователем стоит проблема выбора. Необходимо задать количество интервалов для восстанавливающих признаков и восстанавливаемого признака. Возможно, для этого придется провести небольшой эксперимент. В сложных алгоритмах исследователь должен хорошо представлять принцип работы алгоритма. Допуская возможность эксперимента, для оптимизации точности работы алгоритма необходимо иметь критерии для сравнения различных вариантов построенного решающего правила.

Рассмотренный метод не столь чувствителен к подбору восстанавливающих признаков, хотя проблема с подбором восстанавливающих признаков все-таки существует, потому что излишние «неинформативные» признаки могут «засорять» полезную информацию. Рекомендуется на первых этапах использовать не очень большое количество восстанавливающих признаков, постепенно наращивая их количество. При подборе числовых признаков целесообразно включать сначала признаки с большей корреляцией с восстанавливаемым признаком. В случае использования ранговых признаков можно использовать ранговые коэффициенты корреляции. Не нужно исключать здравый смысл содержательного анализа признаков.

При восстановлении данных, полученных в ходе социально-экономических исследований, может оказаться очень полезным в качестве восстанавливающего рангового признака включать некоторый обобщающий признак, сформированный на основании представлений исследователя о социально-демографическом профиле групп населения. Такой признак формируется на основе нескольких признаков.

Можно привести показательный пример недоучета факторов социологического портрета. Например, при восстановлении при-

знака «возраст» может оказаться, что для наблюдения восстанавливаемого признака подходит возраст 70 и старше лет. И все бы хорошо, и алгоритм сработал корректно, и средние ошибки минимальны, но если принять во внимание, что это данные по студенту дневной формы обучения, то возникают вопросы по корректности подобного восстановления.

Для оценки точности восстановления данных можно использовать следующую методику.

Для сравнения результатов по точности восстановления данных, полученных с помощью различных методов или различных данных, используемых для восстановления, необходимы некоторые критерии качества.

Многие авторы считают, что после восстановления данных должны сохраняться основные свойства выборки (оценки функций плотности, средние и дисперсии признаков). При невысоком проценте восстанавливаемых данных эти параметры практически не изменяются после включения в выборку восстановленных данных.

По мнению автора, наиболее универсальным средством сравнения результатов являются оценки ошибок, рассчитанные методом скользящего экзамена [7].

Суть метода состоит в том, что решающее правило восстановления данных проверяется на данных обучающей выборки, которая содержит и восстанавливающие признаки, и восстанавливаемый признак в полном объеме (полные данные). Процедура скользящего экзамена состоит в том, что из обучающей выборки последовательно отбрасывается по одному наблюдению, которое потом восстанавливается с помощью оставшихся наблюдений. Ошибкой считается отнесение наблюдений восстанавливаемого признака к другим классам. Процедура повторяется n_0 раз. При достаточно больших объемах выборки (в тысячах наблюдений) это может занять достаточно много времени, но при использовании современной вычислительной техники это не проблема. В результате процедуры скользящего экзамена рассчитывается матрица ошибок восстановления:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix}, \qquad (2.2)$$

где p_{ij} – количество значений восстанавливаемого признака, из класса *i*, отнесенных при восстановлении к классу *j*.

Иначе говоря, количество правильных восстановлений будет равно сумме диагональных элементов матрицы:

$$Q = \sum_{t=1}^{k} p_{tt} \tag{2.3}$$

Сумма элементов по строке дает объем класса обучающей выборки:

$$Q_t = \sum_{l=1}^{k} p_{tl} = s_t, \quad t = \overline{1,k} \tag{2.4}$$

Качество восстановления можно оценить показателями Ω – процент ошибок восстановления; Ω_t – процент ошибок по классам:

$$\Omega = 1 - \frac{n_0 - Q}{n_0}, \tag{2.5}$$

$$\Omega_t = 1 - \frac{p_{tt}}{s_t} \tag{2.6}$$

Более подробный анализ ошибок позволяет сделать нормированная матрица ошибок восстановления. Нормированные значения матрицы ошибок производятся путем деления каждой строки на объем класса обучающей выборки $s_t\left(t = \overline{1,k}\right)$. Тогда сумма элементов каждой строки матрицы будет равна единице. Элементы нормированной матрицы ошибок P' обозначим p'_{ij}:

$$P' = \begin{bmatrix} p'_{11} & p'_{12} & \cdots & p'_{1k} \\ p'_{21} & p'_{22} & \cdots & p'_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p'_{k1} & p'_{k2} & \cdots & p'_{kk} \end{bmatrix} \tag{2.7}$$

Ошибки по классам могут быть распределены неравномерно, и, возможно, придется отказаться от восстановления некоторых данных. Матрицы ошибок восстановления могут расширить понимание причин возникновения неответов, что важно для организации последующих исследований при мониторинге социально-экономических процессов. Многие авторы считают проблему выявления причин неответов еще более важной, чем проблему восстановления данных.

Изложенная процедура оценки уровня ошибок восстановления пригодна для любого типа восстанавливаемых признаков. При восстановлении числовых признаков по обучающей выборке можно считать и дисперсии отклонений исходных значений и оценок, полученных при восстановлении данных. Такие оценки могут быть рассчитаны как по всей выборке, так и по классам.

Прежде чем приступать к процедуре восстановления данных, необходимо провести тщательный анализ возможных ошибок в данных и выявить выбросы. Для этого мы используем процедуры, автоматизирующие данный процесс. При больших объемах выборок и значительном количестве признаков без специальных программных средств не обойтись.

Необходимо заметить, что заниматься трудоемкой процедурой восстановления данных имеет смысл, если исследователь предполагает в своей работе использование многомерных статистических методов. Многие исследователи ограничиваются исследованием одномерных признаков, поэтому восстанавливать пропуски в данных им не имеет смысла, поскольку можно внести только искажения в результаты, полученные по полным данным.

Некоторые многомерные статистические методы могут быть реализованы программно с учетом наличия пропусков в части данных. Простейшим примером служит расчет ковариационной матрицы по данным с пропусками. При расчете ковариационной матрицы используются оценки средних, которые могут быть рассчитаны по каждому признаку в отдельности с учетом пропусков. Можно было привести и более сложные примеры, однако заметим, что учет пропусков программно способен привести к существенному усложнению программы, а при их использовании должны быть оговорены обозначения отсутствия данных в таблицах данных. При этом для разных признаков (по типу) должны быть свои условные обозначения. Поэтому такой подход применяется в исключительных случаях. Мы, например, использовали его на предварительных этапах обработки данных при обнаружении выбросов и грубых ошибок.

Относительно рассмотренного алгоритма в качестве выводов по статье можно добавить, что метод может дать более точные результаты по сравнению с некоторыми другими методами за счет расширения диапазона используемых признаков.

Серия экспериментов на модельных данных показала преимущества метода перед другими. В экспериментах мы широко использовали программу моделирования многомерных нормальных распределений [20]. В настоящее время мы продолжаем эксперименты на модельных данных для выявления условий и ограничений применения метода. Доказано, что любой метод очень сильно зависит от имеющихся данных. Если они «плохие», то здесь бессильны самые совершенные методы.

Можно добавить, что разработанные нами программные средства либо содержат встроенные блоки, позволяющие проводить эксперименты, либо включают параметры, автоматизирующие экспериментальную работу, облегчая работу исследователя.

Разработанные программные средства прошли апробацию на больших массивах данных, собранных в исследованиях, посвященных анализу проблем туризма и рекреации в Приморском крае [21].

2.2.2. Повышение качества данных на основе анализа статистической зависимости признаков

В последние годы в России для исследования социально-экономических процессов широко используются онлайн-опросы [22]. Количество онлайн-опросов возросло многократно после того, как получили распространение специальные сервисы по конструированию онлайн-анкет и сопровождению процесса анкетирования в сети Интернет [23, 24]. В публикациях российских ученых продолжаются дискуссии относительно эффективности онлайн-опросов и традиционных опросов на бумажных носителях [25, 26]. В зарубежных научных публикациях последних лет такая тема поднимается все реже [27]. Напрашивается вывод, что зарубежные исследователи практически полностью перешли на опросы в Интернете. Среди зарубежных сейчас достаточно много публикаций, в которых рассматриваются пути повышения качества данных, полученных с помощью онлайн-опросов [28–32]. В российской научной литературе также начали появляться публикации, посвященные проблеме повышения эффективности онлайн-опросов [33–38], которая в первую очередь определяется качеством данных, получаемых с помощью онлайн-опросов.

Понятно, что, создав свою онлайн-анкету и разместив ее в свободном доступе в сети Интернет, наивно полагать, что она привле-

чет достаточное для анализа количество респондентов в обозримом будущем. Для сбора анкет в Интернете, необходимо иметь собственную базу респондентов или использовать заинтересованных чем-либо интервьюеров, которые бы приглашали принять участие в опросе своих знакомых.

С увеличением спроса на онлайн-анкетирование появились специальные интернет-компании, которые осуществляют анкетирование, привлекая респондентов за плату (такие сайты называют «опросниками»). Среди российских «сайтов-опросников» известны следующие: «Анкетка», «Мое мнение», «Интернет-опрос», «Экспертное мнение». В сети представлено и множество зарубежных «сайтов-опросников», которые работают и с русскоязычными респондентами: «GlobalTestMarket», «ThePanel Station», «i-Say». Однако к анкетам, заполненным профессиональными респондентами, нужно относиться с осторожностью [39].

В настоящее время проблема анализа достоверности информации онлайн-опросов недостаточно разработана. Поэтому методические подходы к анализу достоверности различных видов информации весьма востребованы практиками. Ранее нами был разработан ряд методов анализа достоверности анкетных данных, которые с успехом применялись в течение длительного времени при анализе анкет, собираемых на бумажном носителе. В настоящей работе предлагается к рассмотрению новый метод, позволяющий повысить достоверность данных, собираемых с помощью онлайн-анкетирования.

Предметом исследований являются данные, полученные в ходе онлайн-опросов. Такие данные формируют многомерные выборки признаков таблиц статистических данных, которые используются для анализа социально-экономических процессов и явлений.

В большинстве анкетных форм, представленных в Интернете, в качестве основного типа вопросов используют вопросы с набором альтернативных вариантов ответов. Такие анкеты требуют минимального времени заполнения и наиболее понятны большинству респондентов. Тем не менее, при недобросовестном отношении к опросу части респондентов выборки могут содержать недостоверную информацию, например, при выборе респондентом первых попавшихся ответов на вопросы анкеты.

Вопросы анкеты с набором альтернативных ответов порождают номинальные признаки с фиксированным для каждого вопроса набором значений.

Пусть произведена выборка объемом n (количество анкет). Обозначим совокупность номинальных признаков в таблице анкетных данных как $(x_1; x_2; x_3; \dots x_m)$, где m – количество номинальных признаков. Известно фиксированное количество возможных значений для каждого номинального признака $(k_1; k_2; k_3; \dots; k_m)$, где m – количество признаков. Рассмотрим основные принципы, заложенные в основу предлагаемого метода.

Анализ достоверности анкетных данных основан на использовании принципа «скользящего экзамена» [40]. Суть его заключается в том, что вся выборка делится на две части. Первая часть включает все данные за исключением тех, относительно которых осуществляется проверка качества (достоверности) данных. Вторая часть включает данные, подлежащие проверке. Подлежать проверке могут данные, инициированные одним, отдельно взятым интервьюером (пакет данных). В частном случае это может быть только одна анкета (строка в таблице данных). Первую выборку будем называть обучающей (ОВ), а вторую контрольной (КВ). По двум выборкам рассчитываются различные частотные характеристики, используемые для расчета некоторого статистического критерия для данной КВ, сравнение значений которого со значениями, рассчитанными по всей выборке, позволяет выделить нетипичные данные, которые могут рассматриваться как «подозрительные» с точки зрения достоверности. После расчета критерия данные второй выборки (КВ) возвращаются в первую выборку (ОВ), а из нее изымается следующая группа данных, инициированных другим интервьюером. Процедура повторяется до тех пор, пока все пакеты не будут протестированы.

Процедура тестирования качества основана на гипотезе, что большая часть данных все-таки отвечает требованиям по качеству, а некачественные (недостоверные) данные составляют меньшинство. Конечно, доля одного пакета не должна быть преобладающей во всей выборке. При тестировании единичных анкет это правило выполняется автоматически, поскольку вклад каждой анкеты в частотные характеристики выборки незначительный.

Вторая гипотеза, лежащая в основе разработанного метода, состоит в том, что обычно респондент, предоставляющий недостовер-

ную информацию, нарушает некоторые закономерности (внутреннюю логику), присущие исследуемому в анкетном опросе процессу. Например, респондент, безответственно относящийся к ответам на вопросы анкеты, отвечает «как попало», не задумываясь над содержанием вопросов. Если респондент недобросовестно относится к ответу на одни вопросы анкеты, он чаще всего относится также ко всему анкетному опросу. Иными словами, предполагается, что предоставляя недостоверные данные, респондент не преследует цели умышленно дезинформировать исследователя. В противном случае на ответы вопросов анкеты могло бы потребоваться времени даже больше, чем при добросовестном отношении к опросу.

Третья гипотеза состоит в том, что ответы на вопросы анкеты имеют некоторую внутреннюю логику, ответы на одни вопросы анкеты влияют в статистическом смысле на другие ответы. Иногда даже возможные варианты ответов на отдельные вопросы могут входить в полное противоречие. Например, если студент в одном вопросе о месте проживания указывает «кампус университета», а в другом вопросе о времени, затрачиваемом на дорогу до университета, дает ответ «более часа», наблюдается явное противоречие в ответах на эту пару вопросов. Рассмотрим еще один пример, который если теоретически не приводит к противоречию, то имеет крайне низкую вероятность. Если респондент о гендерной принадлежности указывает мужской пол, а в другом ответе о частоте посещения маникюрного салона указывает «один раз в неделю», то этот вариант в принципе возможен, но статистически незначим и может вызвать подозрение исследователя в достоверности и других ответов на вопросы анкеты. Нелогичное поведение респондента может быть вызвано и неосознанным действием, например, при недопонимании вопроса анкеты. Однако такие ответы все равно нельзя признать достоверными. А на будущее исследователь должен учитывать, что отдельные вопросы могут восприниматься неадекватно, и предпринять какие-то действия, чтобы такие ситуации не повторялись.

Предложенный критерий учитывает сразу все пары ответов респондентов. Критерий дает одно интегральное значение для тестируемого пакета анкет (или отдельной анкеты). Такие значения представляют собой некоторую статистику (значение критерия), которая может выводиться в виде графика, на котором легко можно обнаружить «подозрительные» данные.

Выделенные «подозрительные» анкеты или пакеты анкет не отбрасываются автоматически, а подвергаются исследователем детальному содержательному анализу проверки правдоподобия ответов на вопросы анкеты с целью выяснения причин выделения таких анкет в ряду прочих. И только после установления того, что такие анкеты нельзя признать достоверными, они отбрасываются.

Принципы расчета критерия рассмотрим на примере модельных данных. Для удобства изложения примем некоторые допущения.

Будем полагать, что тестируются не пакеты анкет, а каждая анкета по отдельности. При рассмотрении пакетов анкет значения критерия просто суммируются по анкетам пакета и нормируются по объему выборки пакета.

Объем модельной выборки составляет 101 наблюдение (n=101). Это сделано для того, чтобы обучающая выборка включала 100 наблюдений. Выборка представлена пятью номинальными признаками $(x_1; x_2; x_3; x_4; x_5)$.

Набор значений каждого признака может быть задан множеством $(\pi_{j_1}; \pi_{j_2}; \pi_{j_3}; \ldots; \pi_{jt_j})$, где $j = 1,2,\ldots m$, m – количество признаков. Предположим, что количество возможных значений всех признаков одинаково. В примере принято $t_j = 4$. Считается, что предварительно выборка номинальных значений признаков была преобразована путем замены возможных значений признаков их номерами в списке, т.е. $(\pi_{j_1}; \pi_{j_2}; \pi_{j_3}; \pi_{j_4}) = (1; 2; 3; 4)$.

Для упрощения расчетов повторяющиеся наблюдения выборки были сгруппированы. Для каждой уникальной последовательности значений признаков предварительно была рассчитана «частота встречаемости в выборке» Q_s, где $s = 1,2,3,\ldots\alpha$, α – количество уникальных последовательностей ответов, встречавшихся в выборке ($\sum Q_s = 101$). В нашем примере $\alpha = 11$. Сгруппированная выборка представлена в компактном виде в табл. 2.1. Расчет критерия λ_s рассмотрен ниже. Частотные ряды по всем пяти признакам представлены на рис. 2.11.

Таблица 2.1

Сгруппированная выборка модельных данных

Номер уникальной последовательности ответов	Значения признаков выборки					Частота встречаемости в выборке Q_s	Значение критерия λ_s
	x_1	x_2	x_3	x_4	x_5		
1	1	1	4	3	1	21	0,790
2	1	1	4	3	2	19	0,902
3	1	1	3	3	2	18	0,792
4	2	2	3	2	3	10	0,182
5	3	3	2	3	2	10	0,256
6	1	2	1	3	2	8	0,424
7	2	1	3	4	2	7	0,272
8	3	4	3	4	3	3	0,094
9	2	2	3	3	2	2	0,356
10	4	1	3	1	4	2	0,090
11	1	2	2	2	3	2	0,076

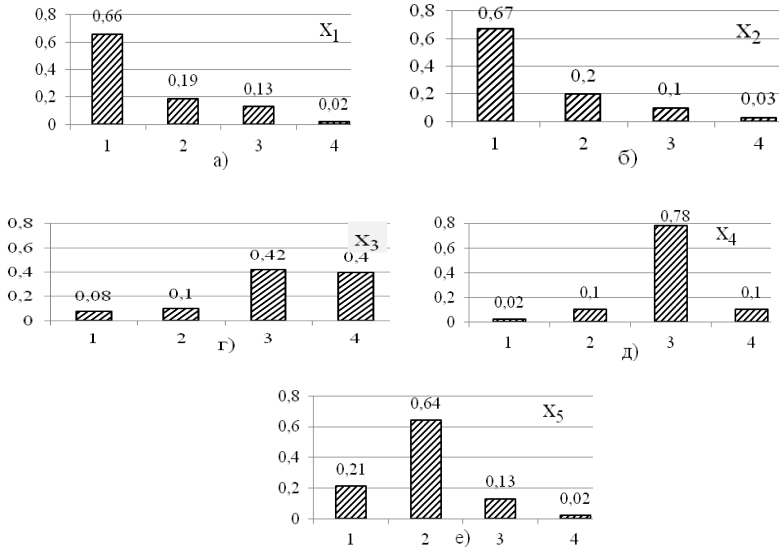


Рис. 2.11. Частотные ряды по пяти признакам обучающей выборки

Сформируем обучающую выборку, в которую включим первые десять уникальных последовательностей ответов. В контрольную выборку включим одиннадцатую последовательность. Для примера сначала мы выбрали последовательность, которая наиболее контрастирует со всеми остальными.

Таким образом, контрольная выборка состоит из одного наблюдения $(x_1; x_2; x_3; x_4; x_5) = (1; 2; 2; 2; 3)$. В обучающую выборку включены все остальные 100 наблюдений исходной выборки.

Сочетания возможных пар признаков выборки могут быть представлены полным графом G с количеством ребер, равным d (m). Для заданного числа признаков d=10 (рис. 2.12).

$$d = m \times \frac{m-1}{2} \qquad (2.8)$$

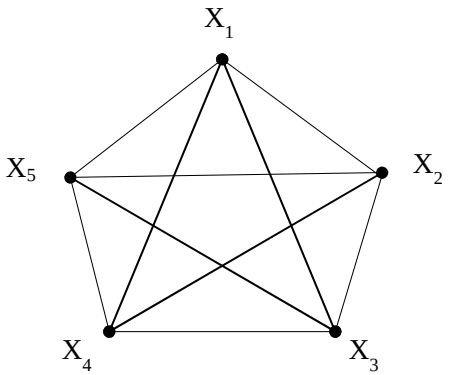


Рис. 2.12. Полный граф сочетания признаков при m=5

Расчет элементов критерия представлен в табл. 2.2. Все возможные пары признаков размещены во втором столбце таблицы. Такие же сочетания пар будут и для любого другого наблюдения выборки. Во втором столбце указаны конкретные значения признаков для рассматриваемого наблюдения (контрольного). Далее рассчитывается количество таких пар в таблице данных в обучающей выборке без учета кратности. Например, пара $(x_1; x_2) = (1; 2)$ встречалась только однажды в уникальном наблюдении под номером 6 в табл. 2.1: $(x_1; x_2; x_3; x_4; x_5) = (1; 2; 1; 3; 2)$. Такая пара может

встречаться и в других уникальных наблюдениях (максимум во всех, что крайне маловероятно). Для данной контрольной выборки обнаружено пересечений с уникальными наблюдениями только по одному разу в четырех случаях (четыре единицы в столбце «Количество пар в ОВ без учета кратности» в табл. 2.2). Каждое уникальное значение имеет свою кратность Q_s. Например, наблюдение под номером 6 в ОВ $(x_1; x_2; x_3; x_4; x_5) = (1; 2; 1; 3; 2)$ имеет кратность 8, а другое наблюдение $(x_1; x_2; x_3; x_4; x_5) = (2; 2; 3; 2; 3)$ под номером 4, имеющее аналогичную пару $(x_2; x_4) - (2; 2)$ в контрольной выборке, имеет кратность 10. Суммарная кратность повторения совпадения каждой пары значений обучающей и контрольной выборок записана в столбце «Количество пар в ОВ с учетом кратности» табл. 2.2.

Далее рассчитывается частота встречаемости каждой пары обучающей выборки в контрольной выборке (столбец «Частота» в табл. 2.2). Для этого «Количество пар в ОВ с учетом кратности» делится на объем выборки (в нашем случае 100).

Теперь необходимо сделать одно важное замечание. Дело в том, что в общем случае все пары неравнозначны. Количество возможных вариантов ответов зависит от количества значений признаков, составляющих пару (в нашем случае для всех признаков $t_j = 4$). Количество возможных вариантов значений каждой пары признаков $(x_i; x_j)$ равно $\rho(x_i; x_j) = t_i \times t_j$ (в нашем случае для всех пар признаков $\rho(x_i; x_j) = 16$). Поэтому для различных пар необходимо ввести поправочный коэффициент (вес пары). Веса для разных сочетаний пар признаков в общем случае рассчитываются по формуле (2.9):

$$V_{ij} = \frac{t_i t_j}{\Sigma t_i t_j}. \qquad (2.9)$$

Суммирование производится по всем возможным d парам полного графа сочетания признаков. В нашем примере все пары будут иметь один и тот же вес $V_{ij} = 0{,}2$. Затем рассчитывается «вклад каждой пары» как произведение частоты пары на вес. Суммирование по вкладам каждой пары дает общее значение критерия λ_s. В данном случае для строки данных $(x_1; x_2; x_3; x_4; x_5) = (1; 2; 2; 2; 3)$ значение критерия равно $\lambda_{11} = 0{,}076$.

Таблица 2.2

Расчет критерия для уникальной последовательности под номером 11

Пара	Значения признаков в КВ	Количество пар в ОВ без учета кратности	Кратность				Количество пар в ОВ с учетом кратности	Частота	Вес пары	Вклад пары
			1	2	3	4				
(x_1; x_2)	(1;2)	1	8				8	0,08	0,2	0,016
(x_1; x_3)	(1;2)	0					0	0,00	0,2	0,000
(x_1; x_4)	(1;2)	0					0	0,00	0,2	0,000
(x_1; x_5)	(1;3)	0					0	0,00	0,2	0,000
(x_2; x_3)	(2;2)	0					0	0,00	0,2	0,000
(x_2; x_4)	(2;2)	1	10				10	0,10	0,2	0,020
(x_2; x_5)	(2;3)	1	10				10	0,10	0,2	0,020
(x_3; x_4)	(2;2)	0					0	0,00	0,2	0,000
(x_3; x_5)	(2;3)	0					0	0,00	0,2	0,000
(x_4; x_5)	(2;3)	1	10				10	0,10	0,2	0,020

Для сравнения рассмотрим таблицу расчета критерия для характерного наблюдения данной выборки $(x_1; x_2; x_3; x_4; x_5) = (1; 1; 4; 3; 2)$ (номер 2 в таблице 2.2). Расчеты значения критерия для данного наблюдения представлены в табл. 2.3. В данном случае значение критерия будет равно $\lambda_2 = 0{,}902$. Значения критерия по всем уни-

кальным наблюдениям представлены в последнем столбце табл. 2.1 (столбец «Значение критерия»).

Таблица 2.3

Расчет критерия для уникальной последовательности под номером 2

Пара	Значения признаков в КВ	Количество пар в ОВ без учета кратности	Кратность					Количество пар в ОВ с учетом кратности	Частота	Вес пары	Вклад пары
			1	2	3	4	5				
$(x_1;x_2)$	(1;1)	3	21	18	18			57	0,57	0,20	0,114
$(x_1;x_3)$	(1;4)	2	21	18				39	0,39	0,20	0,078
$(x_1;x_4)$	(1;3)	4	21	18	18	8		65	0,65	0,20	0,130
$(x_1;x_5)$	(1;2)	3	18	18	8			44	0,44	0,20	0,088
$(x_2;x_3)$	(1;4)	2	21	19				40	0,40	0,20	0,080
$(x_2;x_4)$	(1;3)	3	21	18	18			57	0,57	0,20	0,114
$(x_2;x_5)$	(1;2)	2	18	18				36	0,36	0,20	0,072
$(x_3;x_4)$	(4;3)	2	21	18				39	0,39	0,20	0,078
$(x_3;x_5)$	(4;2)	1	18					18	0,18	0,20	0,036
$(x_4;x_5)$	(3;2)	5	18	18	10	8	2	56	0,56	0,20	0,112

На рис. 2.13 отражен график упорядоченных значений критерия. На горизонтальной оси указаны номера уникальных последовательностей. Из графика видно, что три последовательности (со значениями критериев 0,094; 0,09 и 0,076) выбиваются из общего ряда наблюдений выборки, поэтому их можно рассматривать как «сомнительные» с точки зрения достоверности данных. В реальной ситуации необходимо путем содержательного анализа ответов респондентов выяснить причины, в результате которых были выделены именно эти наблюдения. Поскольку «подозрительных» анкет, как

правило, немного, то такой анализ обычно не вызывает затруднений.

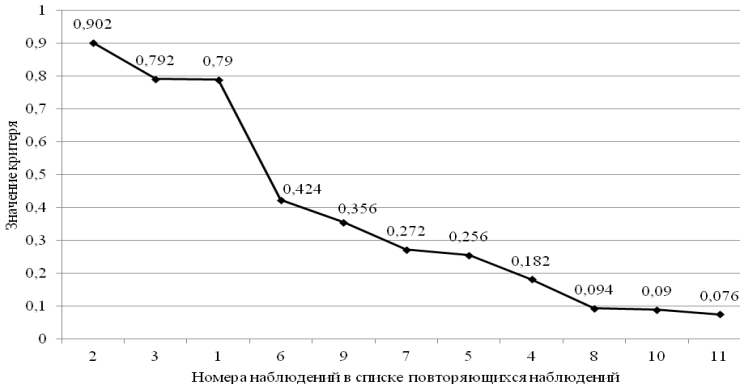


Рис. 2.13. График значений критерия для уникальных последовательностей из табл. 2.1

На основе предложенного метода анализа достоверности анкетных данных была разработана компьютерная программа в среде EXCEL. Кроме модельных данных программа была протестирована и на реальных данных, показав хорошие результаты. Однако программа в настоящее время содержит ряд ограничений на масштаб задачи, поэтому она пока еще не может быть использована для широкого круга анкетных опросов.

Для рекомендации ее к широкому использованию необходимо провести еще ряд дополнительных экспериментов на реальных данных при различном сочетании уровней значений признаков. В дальнейшем предполагается произвести оптимизацию программы с позиций оптимизации вычислительного процесса.

Основным достоинством предложенного метода является его способность учитывать связь между различными признаками исследуемой таблицы данных.

Предложенная технология тестирования достоверности онлайн-анкет позволяет в оперативном режиме контролировать ход опроса. При необходимости исследователь может принять действия по улучшению качества данных (например, провести дополнительные

консультации интервьюеров или скорректировать формулировки некоторых вопросов).

Для анализа достоверности данных онлайн-опросов целесообразно использовать ряд критериев. Например, можно использовать критерии, разработанные нами ранее для различных типов данных.

Если анкета включает вопросы, которые порождают различные типы признаков, то они тоже могут быть включены в список анализируемых признаков после предварительного преобразования значений признаков с понижением мощности шкал. Дополнительные номинальные признаки могут быть получены в результате типологии открытых вопросов.

Возможно, предложенный метод в дальнейшем может быть использован для классификации наблюдений. Например, на графике на рис. 2.13 можно выделить три зоны, которые могут быть положены в основу классификации. Эту возможность легко определить в результате дальнейших исследований.

2.3. Подходы и методики обработки качественной информации

2.3.1. Типологизация плохоструктурированных данных

Для сегментирования рынков по психографическим признакам широко используются качественные данные. В последние годы методы и технологии обработки качественных данных активно развивались. Теоретические основы типологизации рассматриваются в работах [41–43].

В мировой литературе утвердилось понятие «качественные методы исследования», которые иногда называют «мягкими» в отличие от «количественных» и «жестких». В определенном смысле можно говорить также о неформализованных или слабо формализованных подходах в сравнении с жестко формализованными.

Сегментирование по психографическим признакам основывается на теории типологического анализа.

Типологический анализ – метаметодика анализа данных, совокупность методов изучения социального феномена, позволяющих выделить социально значимые, внутренне однородные, качественно отличные друг от друга группы эмпирических объектов, характери-

зующиеся типообразующими признаками, природа которых различна, и интерпретируемые как носители различных типов существования феномена [44, 45].

Типологический анализ в основном базируется на анализе качественных данных. По мнению В.А. Ядова, качественные методы дают возможность глубже понять изучаемое явление и предложить множественную интерпретацию [46].

Основанием типологии служит совокупность суждений (утверждений) о близости (схожести, похожести) объектов, носителей информации об изучаемых социальных феноменах (явлениях, процессах).

Предметом типологии является совокупность основных характеристик социального феномена, ответственная за отнесение эмпирических объектов к однотипной группе.

Рассмотрим некоторые термины, которые используются в типологическом анализе.

Тип (***type***) – вид, форма существования социальных феноменов в науке или повседневной жизнедеятельности людей; сущность, знание о которой всегда относительно. Имеет три условных значения в смысле – типовой, типологический, типический.

Социальный тип (***social type***) – группа людей, схожих в восприятии другими по каким-то основаниям (облику, месту в социальной структуре, поведению и т. д.).

Типовой – в типологическом анализе: распространенный, модальный, часто встречающийся.

Типический – в типологическом анализе: специфический, характерный, редко встречающийся.

Типологический – в типологическом анализе: особенный, общий, объединяющий.

Типизация (***typization***) – конструирование людьми социальной реальности на основе придания окружающим ярлыков, спонтанная классификация.

Типологизация (***typologization***) – процедура систематизации знаний об изучаемых феноменах либо для введения (задания) типов, либо для поиска знаний о типах. Типологизация служит для конструирования типов.

Типология (***typology***) – совокупность типов, результат их конструирования. Способ конструирования типов.

Типообразующий признак – характеристика, свойство социальных феноменов, на основе которых либо конструируются типы, либо формулируются гипотезы об их существовании. Типообразующий признак – это концептуальная переменная.

В конструировании или построении типов различают две ситуации:

1. Исследователь конструирует типы, исходя из знаний, накопленных в интересующей его предметной области. Они конструируются либо априори как результат научного «озарения», либо апостериори, опираясь на накопленные знания, в том числе и на эмпирические (по результатам большого числа исследований).

2. Исследователь опирается на мысленные конструкции, идеальные представления, вытекающие из какой-либо социологической теории. Тогда типы носят идеальный характер. Понятие «идеальный тип» входит во все социологические словари, и его введение в социологии связывают с именем Макса Вебера.

Неструктурированные данные – данные, существующие в виде текстов и полученные в процессе проведения разного вида интервью. Сюда также относятся тексты ответов на открытые вопросы (с неограниченным полем поиска ответов) и любые другие тексты, к которым обращается исследователь как к социологическим данным. Эти данные не формализованные, не организованные специальным образом.

Существует несколько подходов к решению проблемы неструктурированных или неформализованных данных. В саму природу нечисловой информации заложена возможность использования типологического анализа для ее обобщения и структуризации. «При всей уникальности действующего индивида большая часть его индивидуальных смыслов типична, т.е. обладает общностью с другими людьми» [47]. Одной из основных задач является разработка алгоритма построения типологий таким образом, чтобы преодолеть субъективность исследователя, не упустив при этом важную информацию. Алгоритмы предполагают применение сжатия и структурирования информации так, чтобы она сохранила свойства исследуемого объекта.

Можно выделить три подхода построения типологий:

– концепция типологических операций А. Бартона и П.Ф. Лазарсфельда;

– анализ структуры У. Герхардта;

– типологический анализ У. Кукартца.

Первый подход основан на использовании «типологических операций»: редукции, субструкции и трансформации. Через определение признаков и степеней их выраженности строится пространство свойств, лежащих в основе типологии. Используя графическую или табличную формы представления данных, определяются все возможные комбинации и все потенциальные типы. Все типы связываются на одном пространстве. Комбинации признаков могут быть сокращены. Субструкция выявляет само пространство признаков, лежащее в основе типологии, которое может трансформироваться при интерпретации сконструированных типов. Несмотря на то, что данный подход был предложен для конструирования типов в исследованиях количественной стратегии, он имеет центральное значение для обобщения нечисловой информации.

Второй и третий подходы используют те же основные типологические операции, пытаясь при этом преодолеть его главный недостаток: определение критериев отбора признаков для анализа данных. В основе второго подхода лежат «идеальные типы», которые служат базисом для анализа информации, полученной в ходе исследования. На первом этапе проводится сравнение случаев через их реконструкцию, чтобы выявить их особенности. Это привносит в исследование прозрачность процесса обобщения и его результатов. На втором этапе исследуемые случаи группируются с помощью их сопоставления. Такие приемы в целом соответствуют «концепции типологических операций» и позволяют узнать все потенциально возможные комбинации признаков. На последнем этапе выявляются и объясняются смысловые связи внутри и между полученными группами. Для этой цели был разработан анализ структуры и процесса, состоящий из двух шагов обобщения. Главными недостатками данного подхода являются трудность абстрагироваться от субъективных представлений исследователя при обобщении данных и отсутствие алгоритма проверки сконструированных типов.

Типологический анализ в данном подходе имеет ряд особенностей по сравнению с другими методами обобщения качественных данных. В процессе работы абстрагируются от каждого отдельно исследуемого случая и получают типичное событие как результат упорядоченных фаз секвенции. «Структурная герменевтика», наоборот, понимает материал в его единичности, неотделенности от

каждого конкретного исследуемого случая. Типологический анализ находится в точке пересечения между индивидуальной историей и общепринятой. Во втором подходе исследуемые случаи сохраняются по возможности в своей целостности, в третьем подходе при анализе отдельных случаев и их сравнении используется тематическое обобщение.

Третье направление в полной мере нельзя считать отдельным подходом. Это «инструмент, построенный для целей выражения методологических взглядов М. Вебера» [48]. При разработке инструментов типологического анализа (программных средств) предпринимается попытка соединить в исследовании различные способы типологизации нечисловой информации с учетом их достоинств и недостатков.

Известные социологи Н. Филдинг и Р. Ли в предметной области инструментальных средств анализа качественных данных предложили использовать специальный термин «компьютерный, но ассистируемый анализ качественных данных» (Computer Assisted Qualitative Data AnalysiS, CAQDAS). Современный компьютерно-ассистируемый анализ качественных данных является методологической исследовательской областью, объединяющей ученых многих стран. Ассистируемый анализ представлен множеством компьютерных пакетов, в том числе: Atlas.ti, MAXQDA, NVivo, xSight, Qualrus, Ethnograph и др. Эти пакеты являют собой класс компьютерных программ, которые включают в свою архитектуру специальные структуры, называемые функциями кодирования и реконструирования качественных, нечисловых данных (coding and retrieval functions). Функции кодирования и реконструирования данных (ФКР) представляют собой компьютерный инструмент (tool), используемый в человеко-машинном режиме и ассистирующий пользователю при изучении данных, представленных в так называемых нечисловых форматах. В основе ассистирования лежит аппарат аналитических переобозначений, называемых кодами, введение и связывание которых между собой осуществляется самим пользователем. Более подробный анализ методологических разработок компьютерного инструментария анализа качественных данных представлен в работе [48].

Отмечая широкие возможности зарубежных компьютерных инструментальных средств анализа качественных данных, нельзя не

сказать, что они не нашли своего распространения не только у отечественных ученых, занимающихся проблемами сегментирования рынка, но и у отечественных социологов. Эти средства имеют несколько другую направленность и больше предназначены для решения гуманитарных и лингвистических задач.

В данной работе разработана методология анализа качественных данных, которые собираются с целью исследования рынков и, в частности, туристского рынка. Методология включает достаточно простые в применении компьютерные программные инструменты, которые могут быть использованы в рамках компьютерной среды EXCEL, наиболее распространенной среди отечественных исследователей рынков.

Предложенная методология основывается на представлении данных в форме «термов».

Термом называется символьное выражение: $t(X_1, X_2, \ldots X_n)$, где t – имя терма, называемого функтор или «функциональная буква», а $X_1, X_2, \ldots X_n$ – термы, структурированные или простейшие. Для формального описания термов в работе было введено новое понятие – составной признак.

Предложенную методологию анализа данных, используемых при сегментировании, необходимо рассматривать не столько как учение о методах, сколько как учение о взаимодействии методов между собой на разных классах исследовательских практик анализа данных. Структурно эта методология включает приемы и методы сбора и измерения информации, а также математические методы.

Использование комбинации количественных и качественных методов часто является наилучшим решением проблемы сегментирования рынка. Различные методы дополняют и контролируют друг друга, ограничения одного метода уравновешиваются ограничениями другого. Такие свойства называют комплиментарностью и триангуляцией.

Комплементарными называют несходные или даже противоположные теории, концепции, модели и точки зрения, отражающие различные взгляды на действительность.

Триангуляция – возможность использования несколько источников информации. В анализе рынка можно выделить несколько типов триангуляции: триангуляция данных; триангуляция исследователей; триангуляция методов; триангуляция теорий.

2.3.2. Компьютерная технология разработки типологий

Данные, полученные в результате ответов на открытые вопросы, являются неструктурированными и поэтому не могут быть обработаны числовыми методами. В результате обработки такие данные преобразуются к структурированному виду. Этот процесс достаточно трудоемкий и требует применения специальной компьютерной технологии. В основу рассматриваемой в данной работе компьютерной технологии положен алгоритм обработки качественных данных, использующий специальную операцию типизации данных. Рассматриваемая компьютерная технология получила свою реализацию в виде специальной компьютерной программы. Программа снабжена элементами экспертной системы, которые позволяют существенно сократить время обработки данных при повторении аналогичных опросов.

Качественные данные порождают открытые вопросы, в которых респонденту предлагается сформулировать ответ в форме текста. Простейшими формами таких вопросов служат вопросы о занимаемой должности, профессии, месте жительства. Это такие вопросы, в которых исследователь не может сформировать список альтернативных ответов или такой список был бы очень велик. По методике обработки данных к качественным данным можно отнести данные, полученные при ответе на вопрос со списком вариантов ответов, в котором допускается выбор нескольких возможных вариантов ответов. В более сложных ситуациях в открытых вопросах от респондента требуется выразить свое отношение или мнение.

Открытые или неструктурированные вопросы являются наиболее сложными с точки зрения компьютерной обработки анкетных данных. В отличие от закрытых, такие вопросы не содержат подсказок, не «навязывают» тот или иной вариант ответа и рассчитаны на получение неформализованного мнения. Еще чаще, чем открытые вопросы, встречаются полузакрытые вопросы, который кроме определенного числа вариантов ответа содержат позицию «другое – укажите какое (что, где, как)». Известны и иные формы открытых вопросов: «завершение предложения», «подбор ассоциации» и т.д.

Большинство исследователей не применяют компьютерную обработку открытых вопросов, используют в поисковых целях для получения информации для будущих исследований. Между тем ответы на эти вопросы могут оказаться очень информативными.

Чтобы извлечь полезную информацию, содержащуюся в открытых вопросах, необходима специальная компьютерная технология обработки данных. В данной работе представлена технология, которая является составным элементом целого программного комплекса по обработке анкетных данных, предназначенного для работы в среде EXCEL [49].

Основу разрабатываемой технологии составляет собственная модель данных, определяющая форму представления и хранения информации. Технология реализована в виде комплекса программных средств. Рассмотрим основные функции, которые выполняет компьютерная технология обработки качественных данных:

– исправление ошибок, допущенных респондентом или оператором. При наборе оператором текстовой информации количество ошибок существенно возрастает. В первую очередь это обусловлено неразборчивостью подчерка. Кроме того, одни и те же тексты допускают множество вариантов написания. Например, такие ответы о месте жительства, как «г. Владивосток», «Г. Владивосток», «г Владивосток» и «Владивосток», при компьютерной обработке будут восприниматься как различные ответы;

– переход от неструктурированных вопросов к структурированным. Такая операция сводится к выработке единообразных форм высказываний и при необходимости к обобщению отдельных высказываний. В результате должны быть получены данные, пригодные для компьютерной обработки;

– функции экспертной системы, позволяющие накапливать опыт исследователя при структурировании открытых вопросов. Такая система может быть использована самим исследователем при обработке данных в последующих опросах, а также и передаваться другим исследователям, занимающимся сходной проблемой.

Основой технологии является форма представления и хранения информации качественного характера. Данные анкетного опроса принято представлять в виде таблицы «объект-свойство». Такую таблицу легко разместить на отдельном листе EXCEL. Для данных по открытому вопросу, представленных в форме текста, используется один столбец таблицы. Мы считаем, что ответ может быть множественным. Например, отвечая на вопрос «Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купа-

ния?», респондент может выразить ответ в виде нескольких простых высказываний: «играть в волейбол, любоваться природой, ловить рыбу». Признак в таблице «объект-свойство», содержащий данные по такому вопросу, мы называем составным. Другими словами, ответ может состоять из нескольких более простых высказываний. Простые высказывания в ответах респондентов разделяются каким-либо знаком («;» или «,»). В более сложных случаях отдельные простые высказывания могут быть в виде целых предложений. В простейших случаях ответ может состоять только из одного простого высказывания. Допуская такие ответы на открытые вопросы, мы ни в чем не ограничиваем респондента. Пусть пишет ответ, как ему удобно. Никаких правил по форме ответов мы не задаем. При вводе данных оператор вводит ответы «как есть – никакой фантазии». Если оператор будет видоизменять ответ респондента, исправлять, вносить какие-то знаки, то это будет уже обработка данных. Как неквалифицированный человек произведет такую обработку, совершенно непонятно. Здесь и возникает проблема разработки компьютерной технологии обработки таких данных. Обработать тысячи наблюдений или ответов без заявленной технологии практически невозможно. Таким образом, на входе мы имеем составной признак, представленный в текстовой форме.

Теперь определим, что мы будем иметь на выходе предложенной информационной технологии обработки качественных данных. Для этого прежде рассмотрим предположения, заложенные в основу технологии. Начнем с простых высказываний – частных случаев составного признака (или свойства). При открытой форме вопроса можно было бы ожидать, что респонденты не дадут одинаковых ответов. На практике встречается достаточно много одинаковых или сходных по смыслу высказываний, не говоря уже о простых описках и орфографических ошибках. Перечень действительно различных по сути, а не по форме ответов на такие вопросы анкет ограничен. Уже при выборке порядка 700 анкет можно выделить всего от 30 до 50, по сути, различных вариантов ответов, которые можно интерпретировать как значения признака, измеренного в номинальной шкале. При увеличении объема выборки список вариантов практически не изменяется.

Для обработки данных открытых вопросов мы используем операцию типизации, т.е. замену исходного простого высказывания (в

форме текста) на близкое или сходное по значению, или обобщающее, высказывание (в форме текста). Для выполнения операции типизации формируется вспомогательная таблица «Список значений признака». Один из столбцов такой таблицы включает все уникальные значения исходного признака. Если типизации подвергается составной признак, то при расчете таблицы «Список значений признака» учитываются все простые высказывания сложного или составного высказывания. Таблица «Список значений признака» содержит столбец, в котором рассчитаны частоты встретившихся значений. Таблица снабжается автофильтром. Операция типизации применяется не к исходным данным таблицы «объект-свойство», а к данным таблицы «Список значений признака». В начале обрабатываются простые ситуации. Например, различное написание одного слова или различный порядок слов.

Среди сходных высказываний выбирается наиболее удачная (или грамотная) форма написания высказывания, затем такое высказывание копируется в ячейки таблицы «Список значений признака» со сходными высказываниями. Выполняя замену какого-то уникального высказывания на уже существующее из списка значений, мы тем самым сокращаем количество строк таблицы «Список значений признака». После выполнения серии замен целесообразно выполнять операцию «сжатия», которая заключается в пересчете таблицы «Список значений признака». Постепенно таблица «Список значений признака» сокращается и становится более наглядной. Блок-схема алгоритма типизации приведена на рис. 2.14.

После того, как простые ситуации обработаны, приступают к обработке более сложных случаев. В таблице «Список значений признака» отыскивается группа редко встречающихся, но касающихся одной темы высказываний. Для этой группы простых высказываний исследователь подбирает в таблице некоторое обобщающее высказывание и, если такого не находит, сам формулирует новое обобщающее высказывание, отражающее смысл или тему группы простых высказываний.

Например, отвечая на вопрос «Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купания?», наряду с другими ответами различные респонденты давали такие ответы: «воспитание внуков», «воспитание ребенка», «играть с внуками», «учить плавать детей».

Однако эти высказывания встречались достаточно редко (менее 0,1%), поэтому мы заменили их на обобщающее высказывание «заниматься с детьми», которое нашли в таблице «Список значений признака». В принципе, смысл высказываний сохранился.

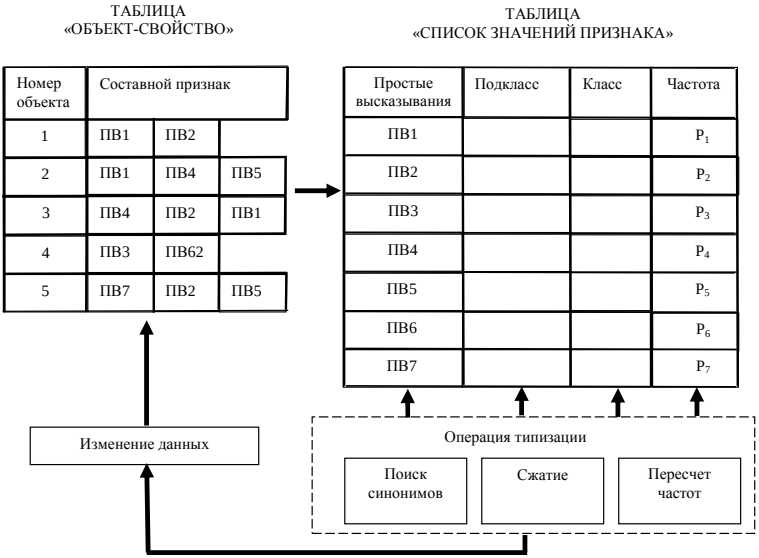


Рис. 2.14. Блок-схема алгоритма типизации

И все-таки чтобы не потерять информацию, особенно при повторном проведении опросов, мы заменяем сходные высказывания на обобщающие с уточнением. Уточнение или нюанс указывается в скобках. Например, в рассмотренном выше случае мы заменили оригинальные значения на:

– «заниматься с детьми (воспитание внуков)»,
– «заниматься с детьми (воспитание ребенка)»,
– «заниматься с детьми (играть с внуками)»,
– «заниматься с детьми (учить плавать детей)».

Для нас важнее характер ответа, который определяет тип личности респондента (потребителя), а не конкретное содержание ответа. Если исходная таблица «Список значений признака» может содержать до нескольких тысяч значений, то после обработки (типизации) такая таблица обычно содержит до трехсот зна-

чений с учетом значений с уточнениями. Созданием такой таблицы заканчивается первый этап типизации (первый уровень). Даже при автоматизации процесса работа требует достаточно много времени, опыта и большой внимательности от исследователя. И, конечно же, эта работа не может быть выполнена за один сеанс. Поэтому при завершении сеанса результаты сохраняются, и в следующем сеансе работа продолжается с того места, где она была остановлена.

Полученный новый признак содержит все еще слишком много значений, чтобы его можно было анализировать. Поэтому этот признак подвергается дополнительной обработке (второй уровень). На этом этапе просто исключаются уточнения, содержащиеся в скобках, и формируется еще один столбец таблицы «Список значений признака», который мы называем подкласс, количество уникальных высказываний в котором будет уже от 30 до 50.

Наличие 30–50 вариантов значений – тоже большое количество для анализа измерений в номинальной шкале. Поэтому исследователь после формирования приемлемого списка действительно различных вариантов ответов должен сгруппировать эти ответы, рассматривая их как некоторые характеристики непересекающихся классов, типов или тем, в зависимости от содержательного смысла признака и постановки задачи, для которой производится типизация. В нашем примере больше подходит определение типа личности. Объединение простых высказываний в классы является третьим уровнем типизации. Для каждого класса исследователь сам формулирует название по характеру объединяемых высказываний.

На практике результаты группировки у разных исследователей получаются очень похожими. Различия могут возникать из-за того, что некоторые высказывания действительно могут занимать промежуточное состояние и относиться сразу к нескольким классам. А вот названия классов каждый исследователь может дать совершенно разные.

Таким образом, в результате обработки данных открытого вопроса мы будем иметь (на выходе):

– три новых представления признака (свойства), которые включаются в исходную таблицу данных и могут быть подвергнуты дальнейшей обработке для получения содержательных выводов;

– таблицу «Список значений признака», которая может быть использована при повторении данного анкетного опроса или для выявления типизаций данных других анкет, предназначенных для исследования данного процесса.

Необходимо отметить, что в результате типизации составных признаков будут сформированы составные признаки. Для их анализа разработаны специальные методы обработки.

Технология обработки открытых вопросов содержит еще один важнейший элемент, позволяющий существенно (на порядок) уменьшить время на типизацию данных при повторных опросах (мониторинге процесса). При пополнении таблицы исходных данных необходимо опять повторять процедуру типизации с учетом новых данных. Для ускорения работы исследователь может использовать два типа словарей, которые создаются для каждого признака, содержащего данные по открытому вопросу: «Словарь замен» и «Словарь ключевых слов». Такие словари формируются для каждого отдельного качественного признака. Кроме того, при обработке данных используется еще один словарь, работающий со всеми качественными признаками и даже с различными анкетами. Это «Словарь избыточной информации». Он используется на первом этапе обработки качественной текстовой информации. С помощью него удаляются или корректируются высказывания, содержащие различную избыточную и несодержательную информацию.

Все словари хранятся в одном файле Access. Они хранят опыт, накапливаемый исследователем в процессе долговременной работы, и представляют собой базу знаний. Структуру словарей и технологию работы с ними рассмотрим позже. Технология обработки качественных данных была реализована в виде комплекса программных средств. Рассмотрим общую схему работы с разработанным комплексом программных средств (рис. 2.15).

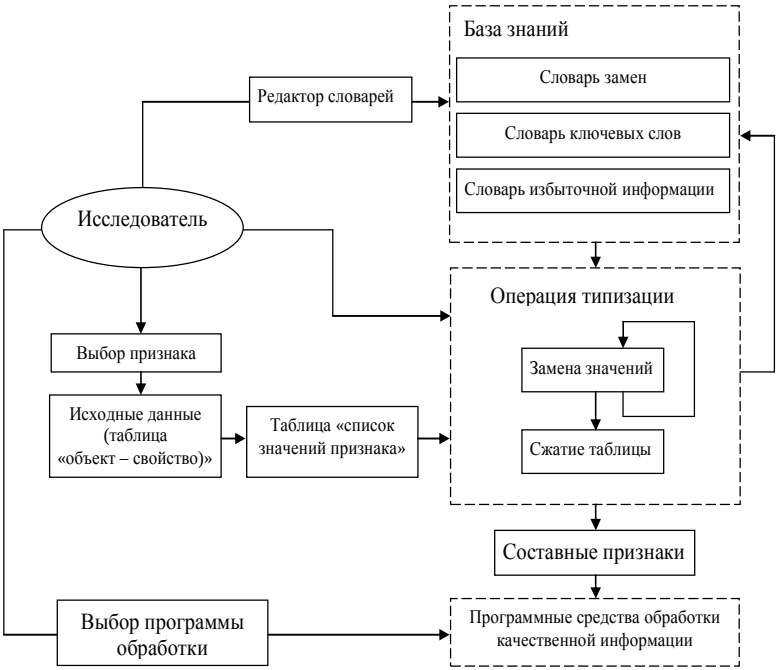


Рис. 2.15. Схема компьютерной технологии обработки качественных данных

На схеме представлены основные элементы технологии:

– таблица «Список значений признака»;

– операция типизации;

– база знаний;

– программные средства обработки качественной информации.

Информационная технология оформлена в виде комплекса программ, объединенных общим интерфейсом. Запуск программы начинается с выбора признака (рис. 2.16).

В окне указывается либо диапазон, либо столбец таблицы данных. После выбора признака назначаются параметры программы (рис. 2.17). В диалоговом окне выбора параметров пользователь должен указать разделитель – символ, позволяющий разбить составной признак на простые.

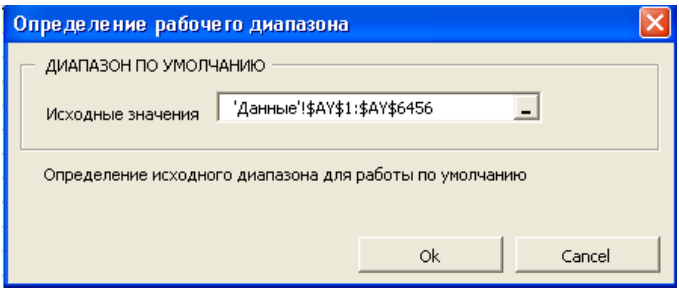


Рис. 2.16. Диалоговое окно выбора составного признака

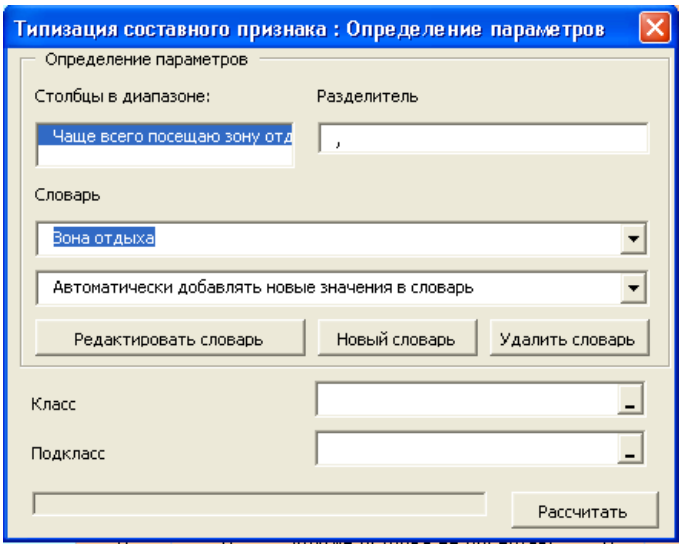


Рис. 2.17. Диалоговое окно определения параметров программы

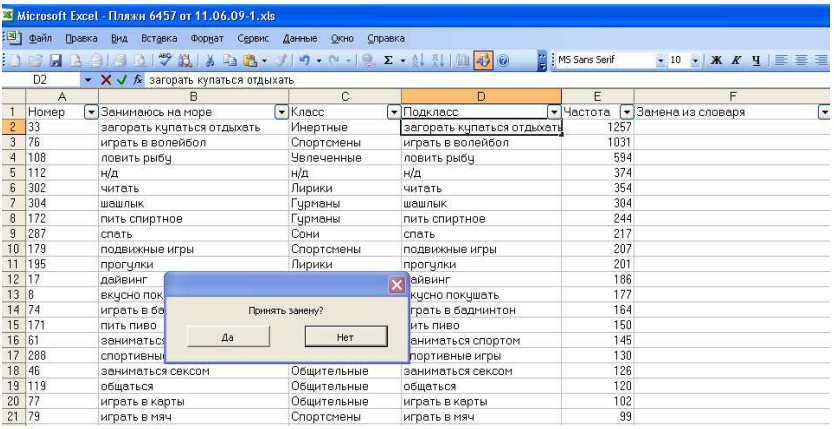
Затем устанавливается «Словарь замен» путем выбора из списка словарей замен. Выбрав словарь, пользователь может включить функцию «Автоматически добавлять новые значения в словарь» либо не подключать ее, и тогда словарь не будет пополняться новыми словами замен, возникающими в процессе работы пользователя с программой типологии.

В диалоговом окне определения параметров (рис. 2.17) пользователь может создавать новые или удалять устаревшие словари замен. Если на предыдущих этапах работы были сформированы обобщающие признаки «Класс» «Подкласс», то, указав их распо-

ложение, пользователь получит эти значения в таблице «Список значений признака».

Таким образом, ему не надо будет вводить класс в подкласс для ранее обработанных данных. При выборе функции «Редактировать словарь» вызывается программа редактирования словарей. Здесь пользователь может выбрать любой словарь и произвести его редактирование. Работа со словарями – отдельный элемент технологии, связанный с функционированием экспертной системы. Поэтому описание принципов работы со словарями будет рассмотрено в следующем параграфе.

После определения параметров пользователь должен нажать кнопку «Рассчитать», и тогда программа произведет расчет таблицы «Список значений признака». Форма таблицы приведена на рис. 2.18. После расчета таблицы «Список значений признака» программа просматривает словарь замен и, если обнаруживает в таблице ранее замененное значение, предлагает произвести замену. Все замены из словаря замен выводятся в отдельном столбце таблицы. После просмотра предлагаемых замен пользователь может принять предлагаемые замены или не принять. Отдельные предложения программы при этом могут быть скорректированы.



	A	B	C	D	E	F
1	Номер	Занимаюсь на море	Класс	Подкласс	Частота	Замена из словаря
2	33	загорать купаться отдыхать	Инертные	загорать купаться отдыхать	1257	
3	76	играть в волейбол	Спортсмены	играть в волейбол	1031	
4	108	ловить рыбу	Увлеченные	ловить рыбу	594	
5	112	н/д	н/д	н/д	374	
6	302	читать	Лирики	читать	354	
7	304	шашлык	Гурманы	шашлык	304	
8	172	пить спиртное	Гурманы	пить спиртное	244	
9	287	спать	Сони	спать	217	
10	179	подвижные игры	Спортсмены	подвижные игры	207	
11	195	прогулки	Лирики	прогулки	201	
12	17	дайвинг		айвинг	186	
13	8	вкусно пок		кусно покушать	177	
14	74	играть в ба		грать в бадминтон	164	
15	171	пить пиво		ить пиво	150	
16	61	заниматьс		аниматься спортом	145	
17	288	спортивны		портивные игры	130	
18	46	заниматься сексом	Общительные	заниматься сексом	126	
19	119	общаться	Общительные	общаться	120	
20	77	играть в карты	Общительные	играть в карты	102	
21	79	играть в мяч	Спортсмены	играть в мяч	99	

Рис. 2.18. Диалоговое окно принятия замен из «словаря замен»

Теперь пользователь может приступить к выполнению операции типизации, используя все средства самого EXCEL и допол-

нительные инструментальные средства, предоставляемые программой типизации.

Работа пользователя по типизации состоит в корректировке таблицы «Список значений признака» (рис. 2.18). Пользователь отыскивает сходные высказывания и приводит их к единому написанию. При этом весьма полезной информацией для пользователя служит столбец с частотами встречаемости отдельных высказываний в исходной таблице данных. Для поиска сходных значений пользователю приходится прибегать к всевозможным способам сортировки таблицы и всем видам фильтров, представленным в EXCEL. Использование фильтров позволяет находить высказывания, включающие похожие элементы. Например, с помощью фильтра, представленного на рис. 2.19, выделяются такие простые высказывания, как «рыбалка», «рыбалкой», «рыбачить», «рыбная ловля» и другие, которые без потери информации могут быть заменены одним высказыванием «ловить рыбу».

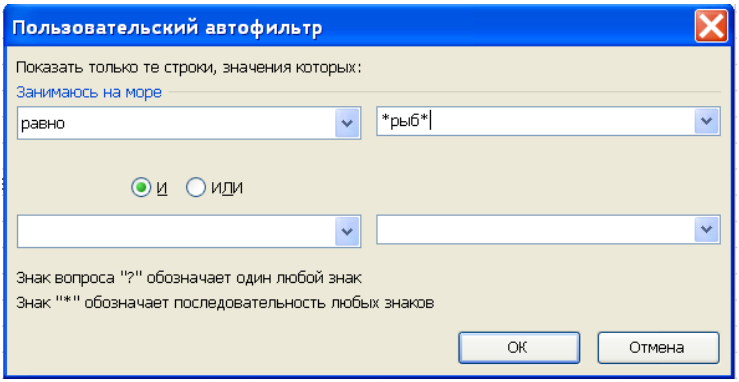


Рис. 2.19. Использование фильтров EXCEL для поиска сходных высказываний

Выполнив серию корректировок, пользователь имеет возможность сжать таблицу «Список значений признака». Для этого предумотрена операция «Сжать» диалогового окна инструментальных средств программы типизации (рис. 2.20).

При выполнении этой команды таблица «Список значений признака» сжимается – одинаковые высказывания собираются в одной строке, а частота встречаемости высказывания увеличивается.

После сжатия пользователь может продолжить работу. Если пользователь допустит ошибку или некорректные действия, то он может воспользоваться командами «Начальные значения», «Назад» «save pint», «Назад».

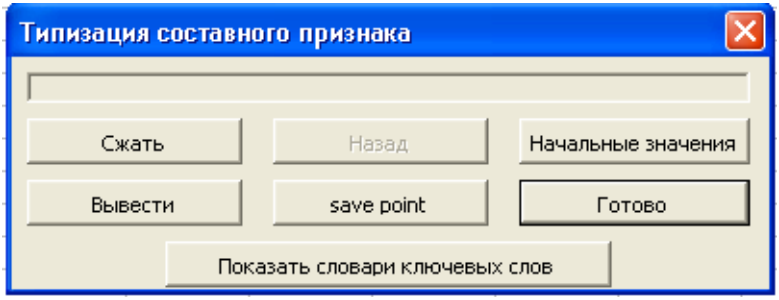


Рис. 2.20. Диалоговое окно выбора инструментальных средств, используемых при выполнении операции типизации

Команда «Начальные значения» отменяет все корректировки таблицы «Список значений признака» и выводит ее в первоначальном виде. Команда «save pint» сохраняет текущее состояние таблицы, тогда после ряда некорректных действий пользователь может вернуться к сохраненной таблице «Список значений признака», воспользовавшись командой «Назад». При окончании сеанса работы с программой типизации необходимо выполнить команду «Готово». В процессе работы пользователь может использовать «словарь ключевых слов». Работа со всеми видами словарей рассматривается далее.

По завершении работы с программой выводится диалоговое окно с вариантами представления результатов работы программы (рис. 2.21).

Пользователь имеет возможность заменить значения в таблице исходных данных на новые значения из таблицы «Список значений признака», полученные в результате выполнения операции типизации. Обновленные значения могут быть размещены на том же листе, что и таблица «Список значений признака». Если пользователь определял обобщающие высказывания в соответствующих столбцах таблицы «Список значений признака» («класс», «подкласс»), то полный развернутый список можно тоже вывести и разместить на указанных листах файла EXCEL.

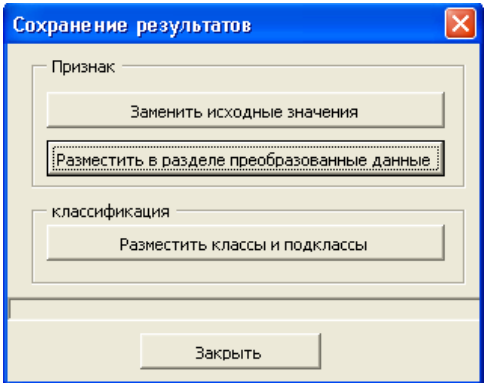


Рис. 2.21. Диалоговое окно вывода результатов работы программы

При больших объемах таблицы исходных данных операция разработки типологии может занять очень много времени. Для повышения эффективности работы пользователя при обработке данных массовых анкетных опросов применяют различные словари, которые хранят знания, накапливаемые в процессе исследования. Формирование базы знаний и методика использования их в процессе обработки данных рассмотрена в следующей главе.

По завершении работы с программой типизации можно выполнить расчеты с помощью различных программ анализа структуры типологий.

2.3.3. Повышение эффективности обработки качественных данных на основе экспертной системы

Повышение эффективности работы компьютерной технологии обработки качественных данных достигается за счет создания и использования базы знаний. Компьютерные технологии, позволяющие использовать базы знаний, относятся к классу экспертных систем. Главная особенность экспертной системы заключается в умении делать правильные предсказания. На рисунке 2.22 приведена схема основных элементов интеллектуальной системы обработки данных [50].

Экспертная система реализуется в виде специального программного средства. Производя всевозможные подсказки пользователю во время его работы, специальные программные средства существенно сокращают время работы пользователя.

Подсказки пользователю производятся с помощью специальных словарей. Формирование словарей происходит в процессе работы пользователя над задачей типизации. Словари хранят опыт пользователя, приобретенный им в процессе решения задач типизации качественных признаков.

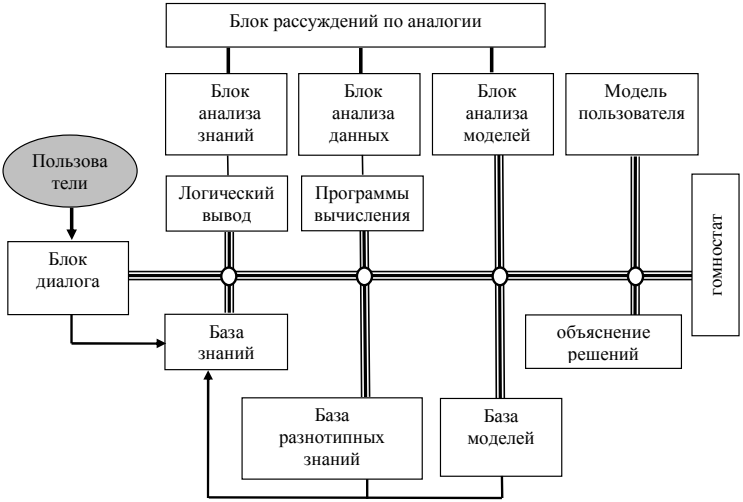


Рис. 2.22. Основные элементы интеллектуальной системы обработки данных

«Словарь замен» формируется автоматически при выполнении пользователем замен одних простых высказываний другими. Словарь пополняется при работе пользователя с программой типизации и хранит все совершенные замены. При сборе новых анкетных данных приходится опять производить операцию типизации, т.е. пользователь должен подбирать аналоги для новых данных. Оказывается, что при повторном сборе информации ситуации, обработанные пользователем на предыдущих этапах, в подавляющем числе случае повторяются. Тогда, подключив «Словарь замен», пользователь получит подсказки по заменам и ему останется только обработать ситуации, которые ранее не встречались. Рассмотрим структуру и функции используемых словарей.

Сопровождение словаря практически не требует дополнительных затрат времени. Периодически целесообразно его просматривать и редактировать. Со временем в словаре начинают накапливаться неактуальные варианты замен. Такие записи необходимо удалять, потому что при очень больших объемах словаря скорость работы программы типизации снижается. Пример «Словаря замен» представлен на рис. 2.23.

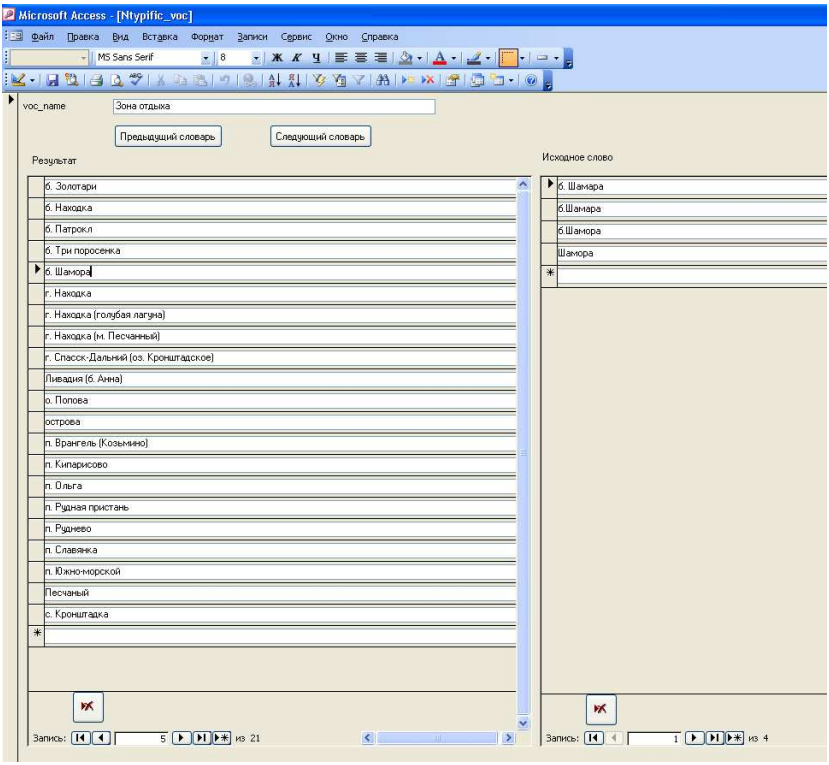


Рис. 2.23. Словарь замен

Рассмотрим функции следующего словаря – «Словаря ключевых слов». Этот словарь оказывается очень полезен, когда фразы содержат много слов. При длинных фразах поиск подходящих синонимов в таблице «Список значений признака» очень затруднен, потому что в начале работы список значений очень велик. Полное совпадение фраз встречается крайне редко. Поэтому очень часто приходится прибегать к фильтрам EXCEL, чтобы выделять более

короткие списки, содержащие определенные сочетания отрывков фраз (рис. 2.24).

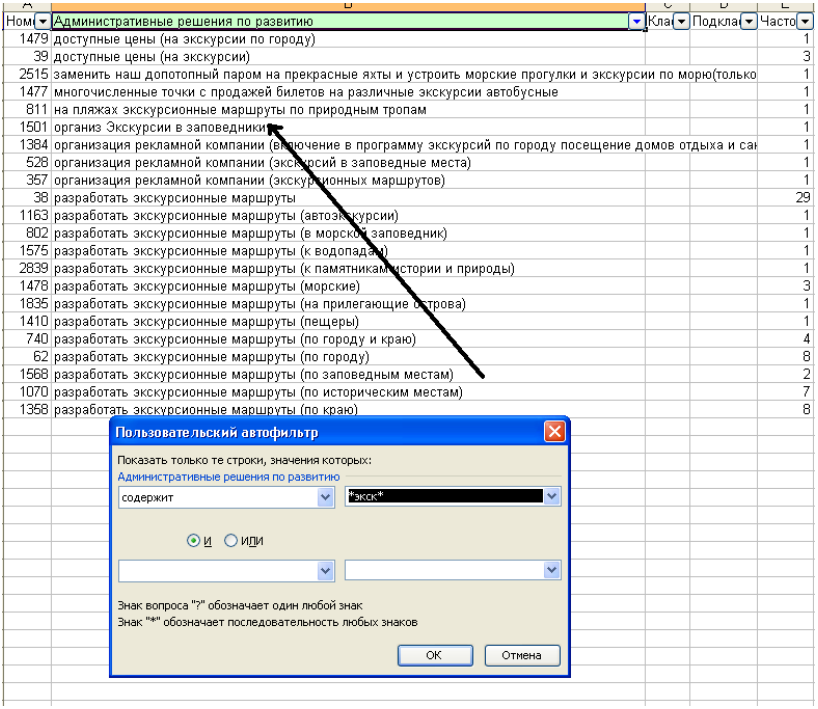


Рис. 2.24. Пример работы с пользовательским фильтром EXCEL при выполнении операции типизации длинных фраз

Такие фразы мы называем ключевыми словами, хотя отрывки фраз ключевыми словами можно назвать условно. В таблице на рис. 2.24 выделены фразы по ключевому слову «экск». В таком сокращенном списке фразе «организ Экскурсии в заповедники» можно легко обнаружить подходящую замену «разработать экскурсионные маршруты (по заповедным местам)». При такой замене смысл фразы совершенно не искажается.

При длительной работе по типизации фраз пользователь накапливает опыт по формулировке ключей. Однако через какой-то промежуток времени опыт утрачивается, и его опять приходится восстанавливать, что ведет к большим непроизводительным затратам времени. Обработка длинных фраз имеет еще одну осо-

бенность. Поскольку полных совпадений высказываний возникает гораздо меньше, чем при обработке простых фраз, приходится гораздо чаще подбирать обобщающие термины-синонимы. При этом для сходных высказываний приходится и чаще использовать уточнения, которые указываются в скобках.

Эти две особенности обработки длинных фраз были учтены при разработке структуры «Словаря ключевых слов» и технологии работы с ним. «Словарь ключевых слов» имеет две функции, облегчающие работу исследователя.

Сначала рассмотрим структуру «Словаря ключевых слов», а затем его функции. Словарь состоит из четырех взаимосвязанных таблиц. Формы четырех таблиц в Access приведены на рис. 2.25.

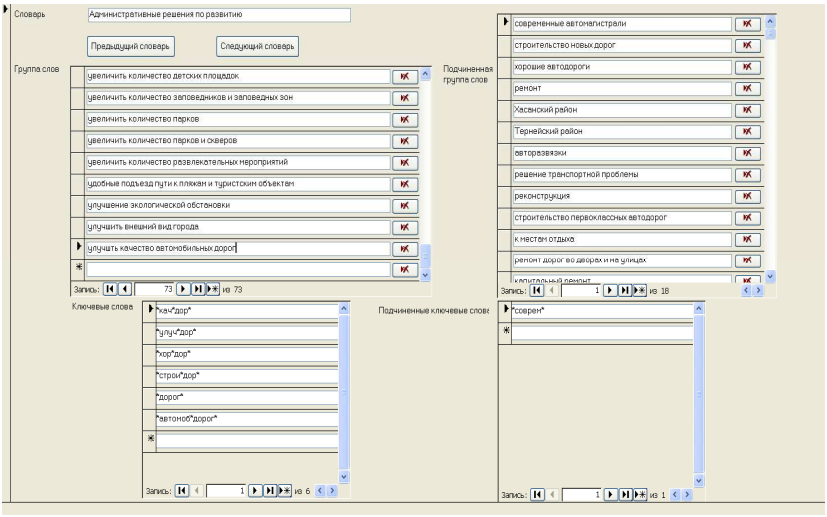


Рис. 2.25. Формы таблиц «Словаря ключевых слов» в Access

В первой (основной) таблице представлены часто встречающиеся фразы-синонимы без уточнения в скобках. Во второй таблице для каждой записи основной таблицы содержатся уточнения, которые приводятся в скобках. В третьей таблице содержатся ключевые слова для фраз из основной таблицы. В четвертой таблице содержатся ключевые слова к уточнениям для основных фраз, которые дополняют ключевые слова основной фразы.

«Словарь ключевых слов» составляется самим пользователем. Вносить данные можно как в процессе выполнения операции типизации, сохраняя повторяющиеся ключи, используемые в фильтрах, так и в специальном редакторе. В первом случае ввод данных выполняется пользователем с помощью команд специального диалогового окна (рис. 2.26), которое открывается по команде «показать словари ключевых слов» в основном диалоговом окне программы типизации (рис. 2.20).

В «Словарь ключевых слов» заносятся фразы из таблицы «Список значений признака», имеющие высокую встречаемость (частоту). Вносимые фразы представляют собой некоторые устойчивые конструкции, которые с достаточной степенью уверенности войдут в окончательный вариант таблицы «Список значений признака».

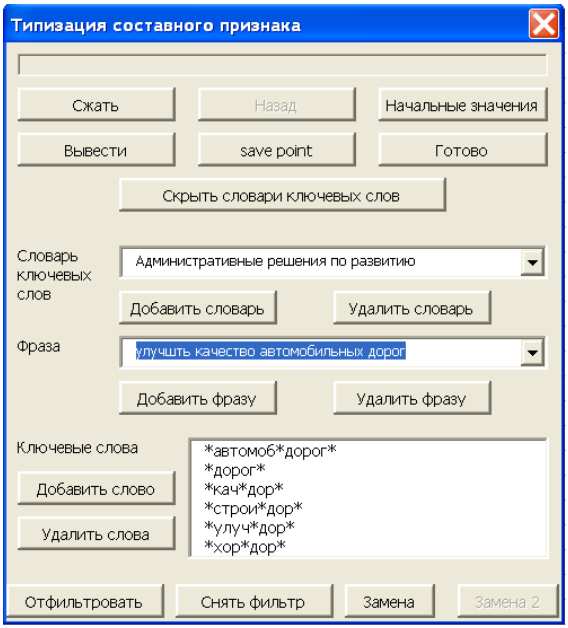


Рис. 2.26. Набор инструментальных средств для работы со «Словарем ключевых слов»

Окончательный вариант представляет собой разработанную типологию. Для этих фраз вносится список ключевых слов для поиска сходных фраз, которые необходимо включить в типологию.

На рисунках 2.27 и 2.28 представлены диалоговые окна с целью пополнения «Словарей ключевых слов». В словарь вносятся фразы, не имеющие уточнений в скобках. Все фразы с уточнениями вносятся в словарь автоматически. Такая функция редактора существенно упрощает работу со словарем, поскольку слова с уточнениями встречаются намного чаще.

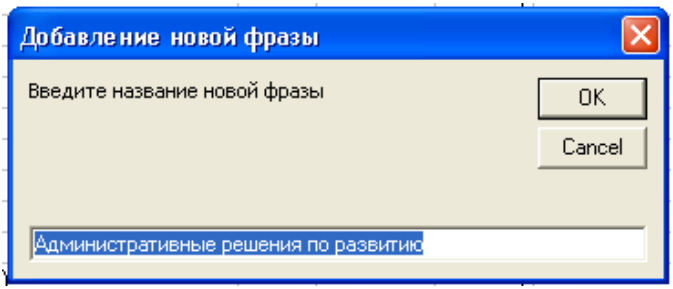


Рис. 2.27. Ввод новых фраз в «Словарь ключевых слов»

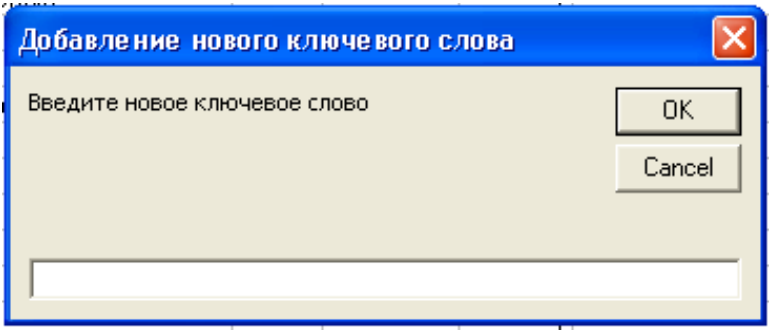


Рис. 2.28. Ввод новых ключевых слов в «Словарь ключевых слов»

Все словари можно редактировать в автономном режиме. Вызов редактора всех словарей осуществляется специальной программой (рис. 2.29).

Рассмотрим функции «Словаря ключевых слов».

Первая функция состоит в использовании ключевых слов, внесенных в словарь для высокочастотных фраз для формирования фильтров с целью поиска аналогичных устойчивых фраз. Отличие работы словаря от работы расширенного фильтра состоит в том, что используемые ранее для поиска подходящих синонимов

ключевые слова сохраняются в словаре, а не формулируются самим пользователем. Словарь создается для облегчения работы в будущем. Когда он будет содержать достаточное количество данных, то может быть использован как база знаний.

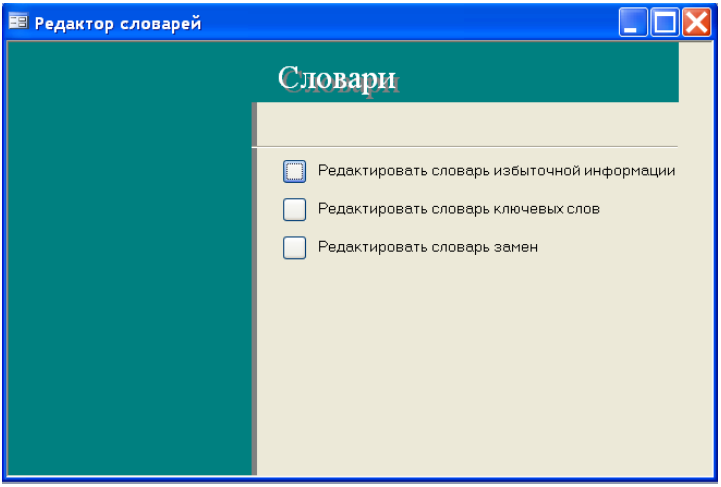


Рис. 2.29. Диалоговое окно вызова редактора словарей

Вторая функция обладает интеллектуальными свойствами. Эта функция вызывается с помощью команды «Замена» (рис. 2.26). Программа просматривает строки таблицы «список значений признака» и для фраз, которые не представлены в «Словаре ключевых слов», подбирает подходящую фразу из словаря. Подбор фразы осуществляется по максимальному количеству совпадений фрагментов анализируемой фразы с ключевыми словами фраз, представленных в «Словаре ключевых слов». Предлагаемые варианты замен выводятся в дополнительном столбце таблицы «Список значений признака» – «Замены из словаря ключевых слов».

Третий словарь базы знаний – «Словарь избыточной информации», он работает со всеми качественными признаками и даже с различными анкетами. Словарь используется на первом этапе обработки качественной текстовой информации. С помощью этого словаря удаляются или корректируются высказывания, содержащие различную избыточную и несодержательную информацию.

Например, с помощью словаря могут быть исключены такие словосочетания, как «Я думаю, что», «По моему мнению», «Это, в свою очередь» и т.п. В этот словарь включаются и слова с типовыми обобщениями. Например, «о», «о.» в данных может быть всегда заменяться на слово «остров». «Словарь избыточной информации» оказывается очень полезным при обработке длинных фраз и предложений. С помощью этого словаря удается существенно сократить таблицу «Список значений признака» на первом этапе работы. С фрагментами фраз, внесенными в «Словарь избыточной информации», можно совершать операции замен, если фраза «Содержит», «Начинается на», «Заканчивается на» или «Точное значение». При выполнении операций имеет значение порядок их выполнения. Поэтому в программе редактора словарей можно изменять порядок записей в словаре, сдвигая отдельные фразы вверх и вниз (рис. 2.30).

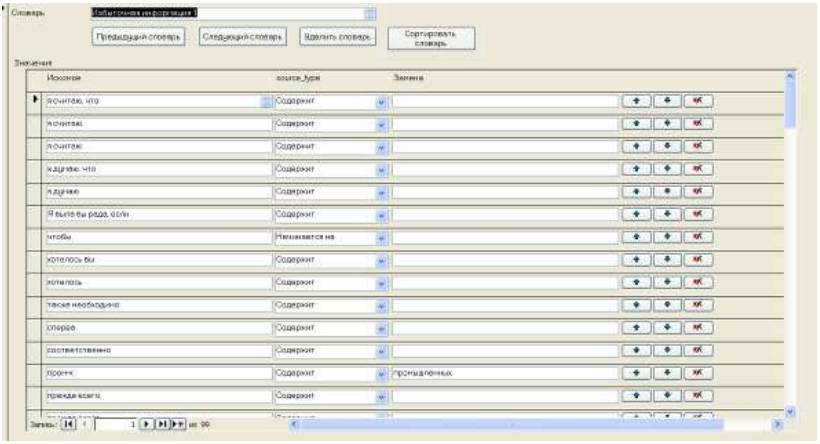


Рис. 2.30. Форма таблицы «Словарь избыточной информации

2.4. Автоматизация обработки при мониторинге социально-экономических процессов

2.4.1. Инструментальные средства обработки данных при мониторинге социально-экономических процессов

В настоящее время актуальной является проблема исследования взаимодействия участников социально-политических и трудовых

процессов, протекающих на уровне региона. Одним из наиболее эффективных средств их исследования служит мониторинг, осуществляемый на основе независимых анкетных опросов жителей, проживающих на территории региона.

Термин «мониторинг» английского происхождения. В России стал употребляться во второй половине XX века в значении «постоянное наблюдение за каким-либо процессом, цель которого – выявление соответствия наблюдаемого процесса стандартам, желаемым результатам, первоначальным предположениям». Мониторинг – это важнейший инструмент оценки эффективности принятия управленческих решений и оценки устойчивости состояния систем под воздействием внешних факторов. В зависимости от назначения мониторинга и уровня социально-экономической системы, в отношении которой он применяется, изменяются структура мониторинга, порядок его проведения и состав исследуемых показателей [51].

Профессиональные анкетные опросы, как правило, проводятся с некоторой периодичностью. Поэтому у исследователей возникает необходимость в выполнении расчетов не по всем данным, а по отдельным группам данных.

Анализ данных мониторинга с использованием различных анкет вызывает ряд трудностей технического характера. Рассмотрим некоторые дополнительные средства анализа данных мониторинга, позволяющие повысить эффективность работы исследователей. Предлагаемые средства предназначены для обработки данных, которые собираются с помощью анкетных форм, а затем экспортируются в EXCEL для их дальнейшей обработки.

Для обработки данных мониторинга было разработано два программных модуля, позволяющих производить анализ данных признаков различной природы.

Программные модули включены в состав комплекса программных средств, являющегося надстройкой к EXCEL. При их разработке были учтены некоторые требования, предъявляемые ранее разработанным комплексом программ [52].

Прежде чем перейти к описанию функций программных модулей, необходимо проанализировать процесс сбора данных в мониторинговых исследованиях и определить связанные с ним понятия.

Сначала определимся с формой представления данных. Будем считать, что данные, полученные в результате опроса, заносятся в некоторую таблицу, которую принято называть «объект-свойство». Главной особенностью сбора данных в мониторинговых исследованиях является то, что в каждой анкете должна быть указана дата сбора информации, которая при компьютерной обработке заносится в отдельный столбец таблицы данных.

В реальных исследованиях ежегодно по одной анкете может опрашиваться значительное количество респондентов (от нескольких сотен до нескольких тысяч). При сборе данных могут быть задействованы множество технических работников, которые непосредственно контактируют с респондентами.

Сбор данных осуществляется в течение некоторого периода (иногда до нескольких месяцев). Считается, что за время сбора данных изменений состояния исследуемой социально-экономической системы не происходит. Социально-экономические системы вообще достаточно инерционны.

Необходимо различать понятия «периоды сбора данных» и периоды, на который распространяются полученные оценки состояния системы – «периоды оценки состояния системы». В большинстве случаев оценки распространяются на календарный год, но могут быть назначены и другие временные рамки.

Данные от технических работников, осуществляющих сбор анкетных данных, для ввода в компьютер поступают неравномерно. Каждая анкета должна иметь свой уникальный номер, по которому ее можно идентифицировать.

При сборе данных на бумажном носителе затрачивается достаточно много времени на ввод данных в компьютер. В этой работе может участвовать несколько операторов. Обычно для удобства ввода данных по конкретной анкете создается специализированная компьютерная форма, облегчающая работу оператора. После ввода в компьютер все данные собираются в единую таблицу и упорядочиваются по номеру анкеты, что очень удобно при сверке компьютерных данных с оригиналом на бумажном носителе с целью обнаружения ошибок. Данные на бумажном носителе сохраняются на весь период, пока существует перспектива сбора информации по данной анкете.

Допустимо считать, что в конечном итоге будет сформирована некоторая таблица EXCEL, данные которой и нужно обработать. Разработанный программный комплекс требует, чтобы «Лист EXCEL», где располагается таблица «объект-свойство», имел название «Данные». Никаких прочих данных или расчетов на листе «Данные» быть не должно, кроме данных таблицы «объект-свойство».

Рассмотрим методику расчета и интерфейс программных модулей.

Прежде чем приступить к анализу данных, полученных в процессе мониторинга, желательно проанализировать процесс сбора данных. На первом этапе желательно представить процесс сбора данных в виде диаграммы, например, по месяцам или кварталам. Для расчета графика сбора данных был разработан специальный программный модуль в среде EXCEL – «Календарный график сбора данных». Интерфейс обращения к программе «Календарный график сбора данных» представлен на рис. 2.31.

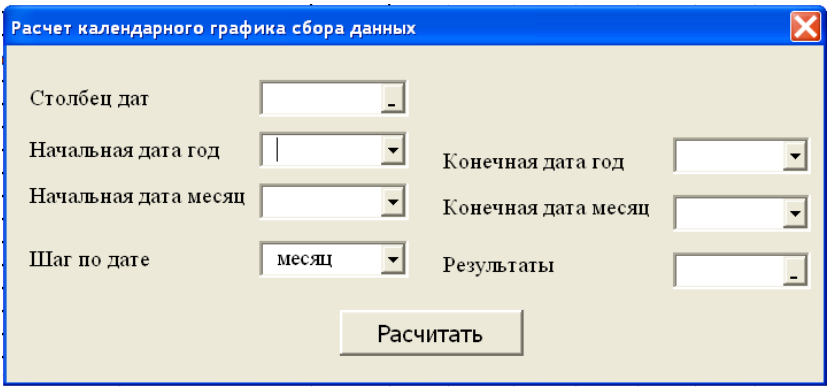


Рис. 2.31. Интерфейс программы «Календарный график сбора данных»

Рассмотрим содержание параметров, которые задаются при обращении к программе:

– «столбец дат» – ссылка на столбец таблицы данных, содержащий даты сбора анкет;

– «начальная дата» – дата, начиная с которой производится расчет графика сбора данных (выбирается из выпадающего списка);

– «конечная дата» – дата, до которой производится расчет графика сбора данных (выбирается из выпадающего списка);

– «шаг по дате» – размер интервалов, по которым рассчитывается календарный график сбора данных (год, квартал, месяц – выбирается из выпадающего списка);

– «результаты разместить» – ячейка листа EXCEL, выбранная для размещения таблицы результатов.

Для демонстрации результатов работы мы выбрали анкету «Исследование времяпрепровождения отпусков жителями Приморского края». Пример таблицы результатов, рассчитанных с помощью программы «Календарный график сбора данных» по этой анкете, представлен на рис. 2.32, а календарный график, построенный по данным этой таблицы на рис. 2.33. Из таблицы и графика на рис. 2.33 следует, что данные анкет в период наблюдений собрались неравномерно.

R	S	T	U	V	W
Номер интервала	Год	Начальная дата интервала	Конечная дата интервала	Количество анкет	Надпись на оси графика
1	2008	01.07.2008	31.07.2008	89	июль. 08
2	2008	01.08.2008	31.08.2008	78	авг. 08
3	2008	01.09.2008	30.09.2008	39	сен. 08
4	2008	01.10.2008	31.10.2008	281	окт. 08
5	2008	01.11.2008	30.11.2008	258	нояб. 08
40	2011	01.10.2011	31.12.2005	212	окт. 11
41	2011	01.11.2011	31.03.2006	292	нояб. 11
42	2011	01.12.2011	30.06.2006	354	дек. 11

Рис. 2.32. Результаты, полученные с помощью программы «Календарный график сбора данных»

Анализ графика процесса сбора данных по анкетному опросу позволяет спланировать работу по решению задач мониторинга исследуемого процесса или явления. График дает информацию о том, каким статистическим материалом располагает исследователь к моменту решения задач мониторинга.

До обработки данных мониторинга необходимо сопоставить даты сбора данных и периоды оценки состояния системы. Например, для анкеты «Исследование времяпрепровождения отпусков жителями Приморского края» сбор данных обычно начинался в конце календарного года и продолжался в течение двух-трех

месяцев, а иногда и более. Для выполнения расчетов необходимо определить временные интервалы сбора данных для каждого периода оценки состояния системы.

При исследовании времяпрепровождения отпусков было выбрано четыре периода сбора данных и четыре периода оценки состояния системы. Соответствие дат сбора анкетных данных и периодов оценки состояния системы представлено в табл. 2.4. Очевидно, периоды сбора данных и периоды оценки системы не совпадают.

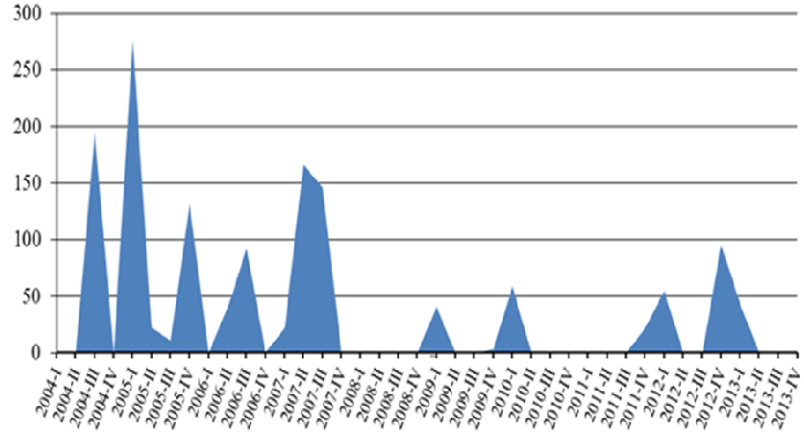


Рис. 2.33. Пример временной диаграммы сбора статистических данных

Расчет частотных рядов какого-либо исследуемого признака (свойства) по периодам оценки состояния системы производится с помощью специального модуля EXCEL «Мониторинг частотных рядов». Интерфейс программы представлен на рис. 2.34.

В программе необходимо определить столбец таблицы данных, в котором размещены даты сбора анкет. Затем определяется таблица соответствия «период-дата», в которой задаются даты начала и конца периодов сбора данных (даты сбора данных в табл. 2.4). Эти данные задаются на любом листе EXCEL, кроме листа «Данные». «Диапазон данных» – это ссылка на один из столбцов таблицы данных, в котором представлены данные исследуемого признака.

Таблица 2.4

Таблица соответствия периодов оценки состояния системы периодам сбора анкетных данных

Номер периода	Периоды оценки состояния системы	Даты сбора данных		Количество анкет
		Начальная дата периода	Конечная дата периода	
1	2008 г.	01.07.2008	31.08.2009	1494
2	2009 г.	01.09.2009	31.08.2010	1510
3	2010 г.	01.09.2010	31.08.2011	1345
4	2011 г.	01.09.2011	31.08.2012	863
Итого		01.07.2008	31.08.2012	5212

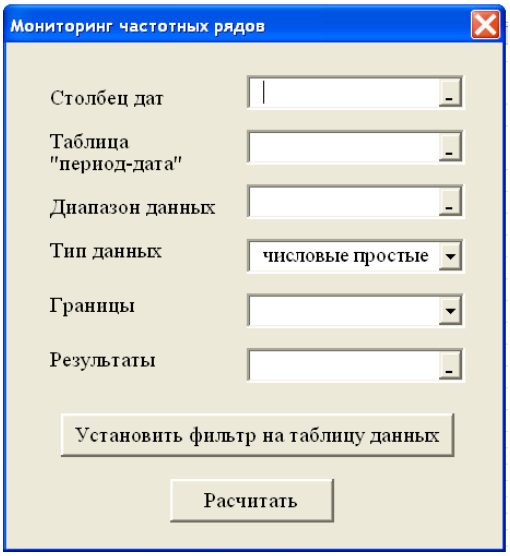


Рис. 2.34. Интерфейс программы «Мониторинг частотных рядов»

С помощью этой программы можно обрабатывать различные типы данных, которые могут встретиться при проведении анкетных опросов. В настоящее время в программе предусмотрена обработка трех типов данных: «числовые данные», «простые текстовые», «составные текстовые». Для различных типов данных используется своя схема

расчета. Поэтому при обращении к программе необходимо задать тип данных. Сначала рассмотрим обработку числовых данных. Под этот тип подпадают и ранговые данные.

Далее в программе определяется диапазон ячеек, в которых размещаются «границы интервалов» для расчета частотного ряда исследуемого признака. Считается, что интервалы следуют один за другим, поэтому задаются только верхние границы интервалов. Самая верхняя граница включается в соответствующий интервал и не входит в следующий интервал. Примером задания верхних границ может служить первый столбец таблицы результатов (рис. 2.35). В этой таблице приведены результаты обработки данных признака по затратам времени на занятие спортом студентов, полученные с помощью программы «Мониторинг частотных рядов». Параметр «результаты разместить» определяет ячейку листа EXCEL, выбранную для размещения рассчитанной таблицы (рис. 2.35).

R	S	T	U	V	W
129					
130	Верхние границы интервалов	Периодов оценки состояния системы			Итого
131		1	2	3	
132		01.07.2004 — 31.12.2006	01.01.2007 — 31.12.2009	01.01.2010 — 31.12.2012	
133	0	205	87	29	321
134	1	72	38	18	128
135	3	240	130	41	411
136	5	95	51	16	162
137	7	67	21	16	104
138	9	15	15	7	37
139	11	26	15	2	43
140	13	12	5	1	18
141	15	11	8	2	21
142	17	5	0	0	5
143	56	18	6	3	27
144	Итого	766	376	135	1277

Рис. 2.35. Результаты работы программы «Мониторинг частотных рядов»

При анализе данных мониторинга может возникнуть необходимость построить серию частотных рядов не по всем данным, а по данным, отвечающим условиям для некоторой группы респондентов, например, только для женщин или только для мужчин. Для определения условий, описывающих такую группу, служит кнопка «установить фильтр на таблицу данных». По этой команде вызыва-

ется стандартный «пользовательский автофильтр» EXCEL, настраиваемый пользователем.

При обработке текстовых данных границы интервалов не задаются. Список уникальных текстовых значений для построения частотного ряда определяется программой. Для того чтобы избежать получения громоздких таблиц, на промежуточной стадии программа сообщает количество найденных уникальных текстовых значений. Исследователь должен определить, нужно ли продолжать расчеты или необходимо произвести дополнительные преобразования по унификации данных (сокращение количества уникальных вариантов текстовых ответов).

Кроме простых текстовых данных программа рассчитывает частотные ряды для составных текстовых данных. Составной признак возникает тогда, когда при ответе на вопрос респондент может указать сразу несколько ответов. Например, при ответе на вопрос анкеты «Какие крупные города вы посетили за последние три года?» респондент может указать сразу несколько городов. Таким образом, составной признак включает несколько простых ответов. Для идентификации составного признака в компьютерном представлении вводится какой-либо единый знак разделителя. Простой ответ может состоять из нескольких слов или даже может быть сформулирован в форме целого предложения. При обработке более сложных случаев необходимо производить предварительную обработку составных признаков, производя переход от неструктурированных данных к структурированным. Методика обработки таких данных представлена в работах [53, 54].

Предлагаемая технология предназначена для массовых исследований и при большом количестве исследуемых признаков. В простых случаях отдельные задачи мониторинга можно решить и с помощью стандартного набора инструментальных средств EXCEL (например, с использованием «сводных таблиц»). В различных ситуациях обработки мониторинговых данных придется сделать серию дополнительных операций, формировать промежуточные данные и т.п.

Рассмотренные модули являются составной частью целого комплекса специализированных программных средств обработки анкетных данных. Специализация программных модулей на определенном классе задач позволяет повысить эффективность работы

исследователя, существенно сокращая затраты времени на обработку данных. Рассмотренная технология прошла апробацию при обработке данных целой серии массовых социологических опросов, проводимых на территории муниципальных образований Приморского края силами научных подразделений Владивостокского государственного университета экономики и сервиса.

2.4.2. Мониторинг системы администрирования организации бизнес-структур региона

В последние годы в научной литературе все больше внимания уделяется повышению эффективности управления муниципальными образованиями региона [55–58]. Региональные условия накладывают отпечаток на общую проблему. Это и большая разница в экономических условиях, демографической ситуации, географическом положении, климатических условиях. Поэтому большинство исследователей рассматривают проблему через призму своего региона. Особенно неблагоприятная ситуация складывается на Дальнем Востоке, что усиливает миграционные процессы [59]. На огромной территории Приморского края (площадь края превосходит размеры средней европейской страны) проживает менее двух миллионов человек. Население крайне неравномерно распределено. В зоне влияния владивостокской агломерации находится более двух третей населения края.

В Приморском крае особенно остро проблема эффективности управления стоит в сельских муниципальных образованиях. Молодежь не задерживается в сельской местности. Накопившиеся территориальные проблемы и недостаток финансов не оправдывают безынициативность в использовании имеющихся в настоящее время возможностей развития. Перспектива развития муниципальных образований напрямую зависит от активности и мобильности субъектов малого бизнеса, который практически остановился в своем развитии на большей части территории края. При формировании стратегии развития муниципальных образований должны быть задействованы все возможности по консолидации различных структур местного сообщества [60].

В целях изучения проблемы эффективности администрирования организации бизнес-структур на территории муниципальных образований края нами был предпринят опрос населения

муниципальных образований. Опрос был проведен в декабре 2012 года (опрошено 1600 респондентов) и декабре 2013 (опрошено 1060 респондентов). Иными словами, полный объем выборки составил 2660 анкет. Рассмотрим некоторые результаты проведенного исследования.

Для оценки возможностей организации бизнеса респондентам был предложен вопрос: «Дайте оценку возможности организации собственного бизнеса на территории вашего муниципального образования». Респонденты могли дать ответ по 15 градациям (табл. 2.5).

Таблица 2.5

Форма таблицы, используемой для ответа на ранговые вопросы

← очень плохо отлично →														
1			2			3			4			5		

Сгруппированные данные по различным категориям населения представлены на рис. 2.36. Опросы подтвердили более низкие возможности организации бизнеса в сельской местности. Оценили возможности как «плохо» и «очень плохо» 58% сельских жителей.

Более благоприятно оценивают обстановку в небольших приморских городах (неудовлетворенность решением данной проблемы высказали 34% населения). В краевой столице Владивостоке обстановка несколько хуже (41%). Последнее можно объяснить более жесткой конкуренцией в сфере бизнеса.

Интерес представляет исследование причин, препятствующих открытию бизнеса на территории различных муниципальных образований. Многие из причин обсуждаются в научных публикациях, и можно было бы привести список вариантов для оценки их рейтингов. Но мы не хотели навязывать мнение респондентам и поэтому задали вопрос в открытой форме: «Какие причины, на ваш взгляд, более всего препятствуют частному предпринимательству на территории вашего муниципального образования?». Таким образом, мы предоставили респонденту полную свободу в своих высказываниях. Однако при таком подходе возникает необходимость обработки больших массивов

качественных данных с целью их систематизации и структурирования. Для решения этой задачи нами была разработана специальная информационная технология обработки качественных данных, которую мы постоянно совершенствуем. Принципы работы технологии изложены в работах [61, 62].

Рис. 2.36. Частотные ряды распределения оценок по возможности организации собственного бизнеса

Компьютерному анализу были подвергнуты 2660 анкет (378 респондентов уклонились от ответа или дали неинформативный ответ). Общее количество ответов на вопрос составило более 4 тысяч предложений, то есть в среднем каждый респондент давал немногим менее 2 предложений на заданный вопрос. Данные прошли два этапа обработки. На первом этапе было выделено 1520 различных предложений ответа на вопрос, содержание ответов практически не изменялось (корректировке подвергалась только форма ответа). На втором этапе близкие по содержанию предложения были сгруппированы в 40 групп ответов.

После перехода от неструктурированных данных ответов респондентов к структурированным можно рассчитать частоту встречаемости групп ответов по всем ответам респондентов. Обычно мы объединяем группы ответов в классы. Объединение производится по близости смысловой нагрузки влияния на изучаемый объект или проблему. При этом мы добиваемся наглядности результатов, но

при формировании классов (или типологий) допускается некоторая субъективность группировки. В данном случае явных типологий выделить не удалось, поэтому для наглядности мы разделили группы ответов на три класса по частоте встречаемости. В первый класс было объединено 8 вариантов ответов (групп). Это самая представительная группа, а, следовательно, и более важная, вобрала 68% всех предложений. Вторая группа из 16 предложений вобрала около 24% всех предложений и третья из 16 предложений объединила в своем составе 8% предложений. Частотные ряды распределения ответов в трех различных классах ответов приведены на рис. 2.37–2.39. При анализе значимости различных причин, препятствующих организации бизнеса, необходимо учитывать, что оценки рассчитываются относительно классов. Другими словами, относительная значимость причин, объединенных в первый класс, несопоставимо больше значимости любой причины второго или третьего класса. Соотношение причин, препятствующих организации и развитию бизнеса, можно пояснить на примере. Так, первые четыре причины первого класса имеют долю в 67% в своем классе и 45% среди всех прочих причин. На первое вышло бремя налогового законодательства, респонденты ставят его даже несколько выше, чем наличие стартового капитала для организации бизнеса (рис. 2.37).

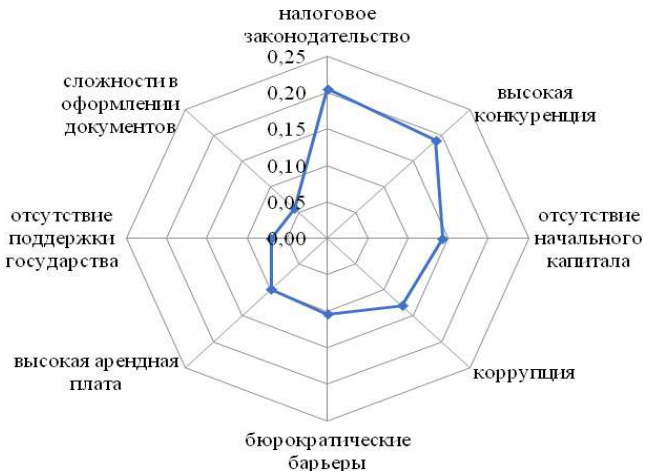


Рис. 2.37. Распределение ответов респондентов по частоте указания причин, препятствующих организации бизнеса (первый класс)

Во втором классе присутствует еще одна причина, связанная с законодательством по регулированию бизнеса («несовершенство законодательства»), но она менее значима чем причины, объединенные в первый класс (но более значимая, чем все прочие причины в третьем классе). Коррупция ставится по важности близко к бюрократическим барьерам, хотя эти причины идут рука об руку и вместе сильно перекрывают все остальные препятствия экономического плана. А ведь эти причины в чистом виде генерируются властными структурами и надзорными органами, т.е., если по существу наладить контроль органов власти муниципальных образований, можно существенно поднять уровень развития предпринимательства. Таким образом, администрация муниципальных образований должна, прежде всего, начать работу по развитию бизнеса с наведения порядка внутри себя. Необходимо вернуть доверие людей к органам власти. Причины, включенные в состав второго класса, не имеют ярких выбросов, то есть они приблизительно равнозначны (рис. 2.38).

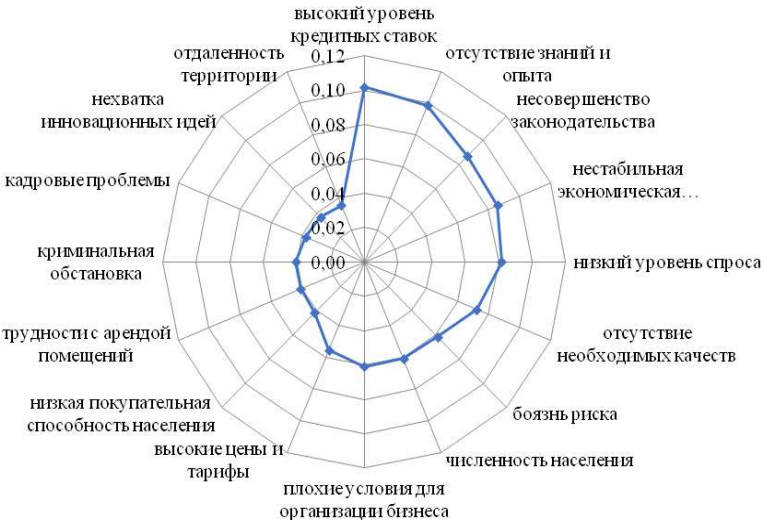


Рис. 2.38. Распределение ответов респондентов по частоте указания причин, препятствующих организации бизнеса (второй класс)

В этом классе выделяется причина «высокий уровень кредитных ставок». Таким образом, подтверждается утверждение о том, что высокие процентные ставки сдерживают деловую активность бизнеса. Обещанные льготы для малого бизнеса на Дальнем Востоке пока не дают положительного результата.

Интересным результатом является высокая оценка значимости показателя «отсутствие знаний и опыта». Этому вопросу не уделяется должного внимания со стороны администрации края. Между тем решение его не столько требует высоких затрат, сколько больше лежит в организационной плоскости и доступности информации.

Большое беспокойство вызывает показатель «низкий уровень спроса», что объясняется как низкой покупательской способностью населения, так и неразвитостью окружающей бизнес-среды. Заметим, что низкий уровень зарплат выступает одной из главных причин, побуждающих население Дальнего Востока покидать регион [58].

В третьем классе ряд причин близок по смыслу к причинам, объединенным в два другие класса (рис. 2.39).

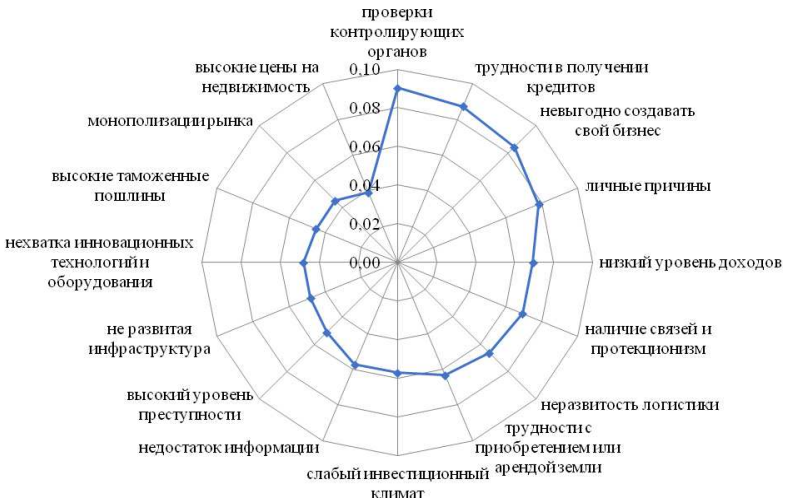


Рис. 2.39. Распределение ответов респондентов по частоте указания причин, препятствующих организации бизнеса (третий класс)

Например, группа ответов «криминальная обстановка» близка по смыслу к группе ответов «высокий уровень преступности», а группа ответов «бюрократические барьеры» не далека от группы «проверки контролирующих органов».

Однако, если проанализировать вопросы, входящие в близкие группы, можно заметить некоторые отличия в содержании ответов. Возможно, в дальнейшем мы разработаем более лаконичную классификацию, а на данном этапе мы хотели подчеркнуть разнообразие ответов.

Мы считаем проблему структурирования данных как важнейшую. При выработке управленческих решений можно по-разному объединять группы ответов, возможно даже в пересекающиеся классы.

Подобные исследования относительно инновационного малого бизнеса представлены в работе [63]. На основе эмпирического исследования в ней рассматриваются проблемы развития малого инновационного бизнеса на примере 70 компаний, расположенных на территории Новосибирской области. Основное внимание сфокусировано на трех аспектах: что мешает развитию малых фирм в инновационной сфере, каковы факторы их успеха, какие формы поддержки инновационного бизнеса являются предпочтительными. Наши исследования во многом созвучны с данной работой и согласуются с предложенными выводами. В настоящее время мы тоже проводим обследование предприятий малого бизнеса Приморского края, используя собственные методики.

Наши исследования показали, что население муниципальных образований, давая неудовлетворительные оценки возможности организации бизнеса на их территории, часто не дают положительных оценок работе администрации муниципального образования в целом (рис. 2.40). Это утверждение подкрепляется анализом данных ответа на следующий вопрос: «Дайте оценку вашей удовлетворенности работой администрации вашего муниципального образования». Для ответа предлагалась 15-разрядная шкала (табл. 2.5).

Наибольшие претензии к работе администрации муниципальных образований имеют жители сельской местности. И здесь дело не только в том, что администрации плохо справляются со своими обязанностями или имеют низкое финансовое обеспечение. В малых поселениях администрация больше на виду, а в крупных

образованиях промахи администрации теряются на уровне социально-экономических проблем регионального характера или страны в целом.

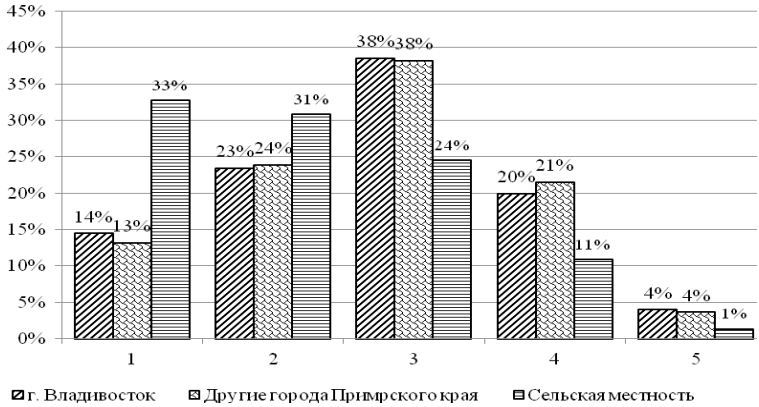


Рис. 2.40. Частотные ряды распределения оценок удовлетворенности работой администрации муниципальных образований на территории Приморского края

Полноценное и самостоятельное развитие муниципальных образований невозможно без создания благоприятной бизнес-среды. Поэтому многие исследователи, рассматривая принципы формирования стратегии социально-экономического развития муниципального образования, во главу угла ставят развитие малого и среднего бизнеса [64–67].

Решить проблему повышения доверия к органам власти и ответственности руководителей муниципальных образований можно путем повышения информатизации общества и внедрения электронного правительства [68]. Многие административные услуги могут быть оказаны через сеть. Активно стартовавший процесс оказания услуг через сеть в последнее время стал тормозиться. В настоящее время настоятельно необходим поиск новых путей использования компьютеризации для взаимодействия населения и органов власти. Внедрение компьютерных технологий для создания средств информационно-аналитической поддержки должно повысить качество стратегического планирования на всех уровнях управления региона [69]. На Дальнем Востоке технические средства во многом отстают от центральных регионов, а тем более от сопредельных государств.

Для развития экономики края с низкой плотностью населения необходимо в кратчайшие сроки повысить мобильность населения, расширяя скелет дорог и обеспечивая строительство высокоскоростных железнодорожных магистралей. Привлечение на эти работы иностранной рабочей силы вполне оправдано, поскольку работа носит временный характер и может быть решена вахтовым методом. Тем более, в недалеком будущем возможно удорожание рабочей силы, привлекаемой сегодня к дорожному строительству.

Крупные проекты требуют больших инвестиций. Такие инвестиции могли бы прийти из сопредельных государств АТР. Деловые круги этих стран сдерживает слишком большая разница условий ведения бизнеса в нашей стране и странах АТР. Проекты по созданию СЭЗ на Дальнем Востоке так и не были реализованы. Сегодня в правительстве говорят о возможности создания офшорной зоны на Дальнем Востоке. Но конкретных решений все-таки нет. Между тем время работает против нас. Мы говорим о крупных проектах, поскольку малый бизнес в регионах с низким спросом должен идти в кильватере крупного бизнеса.

Многие исследователи считают, что для организации конкурентоспособного малого бизнеса предпринимателям не хватает специальных знаний, которые могут распространяться через специальные региональные структуры [70, 71].

Одним из направлений по созданию благоприятных условий для предпринимательства считается механизм государственно-частного партнерства. Однако эффективных моделей такого партнерства пока в нашей стране не создано [72]. По мнению автора, налаживание государственно-частного партнерства возможно на пути создания профессиональных объединений предпринимателей, отражающих интересы определенных кругов малого бизнеса. Такие «гильдии» могли бы не только представлять бизнес, но и взять на себя некоторые контролирующие функции стандартов качества, с которыми у нас сейчас очень большие проблемы. По мнению ряда международных экспертов и отечественных авторов, большие перспективы малого предпринимательства в Приморском крае связаны с развитием туристского бизнеса [73]. В этом направлении необходимо расширять кооперацию с соседними регионами и сопредельными государствами [74].

В работе получены количественные оценки, которые могут быть полезны при разработке программ развития муниципальных

образований Приморского края. В настоящее время сбор данных в этом направлении продолжается, что позволит в будущем давать оценки эффективности некоторых решений по созданию благоприятных условий для формирования предпринимательской среды. Некоторые разработанные типологии могут быть полезны для анализа ситуации в других регионах. В дальнейшем разработанные технологии могут найти применение при налаживании обратной связи в рамках программ электронного правительства.

2.5. Автоматизация разработки когнитивных моделей

В настоящее время Россия ставит перед собой крупные задачи по преобразованию экономики. Решение проблем развития страны затруднено мощным внешним противодействием. Ситуация усложняется тем, что уровень доверия населения ко всем уровням власти страны достаточно невысок. Основными причинами недоверия к властным структурам являются высокий уровень коррупции, низкая ответственность представителей власти перед обществом, проникновение во властные структуры большого количества людей, ставящих свои интересы выше общественных, огромный разрыв между богатыми и бедными. Повысить уровень доверия к властным структурам могут налаживание социального партнерства и привлечение широких масс населения регионов к обсуждению проблем и выработке управленческих решений. Процесс социального партнерства представляет собой процесс согласования принимаемых решений с представителями различных социальных групп, проживающих в регионах (предпринимателей, трудящихся, молодежи и пожилых людей).

Новая стратегия развития России ориентируется на расширение экономических и политических связей со странами Азиатско-Тихоокеанского региона. Поэтому ускоренное развитие Сибири и Дальнего Востока в последнее время является одним из важнейших приоритетов правительства страны. Научный анализ государственной политики развития Дальнего Востока и Байкальского региона приводится в работах известных российских экономистов [75–78]. Как отмечает П.А. Минакир, сегодня для развития Дальнего Востока «недостаточно просто институциональных новаций,

нужны масштабные инвестиции … в инфраструктуру коммунальную, социальную и инфраструктуру рынка» [77, с. 9].

Для решения глобальных задач по развитию Дальнего Востока в новых экономических условиях потребуются мощные трудовые ресурсы. Поэтому проблема сохранения населения и привлечения квалифицированной рабочей силы имеет наивысший приоритет. Решение этой проблемы связано с повышением качества жизни населения. Без ее решения невозможна эффективная трудовая мотивация. Эффективная трудовая мотивация зависит от социального самочувствия населения, проживающего на территории региона.

Социальное самочувствие – эмоциональный аспект оценки представителями социальной группы своего общественного положения, уровня удовлетворения социально-экономических и духовных потребностей, интересов. В социальном самочувствии выражается обобщенная оценка общественных настроений группы: экономических, политических, идеологических, национальных и др. [79–81].

Для повышения эффективности управленческих решений, направленных на улучшение социального самочувствия населения, в настоящей работе предлагается использовать когнитивный подход. Для построения когнитивной модели предварительно необходимо произвести анализ проблем, влияющих на качество жизни и социальное самочувствие населения, проживающего в регионе.

Рассмотрим оценку населением региона проблем, влияющих на качество жизни на примере Приморского края. Информационной базой создания когнитивных моделей выработки решений социально-экономических проблем региона могут служить мнения, сформированные в результате обсуждения проблем в экспертных группах, инновационных семинарах, опросах населения региона.

Во Владивостокском государственном университете экономики сервиса более десяти лет проводится мониторинг социально-экономических процессов в регионе. Ежегодно по различным анкетам проводится опрос 2–3 тысяч жителей Приморского края. Для оценки проблем, влияющих на качество жизни населения региона, в ряд анкет был включен следующий вопрос с фиксированным списком ответов: «Укажите три наиболее важных для вас фактора, влияющих на качество жизни». Список из 13 факторов, влияющих на качество жизни, был определен на предварительных этапах исследования (пробные опросы).

Отвечая на вопрос анкеты, респондент должен был отметить только три из представленных вариантов ответа. Данные ответов на вопросы можно представить в виде матрицы бинарных значений (содержащих значения 0 или 1). Строки матрицы соответствуют опрошенным респондентам (объектам), а столбцы – предложенным вариантам ответов (свойствам). Сумма единичных значений в матрице равна количеству анкет, умноженному на три (каждая строка матрицы содержит три единицы).

По данным ответов респондентов были рассчитаны индексы значимости факторов качества жизни по формуле:

$$I_j = \frac{r_j}{3n_0}, \qquad (2.10)$$

где I_j – индексы значимости факторов, влияющих на качество жизни, $j = \overline{1,k}$;

k – количество факторов в представленном списке факторов;

r_j – количество единичных значений в j -м столбце матрицы данных;

n_0 – объем выборки.

Необходимо отличать подход оценки факторов (проблем), влияющих на качество жизни, используемый в данной работе, от подхода, основанного на интегрированных показателях оценки качества жизни, используемого для расчета официальных оценок качества жизни [82, 83].

На рисунке 2.41 приведены результаты расчетов индексов значимости факторов, указанных в списке вопроса в качестве возможных вариантов ответов. Оценки рассчитаны по данным анкетных опросов за четыре года (2016, 2017, 2018, 2019). За указанный период с помощью этого вопроса было опрошено свыше 4 тысяч респондентов.

На первое место жители Приморского края ставят «возможность трудоустройства». На втором месте показатель «будущее детей», что свидетельствует о большой неуверенности жителей в завтрашнем дне. Здравоохранение и продолжительность жизни – взаимосвязанные показатели и вместе они выходят на первое место. Кроме того, на здоровье оказывают влияние качество питания и окружающая среда. Как видно из диаграммы, приоритеты факторов качества жизни, определенные населением, со временем несколько изменяются.

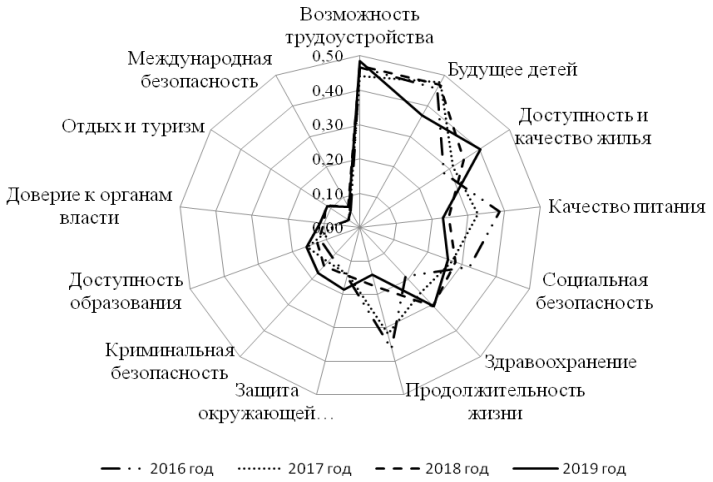


Рис. 2.41. Оценка значимости факторов, определяющих качество жизни, жителями Приморского края по временным периодам

Более заметные различия в оценке факторов качества жизни наблюдаются у респондентов различных возрастных категорий (рис. 2.42).

Рис. 2.42. Оценка значимости факторов, определяющих качество жизни, жителями Приморского края по возрастным категориям

Представительство респондентов различных возрастных категорий в проведенных опросах оценивается следующим образом: до 30 лет – 48%; 31–55 лет – 46%; более 55 лет – 6%. Возможность трудоустройства в различных возрастных категориях занимает лидирующую позицию. Более зрелых людей (старше 55 лет) больше беспокоят проблемы здравоохранения и качества питания. Средние возраста (от 31 до 55 лет) волнует проблема будущего детей.

Улучшить качество жизни можно путем преодоления проблем, которые давят на человека. Для выявления структуры проблем, которые более всего беспокоят жителей края, в анкету был включен следующий открытый вопрос: «Назовите социально-экономические, политические и другие проблемы, которые более всего беспокоят вас и окружающих людей». Важно заметить, что в рассматриваемом исследовании при ответе на открытый вопрос респонденты выражали свое мнение в произвольной текстовой форме.

Вопрос по проблемам региона был использован в двух типах анкет. Опросы по этим анкетам производились в период с октября 2013 года по апрель 2019 года. Распределение количества анкет по годам приведено на рис. 2.43. Всего за этот период было опрошено 6718 жителей Приморского края.

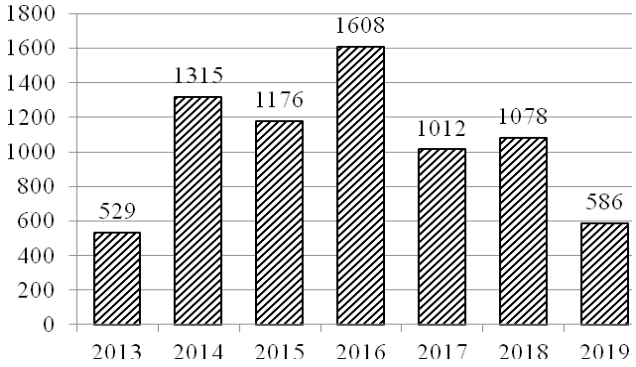


Рис. 2.43. Распределение опрошенных респондентов по годам

Всего в ответах респондентов было указано 18 885 проблем, которые беспокоят население края. Многие варианты ответов повторялись или были достаточно близки по форме и содержанию. Для обработки качественных данных была использована специаль-

ная компьютерная технология, описанная в работе [84]. Компьютерная технология позволяет автоматизировать работу исследователя по типологии данных качественных ответов. Типологизация проводится в три этапа:

– объединение одинаковых по форме и содержанию ответов (уникальные значения);

– объединение близких по смыслу ответов (формирование подклассов);

– объединение близких по тематике ответов (формирование классов).

В результате обработки первичных данных было выделено 5240 уникальных вариантов ответа. По всем вариантам ответов были сформированы 80 групп (подклассов) вариантов ответов, включающих очень близкие по смыслу ответы. Выделенные группы ответов были объединены в 18 близких по тематике классов ответов.

Первые два этапа обработки практически не приводят к искажению первичной информации. Третий этап объединения данных носит субъективный характер и зависит от понимания и интерпретации проблем самим исследователем. Различные исследователи по одним и тем же данным могут давать различные классификации. Но результаты получаются достаточно близкими по структуре (за исключением названий классов, которые присваиваются самим исследователем). Отличия в структуре классов могут возникать за счет того, что некоторые группы ответов занимают пограничные положения и могут быть отнесены и к одному классу, и к другому. Сформированная в результате обработки первичных данных структура вариантов ответов населения Приморского края по проблемам, которые их больше всего беспокоят, представлена в табл. 2.6.

Наибольшая частота встречаемости ответов относится к классу «Ценовая политика», объединившему 10 групп ответов респондентов (рис. 2.44). Большинство жителей Приморского края отмечают остроту жилищной проблемы. Жители края считают, что цены на коммунальные услуги слишком высокие.

Проблема качества жизни, по мнению населения края, является второй по значимости после проблем с ценами. В этом классе проблем с большим отрывом лидируют ответы, связанные с качеством медицинского обслуживания (рис. 2.45). Ответы «плохое качество ме-

дицинского обслуживания» и «высокие цены на медицинские услуги» в этой группе составляют более половины (52%).

Таблица 2.6

Структура данных ответов населения Приморского края по проблемам, которые их больше всего беспокоят

№	Классы ответов	Частоты классов
1	Ценовая политика	0,18
2	Низкое качество жизни	0,11
3	Забота о будущем поколении	0,1
4	Инфраструктура и благоустройство	0,09
5	Низкая заработная плата	0,06
6	Социальная напряженность и бедность	0,06
7	Коррупция и преступность в органах власти	0,06
8	Экологические проблемы и природоохрана	0,06
9	Трудоустройство и занятость	0,05
10	Низкая социальная защищенность	0,05
11	Неудовлетворенность работой властных структур	0,05
12	Преступность и личная безопасность	0,05
13	Нестабильность экономической ситуации	0,03
14	Несовершенное законодательство	0,02
15	Доступность отдыха и развлечений	0,01
16	Международная обстановка	0,01
17	Низкое качество товаров и услуг	0,01
18	Глобальные катаклизмы и катастрофы	0,003

При анализе конкретной проблемы можно обратиться к таблицам, в которых содержится расшифровка по каждой группе ответов, т.е. просмотреть конкретные ответы респондентов.

В реальных ситуациях управления социально-экономическими процессами очень часто возникает задача, состоящая не в том, чтобы сделать выбор между альтернативными решениями с целью оптимизации, а в том, чтобы проанализировать ситуацию для выявле-

ния реальных проблем и определения причин их появления [85, 86, 87]. Понимание проблемы – обязательное предварительное условие нахождения приемлемого решения.

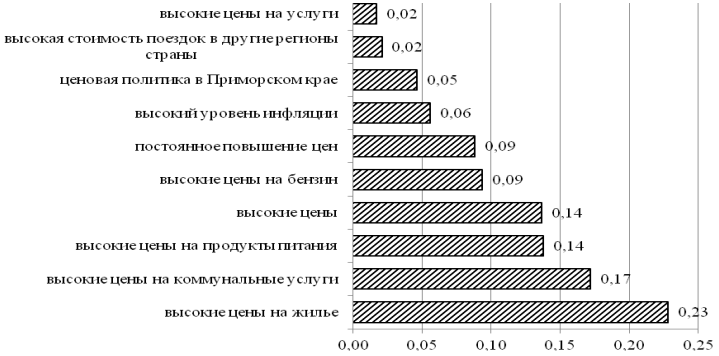


Рис. 2.44. Частотный ряд распределения ответов жителей Приморского края в классе ответов «Ценовая политика»

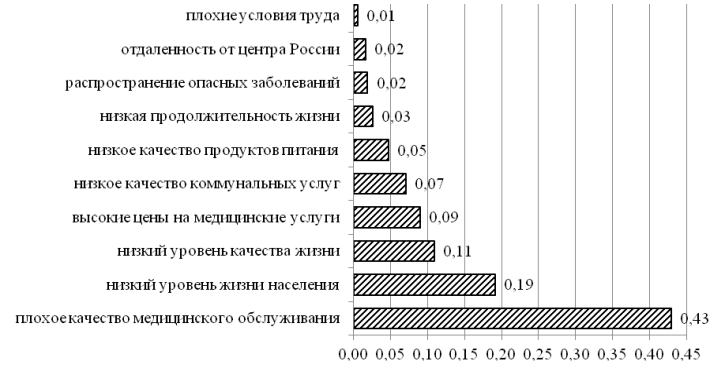


Рис. 2.45. Частотный ряд распределения ответов жителей Приморского края в классе ответов «Низкое качество жизни»

Большинство социально-экономических проблем не существуют независимо друг от друга, а находятся во взаимодействии. Оценить взаимосвязь проблем и определить стратегию их комплексного решения можно с использованием когнитивного подхода.

В последние годы наблюдается возрастание интереса научного сообщества к развитию методов принятия решения при управлении, основанных на принципах когнитивного подхода. Когнитивный подход основан на анализе и логическом обосновании процессов «восприятия, мышления, познания, объяснения и понимания» систем управления различной природы. Когнитивное моделирование нашло применение при разработке управленческих решений в экономических, социологических, экологических и других плохоформализуемых слабоструктурированных системах. Когнитивный подход используется в решении проблем понимания естественного языка, компьютерного перевода, теории искусственного интеллекта, компьютеризации всех сфер общественной деятельности [88, 89].

В работе [90] приведены главные особенности слабоструктурированных системам:

– многоаспектность происходящих в них процессов (экономических, социальных и т.п.) и их взаимосвязанность; в силу этого невозможно вычленение и детальное исследование отдельных явлений – все происходящие в них явления должны рассматриваться в совокупности;

– отсутствие достаточной количественной информации о динамике процессов, что вынуждает переходить к качественному анализу таких процессов;

– изменчивость характера процессов во времени.

Очень часто при анализе таких систем приходится учитывать десятки факторов различной природы.

В силу высокой неопределенности в исследовании процессов, влияющих на качество жизни, ряд исследователей использовали когнитивный подход [91–93].

Производство новых знаний в рамках когнитивного подхода требует разработки компьютерных технологий представления, хранения, обработки, интерпретации новых знаний [94]. В настоящей работе предлагается компьютерная технология автоматизации построения когнитивной модели на примере анализа взаимодействия проблем, влияющих на качество жизни населения региона.

Для оценки взаимного влияния проблем качества жизни к опросам были привлечены студенты экономических специальностей Владивостокского государственного университета экономики и сервиса. Это вызвано тем, что студенческая аудитория обладает достаточным уровнем современных знаний в области экономической теории и более под-

готовлена к оценке сложных социально-экономических категорий. Экспертный опрос проводился в два периода: октябрь 2018 года; октябрь 2019 года. Всего в опросе участвовало около 250 студентов.

В рамках опроса студентам предлагалось дать экспертные оценки взаимодействия проблем, влияющих на качество жизни населения региона, заполнив форму, представленную в табл. 2.7.

Таблица 2.7

Форма таблицы экспертных оценок взаимодействия проблем, влияющих на качество жизни населения региона

№	Проблемы	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Продолжительность жизни	≡												
2	Защита окружающей среды и экология		≡											
3	Качество питания			≡										
4	Доступность и качество жилья				≡									
5	Социальная безопасность					≡								
6	Криминальная безопасность						≡							
7	Возможность трудоустройства							≡						
8	Доступность образования								≡					
9	Будущее детей									≡				
10	Здравоохранение										≡			
11	Доверие к органам власти											≡		
12	Отдых и туризм												≡	
13	Международная безопасность													≡

В каждой строке таблицы необходимо было отметить три позиции (ячейки), соответствующие проблемам, которые получат положительный импульс при разрешении проблемы, указанной в строке таблицы. Например, отметка позиции в третьей строке («Качество питания») первого столбца будет означать, что респондент считает, что повышение показателя «Качество питания» положительно отразится на проблеме «Продолжительность жизни».

Необходимо различать термины «влияние» и «зависимость» проблем. Например, если проблема «Качество питания» очевидно оказывает влияние на показатель «Продолжительность жизни», то показатель «Продолжительность жизни» не обладает очевидным влиянием на проблему «Качество питания». Рассматриваются только позитивные влияния проблем. В силу последнего замечания матрица данных по оценке влияний проблем качества жизни (табл. 2.7) не будет симметричной.

Суммирование данных всех таблиц, заполненных респондентами, дает частотные ряды согласованности мнений по оценке влияния рассматриваемых проблем.

При детальном анализе проблем можно усмотреть какое-то взаимное влияние всех проблем, представленных в таблице. И здесь нет ничего удивительного, поскольку все исследуемые проблемы объединяет одна тема – качество жизни. Например, решение любых проблем, влияющих на качество жизни населения, повышает доверие к органам власти. Поэтому основная задача экспертного опроса – выделить наиболее существенные влияния. Тогда ситуация становится пригодной для анализа.

Очевидно, что при выделении существенных влияний проблем единого мнения у респондентов быть не может. Кроме того, в данных оценок респондентов присутствует составляющая, которая зашумляет оценки. Не все респонденты хорошо представляют себе рассматриваемые проблемы и постановку задачи. Респонденты отличаются своим кругозором и познаниями окружающей социальной действительности, не всегда ответственно относятся к самому опросу. В качестве примера частотного ряда распределения экспертных оценок, представленных опрошенными студентами, рассмотрим частотный ряд влияния проблемы качества питания (рис. 2.46).

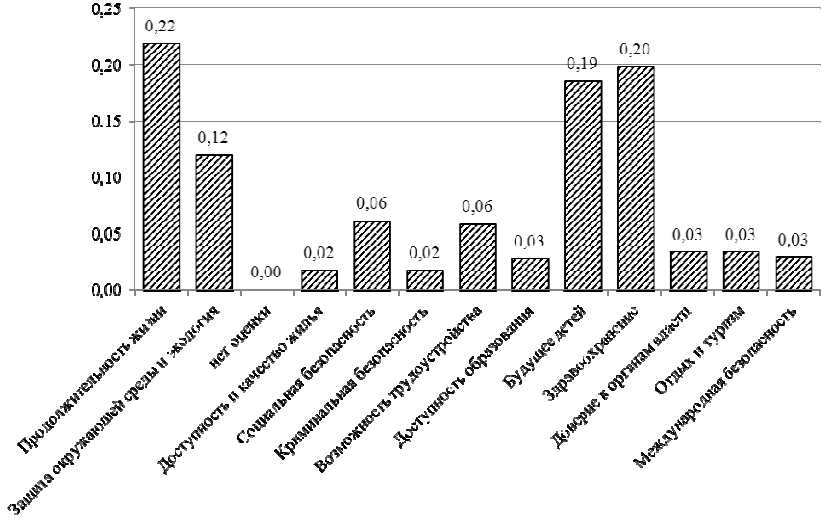


Рис. 2.46. Частотный ряд влияния проблемы качества питания на прочие проблемы

Из диаграммы видно, что по оценкам респондентов улучшение качества питания более всего ведет к снижению остроты трех проблем: «продолжительность жизни», «будущее детей», «здравоохранение». Все остальные оценки находятся в пределах погрешности и могут рассматриваться как шум. Все частотные ряды оценок влияния проблем сведены в таблицу (табл. 2.8). Данные табл. 2.8 используем для построения когнитивной модели. Для этого назначаем некоторое пороговое значение частоты оценки для отбора существенных влияний. Пороговое значение подбирается экспериментально. При снижении порогового значения в модель начинают попадать малозначимые влияния, которые только затрудняют анализ проблемы в целом. На рисунке 2.47 представлена когнитивная модель при пороговом значении 0,155.

Таблица 2.8

Частотные ряды встречаемости связей в ответах респондентов

№	Частота связи в ответах респондентов												
	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13
1	0,00	0,12	0,15	0,04	0,10	0,06	0,06	0,02	0,16	0,20	0,03	0,03	0,02
2	0,16	0,00	0,13	0,03	0,06	0,03	0,01	0,02	0,16	0,11	0,05	0,20	0,05
3	0,22	0,12	0,00	0,02	0,06	0,02	0,06	0,03	0,19	0,20	0,03	0,03	0,03
4	0,12	0,03	0,02	0,00	0,18	0,09	0,13	0,05	0,14	0,04	0,16	0,02	0,03
5	0,13	0,03	0,02	0,05	0,00	0,15	0,05	0,07	0,11	0,08	0,19	0,03	0,09
6	0,16	0,02	0,01	0,03	0,13	0,00	0,04	0,03	0,16	0,05	0,20	0,04	0,13
7	0,06	0,02	0,07	0,13	0,15	0,04	0,00	0,14	0,14	0,04	0,12	0,07	0,02
8	0,04	0,02	0,02	0,08	0,19	0,05	0,17	0,00	0,24	0,04	0,10	0,02	0,03
9	0,14	0,12	0,08	0,05	0,14	0,07	0,06	0,10	0,00	0,13	0,06	0,01	0,04
10	0,23	0,08	0,13	0,02	0,11	0,04	0,03	0,02	0,15	0,00	0,09	0,05	0,04
11	0,04	0,02	0,01	0,07	0,25	0,22	0,05	0,04	0,06	0,05	0,00	0,02	0,18
12	0,18	0,15	0,05	0,02	0,09	0,07	0,07	0,02	0,07	0,09	0,04	0,00	0,16
13	0,08	0,06	0,01	0,02	0,18	0,18	0,02	0,03	0,08	0,04	0,18	0,11	0,00

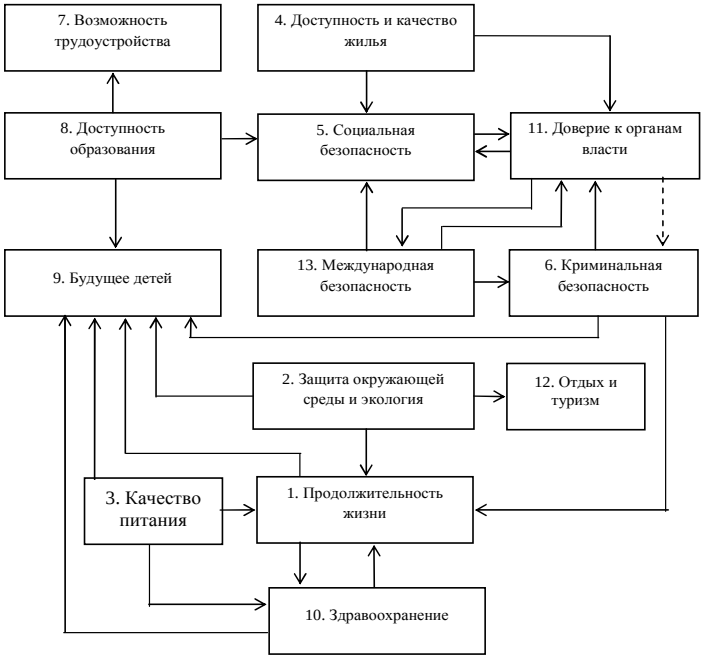


Рис. 2.47. Когнитивная модель взаимосвязей проблем, влияющих на качество жизни

Теперь большинство влияний на схеме рис. 2.47 выглядят вполне логичными. Однако связи, включенные в модель, неравнозначные. Для оценки связей был проведен второй этап экспертной оценки связей. На втором этапе респондентам предлагалось дать оценки связям по когнитивной модели рис. 2.47. Для оценки значимости связей респонденты должны были указать свои оценки в таблице, описывающей когнитивную модель. Связи оценивались по 6-балльной системе. Незначимой связи соответствует оценка «0». В опросе участвовала группа студентов из 50 человек. Усредненные оценки связей представлены в табл. 2.9.

Таблица 2.9

Средние оценки связей

№	Связь		Средняя оценка связи	№	Связь		Средняя оценка связи
1	1	9	3	14	6	11	2
2	1	10	0,7	15	8	5	3
3	2	1	4	16	8	7	2,5
4	2	9	4,5	17	8	9	4
5	2	12	4	18	10	1	5
6	3	1	5	19	10	9	4,5
7	3	9	5	20	11	5	1,5
8	3	10	2,5	21	11	6	0,2
9	4	5	4	22	11	13	0,9
10	4	11	4,5	23	12	1	5
11	5	11	4,5	24	13	5	2
12	6	1	2,5	25	13	6	4,5
13	6	9	3	26	13	11	4,5

В пространстве меньшей размерности экспертам гораздо легче ориентироваться. При повторном опросе некоторые связи могут быть исключены из модели. Например, связь «доверие к органам власти» – «криминальная безопасность» можно признать незначимой (получила оценку 0,2), и она должна быть исключена из модели. Из модели

можно исключить еще две связи с низкими оценками: «продолжительность жизни» – «здравоохранение» (0,7); «доверие к органам власти» – «международная безопасность» (0,9). Все остальные связи имеют средние оценки выше единицы.

Оценки связей могут уточняться с привлечением небольшой группы компетентных экспертов, профессионально занимающихся исследованием социально-экономических процессов. Модель существенно облегчает работу таких экспертов.

После построения когнитивной модели она может быть использована для разработки приоритетов принятия решений. Для анализа когнитивной модели можно использовать количественные методы анализа графов.

Кризисные явления в экономике страны, усугубленные санкциями, предпринятыми против России США и поддержанными странами Европы, не могли не отразиться на социальном самочувствии населения страны. Санкции по-разному отразились на регионах страны [95]. Приморский край относится к одному из регионов, наиболее тяжело воспринимающих экономический кризис. По оценкам экспертов во Владивостоке самая высокая стоимость потребительской корзины в стране. В крае нарастает неуверенность населения в возможностях правительства решить проблемы региона.

Недооценка социальных факторов является причиной возникновения социального стресса. По мнению Б.Т. Величковского, основная причина возникновения социального стресса видится в утрате населением эффективной трудовой мотивации [96]. Эффективная трудовая мотивация позволяет своим трудом обеспечить достойное существование себе и своей семье.

Ряд отечественных авторов придерживаются мнения, что российская государственная, а за ней и региональная, социальная политика базируется на низкой оценке человеческой капитала [97]. Частично это объясняется воздействием распространившейся в России теории «экономически эффективного населения». Главным становится не человек, а экономический рост, который определяет все аспекты социальной политики, включая интенсивность, масштабы и направления модернизации.

Принятие конкретных решений по развитию Приморского края происходит крайне медленно. В период кризиса, наоборот, необходимо быстро принимать решения, чтобы население видело, что оно не брошено на произвол судьбы и каждый должен выживать, как может.

Академик П.А. Минакир в своих последних работах подтверждает необходимость принятия оперативных решений, особенно в кадровой политике. Относительно политики правительства в Дальневосточном регионе он делает заключение, что «следует избавиться от страха, что от перемен может стать еще хуже. Не станет. Хуже некуда» [77, с. 11].

В новых условиях как никогда требуется налаживать контакт властных структур и населения региона.

ЗАКЛЮЧЕНИЕ

Расширение возможностей по сбору данных с развитием Интернета привело к росту потребности в разработке инструментальных средств, позволяющих автоматизировать процесс обработки данных. При больших объемах информации даже тривиальные задачи требуют значительных затрат времени. Большие массивы информации существенно расширили спектр задач, которые можно решать на основании этих данных. Решение множества задач невозможно без использования системного подхода. Системный подход требует совместимости программных средств, используемых для анализа данных. При этом возрастает роль моделей данных, как логических, так и физических. Специфика современных информационных потоков определила основные приоритеты выбора методов обработки данных, представленных в данной работе.

С возрастанием количества источников информации и способов ее сбора особое значение приобретает проблема исследования качества информации. Современные технологии сбора информации привнесли новые источники ошибок, для выявления которых потребовались специализированные средства. В этой связи в работе рассмотрены методические подходы к анализу качества информации, в том числе выявлению недостоверных данных.

Отличительной особенностью современных информационных потоков является возрастание объемов качественной информации, в связи с чем в монографии предложены методические подходы и инструменты ее анализа. Представленные алгоритмы и программные средства позволяют систематизировать качественную информацию, что обеспечивает возможность получения количественных оценок. Особое внимание в работе уделено использованию качественной информации для разработки когнитивных моделей, которые являются эффективным инструментом принятия управленческих решений.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Мартышенко, С.Н. Моделирование многомерных данных и компьютерный эксперимент / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Техника и технология. 2007. №2. С. 47–52.

2. Беликова, Ю.В. Сравнительный анализ сервисов для проведения онлайн-опросов / Ю.В. Беликова // Актуальные научные исследования в современном мире. 2016. № 5-4 (13). С. 36–41.

3. Бондаренко, В.А. Современные тенденции в опросах потребителей с использованием компьютеро-ориентированной коммуникации / В.А. Бондаренко, О.В. Иванченко // Экономика и предпринимательство. 2016. № 1-1 (66-1). С. 605–608.

4. Козлова, О.А. Характеристика качественных и количественных методов онлайн-исследований / О.А. Козлова, Е.Г. Климанова // Вестник Омского университета. Сер.: Мировая экономика и международный бизнес. 2011. № 2. С. 50–54.

5. Иванова, В.А. Особенности проведения дистанционных опросов онлайн / В.А. Иванова // Маркетинг в России и за рубежом. 2015. № 1. С. 11–16.

6. Цалко, Т.В. Цифровизация в маркетинговых исследованиях (на примере онлайн-опросов) / Т.В. Цалко // ЦИТИСЭ. 2019. № 1 (18). С. 42.

7. Беляева, О.В. Использование онлайн-опросов в социологических исследованиях профессионализации молодежи / О.В. Беляева // Известия Юго-Западного государственного университета. Сер.: Экономика. Социология. Менеджмент. 2013. № 4. С. 176–179.

8. Терентьев, Е.А. Влияние визуализации опросного инструментария в онлайн-исследованиях на качество данных / Е.А. Терентьев, А.И. Нефедова, И.А. Груздев // Мониторинг общественного мнения: Экономические и социальные перемены. 2016. № 5. С. 1–15.

9. Галицкий, Е.Б. Потенциальные источники ошибок в данных онлайн-опросов / Е.Б. Галицкий, П.В. Мальцева // Практический маркетинг. 2013. № 10 (200). С. 2–8.

10. Мартышенко, Н.С. Вопросы анализа и прогнозирования пространственного развития рекреации и туризма / Н.С. Мартышенко // Регион: экономика, социология. 2010. №3. С. 167–175.

11. Зангиева, И.К. Проблема пропусков в социологических данных: смысл и подходы к решению / И.К. Зангиева // Социология: методология, методы, математическое моделирование. 2011. № 33. С. 28–56.

12. Мартышенко, С.Н. Многомерные статистические методы повышения достоверности маркетинговых данных / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Практический маркетинг. 2007. №1. С. 20–30.

13. Майков, К.А. Обзор методов предобработки, используемых для решения задач классификации в условиях неполноты данных / К.А. Майков, П.А. Гаврилов // Вестник Рязанского государственного радиотехнического университета. 2016. № 55. С. 140–145.

14. Юдин, Г.Б. Территориальная локализация и уровень неответов в массовом опросе / Г.Б. Юдин // Социологический журнал. 2008. № 1. С. 49–72.

15. Абраменкова, И.В. Методы восстановления пропусков в массивах данных / И.В. Абраменкова, В.В. Круглов // Программные продукты и системы. 2005. № 2. С. 4.

16. Злоба, Е. Статистические методы восстановления пропущенных данных / Е. Злоба, И. Яцкив // Computer Modelling & New Technologies. 2002. Vol. 6, № 1. P. 51–61.

17. Рыженкова, К.В. Методы восстановления пропуска данных при проведении статистических исследований / К.В. Рыженкова // Интеллект. Инновации. Инвестиции. 2012. № 3. С. 127–133.

18. Kalton, G. The Treatment of Missing Survey Data / G. Kalton, D. Kasprzyk // Survey Methodology. 1986. № 12. P. 1–16.

19. Мартышенко, Н.С. Исследование локальных рынков Приморского края в период кризиса / Н.С. Мартышенко. – М.: Профессор, 2015. – 94 с.

20. Мартышенко, Н.С. Алгоритмизация процесса анализа достоверности данных анкетных онлайн-опросов / Н.С. Мартышенко,

С.Н. Мартышенко // Программные системы и вычислительные методы. 2018. № 4. С. 76–85.

21. Мартышенко, Н.С. Исследование структуры потребления туристских продуктов в Приморском крае / Н.С. Мартышенко // Региональная экономика: теория и практика. 2010. № 26. С. 60–68.

22. Насретдинова, М.М. Актуальность онлайн-исследований в России / М.М. Насретдинова // Психология, социология и педагогика. 2014. № 6 (33). С. 24.

23. Перцовский, Н.И. Использование онлайн-исследований для обоснования управленческих (маркетинговых) решений / Н.И. Перцовский, Р.Р. Галимов // Управленческие науки. 2012. № 2. С. 75–80.

24. Ефимова, Д.М. Сравнительный анализ сервисов для продвижения опроса в сети Интернет / Д.М. Ефимова, С.В. Ермолаев // Вестник Российского экономического университета им. Г.В. Плеханова. Вступление. Путь в науку. 2014. № 1-2 (9). С. 88–95.

25. Некрасов, С.И. Сравнение результатов онлайн- и оффлайн-опросов (на примере анкет разной сложности) / С.И. Некрасов // Социология: методология, методы, математическое моделирование. 2011. № 32. С. 53–74.

26. Шкурин, Д.В. Сравнительная оценка качества данных офлайн- и онлайн-опросов / Д.В. Шкурин // Дискуссия. 2015. № 8. С. 101–105.

27. Hohwü, L. et al. Web-based versus traditional paper questionnaires: a mixed-mode survey with a Nordic perspective / L. Hohwü et al. //Journal of medical Internet research. 2013. Vol. 15, № 8.

28. Alessi, E.J. Conducting an internet-based survey: Benefits, pitfalls, and lessons learned / E.J. Alessi, J.I. Martin // Social Work Research. 2010. Vol. 34, № 2. P. 122–128.

29. Hunter, L. Challenging the reported disadvantages of e-questionnaires and addressing methodological issues of online data collection / L. Hunter // Nurse researcher. 2012. Vol. 20, № 1. P. 11–20.

30. McPeake, J. Electronic surveys: how to maximize success / J. McPeake, M. Bateson, A. O'Neill // Nurse researcher. 2014. Vol. 21, № 3. P. 24–26.

31. Phillips, K. Data Use: An evaluation of quality-control questions / K. Phillips // Quirk's Marketing Research Review. December.

2013. URL: http://www.quirks.com/articles/2013/20131205.aspx (дата обращения 23 сентября 2019).

32. Puleston, J. Dimensions of online survey data quality. What really matters? / J. Puleston, M. Eggers // Congress 2012 – Accelerating Excellence – Celebrating 65 Years And Beyond. Atlanta, 2012.

33. Корытникова, Н.В. Параметры проверки и контроля качества онлайн-опроса с использованием параданных / Н.В. Корытникова // Мониторинг общественного мнения: экономические и социальные перемены. 2018. № 3 (145). С. 65–77.

34. Мавлетова, А.М. Влияние элементов приглашения на увеличение доли откликов в онлайн-опросах / А.М. Мавлетова, Н.Г. Малошонок, Е.А. Терентьев // Социология: методология, методы, математическое моделирование. 2014. № 38. С. 72–95.

35. Малошонок, Н.Г. Влияние дизайна на качество данных в онлайн-опросах студентов / Н.Г. Малашонок, Е.А. Терентьев // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 6 (124). С. 15–27.

36. Моисеев, С.П. Выборка, направляемая респондентом в онлайн-опросе: к вопросу о динамике и качестве / С.П. Моисеев, Ю.К. Савинкова // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 6 (124). С. 43–50.

37. Насонова, Н.А. Пример создания репрезентативного интернет-опроса при социологических исследованиях / Н.А. Насонова, Т.В. Кожевникова // Научно-техническое и экономическое сотрудничество стран АТР в XXI веке. 2013. Т. 2. С. 311–319.

38. Федоровский, А.М. Качество онлайн-опросов. Методы проверок / А.М. Федоровский // Мониторинг общественного мнения: экономические и социальные перемены. 2015. № 3 (127). С. 28–35.

39. Фарахутдинов, Ш.Ф. Профессиональные респонденты – камень преткновения онлайн-опросов в современной России / Ш.Ф. Фарахутдинов // Телескоп: журнал социологических и маркетинговых исследований. 2011. № 2. С. 45–47.

40. Неделько, В.М. Исследование погрешности оценок скользящего экзамена / В.М. Неделько // Машинное обучение и анализ данных. 2013. Т. 1, № 5. С. 526–533.

41. Саганенко, Г.И. Системы, форматы и познавательный потенциал открытых вопросов / Г.И. Саганенко // Журнал социологии и социальной антропологии. 2001. Т. 4. С. 171–194.

42. Татарова, Г.Г. Качественные методы в структуре методологии анализа данных / Г.Г. Татарова // Социология: методология, методы, математические модели. 2001. № 14. С. 33–52.

43. Толстова, Ю.Н. Качественная и количественная стратегии. Эмпирическое исследование как измерение в широком смысле / Ю.Н. Толстова, Е.В. Масленников // Социологические исследования. 2000. № 10. С. 101–109.

44. Залесский, П.К. Мировой и российский опыт типологии потребителей по стилю жизни / П.К. Залесский // Маркетинговые исследования. 2005. № 1 (2). С. 3–11.

45. Татарова, Г.Г. Основы типологического анализа в социологических исследованиях: учеб. пособие / Г.Г. Татарова. – М.: Новый учебник, 2004. – 206 с.

46. Ядов, В.А Стратегии и методы качественного анализа данных / В.А. Ядов // Социология: методология, методы, математическое моделирование. 1991. № 1. С. 14–31.

47. Готлиб, А.С. Введение в социологическое исследование. Качественный и количественный подходы. Методология. Исследовательские практики: учеб. пособие / А.С. Готлиб. – М.: Флинта: МПСИ, 2005. – 384 с.

48. Каныгин, Г.В. Инструментальные средства и методологические принципы анализа качественных данных / Г.В. Каныгин // Социология: методология, методы, математические модели. 2007. № 25. С. 70–98.

49. Мартышенко, С.Н. Средства разработки типологий по данным анкетных опросов в среде EXCEL / С.Н. Мартышенко, Н.С. Мартышенко, Д.А. Кустов // Академический журнал Западной Сибири. 2007. №1. С. 75–77.

50. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: Изд-во Института математики, 1999. – 270 с.

51. Малышев, М.Л. Мониторинг социально-трудовой сферы: учеб. пособие / М.Л. Малышев. – М.: Союз; Перспектива, 2007. – 276 с.

52. Мартышенко, С.Н. Современные методы обработки маркетинговой информации / С.Н. Мартышенко, Н.С. Мартышенко. – Владивосток: Изд-во ВГУЭС, 2014. – 148 с.

53. Мартышенко, С.Н. Методические подходы к организации мониторинга социально-экономических процессов на территории муниципальных образований / С.Н. Мартышенко // Современные проблемы науки и образования. 2013. № 5. С. 409.

54. Мартышенко, Н.С. Мониторинг индикаторов социального самочувствия жителей Приморского края в условиях экономического кризиса / Н.С. Мартышенко, Е.Г. Гусев // Политика и общество. 2015. № 11 (131). С. 1517–1529.

55. Гутникова, Е.А. Актуальные проблемы социально-экономического развития муниципалитетов / Е.А. Гутникова // Проблемы развития территории. 2011. Т. 54, № 2. С. 34–45.

56. Куратченко, Е.В. Оценка эффективности управления развитием муниципальных образований (на примере Алтайского края) / Е.В. Куратченко // Регион: Экономика и Социология. 2008. № 3. С. 233–240.

57. Тараскина, А.В. Использование оценки социально-экономической напряженности в управлении сбалансированным развитием муниципальных образований в регионе / А.В. Тараскина, М.А. Гаджиев // Экономика и управление: проблемы, решения. 2018. Т. 6, № 8. С. 91–100.

58. Маркварт, Э. Муниципальное хозяйство как объект управления / Э. Маркварт // Менеджмент и бизнес-администрирование. 2011. № 3. С. 150–152.

59. Мотрич, Е.Л. Миграция в воспроизводстве населения на российском Дальнем Востоке / Е.Л. Мотрич // Уровень жизни населения регионов России. 2013. № 1. С. 25–33.

60. Маршалова, А.С. Проблемы управления социально-экономическим развитием муниципальных образований / А.С. Маршалова, А.С. Новоселов // Регион: Экономика и Социология. 2009. № 1. С. 167–179.

61. Мартышенко, Н.С. Методы обработки нечисловых данных в социально-экономических исследованиях / Н.С. Мартышенко, С.Н. Мартышенко // Вестник Тихоокеанского государственного экономического университета. 2006. №4. С. 48–57.

62. Мартышенко, С.Н. Информационная технология повышения эффективности обработки качественной информации / С.Н. Мартышенко, Е.А. Егоров // Информационные технологии моделирования и управления. 2009. № 6 (58). С. 753–760.

63. Кравченко, Н.А. Перспективы развития малого инновационного бизнеса (на примере Новосибирской области) / Н.А. Кравченко, С.А. Кузнецова, А.Т. Юсупова // Проблемы современной экономики. 2011. № 1. С. 112–115.

64. Мельник, М.В. Комфортная среда для развития малого бизнеса / М.В. Мельник // Инновационное развитие экономики. 2012. № 7. С. 3–11.

65. Сидорова, Н.И. Анализ и оценка развития предпринимательства: региональный аспект / Н.И. Сидорова // Вестник Астраханского государственного технического университета. Сер.: Экономика. 2012. № 1. С. 126–133.

66. Алиев, Б.Х. Государственное регулирование и поддержка малого бизнеса в условиях кризиса / Б.Х. Алиев, Х.М. Мусаева // Финансы и кредит. 2010. № 32. С. 16–23.

67. Буханцева, С.Н. Формирование эффективной предпринимательской среды / С.Н. Буханцева // Фундаментальные исследования. 2013. № 4-2. С. 471–475.

68. Швецов, А.Н. Современные ИКТ в деятельности российских органов власти: преобразят ли они государственное и муниципальное управление? / А.Н. Швецов // Российский экономический журнал. 2011. № 3. С. 21–45.

69. Безрукова, Т.Л. Контроллинг как система информационно-аналитической поддержки стратегического управления / Т.Л. Безрукова, П.А. Петров // Менеджмент и бизнес-администрирование. 2011. № 3. С. 32–39.

70. Бутуханов, И.Н. Управление развитием предприятий малого и среднего бизнеса на муниципальном уровне / И.Н. Бутуханов // Перспективы науки. 2012. № 39. С. 106–108.

71. Блинов, А.О. Формирование организационно-экономического механизма поддержки малого бизнеса как основа стратегического развития территории / А.О. Блинов, С.А. Нефедкина // Известия Сочинского государственного университета. 2011. № 4. С. 23–30.

72. Сибурина, Т.А. Механизмы государственно-частного партнерства: российский и зарубежный опыт, особенности применения в социальной сфере / Т.А. Сибурина // Менеджмент и бизнес-администрирование. 2010. № 1. С. 42–82.

73. Мартышенко, Н.С. Формирование туристского кластера и управление его развитием на территории Приморского края / Н.С. Мартышенко // Регион: системы, экономика, управление. 2008. № 2. С. 122–132.

74. Мартышенко, Н.С. Конкурентное позиционирование предложения территориального туристского продукта Приморского края в Северо-Восточной Азии / Н.С. Мартышенко // Экономика и предпринимательство. 2011. № 5. С. 153–163.

75. Жихаревич, Б.С. Заявленные и реальные приоритеты региональных и местных властей: подход к выявлению и сопоставлению / Б.С. Жихаревич, Н.Б. Жунда, О.В. Русецкая // Регион: Экономика и Социология. 2013. № 2 (78). С. 108–132.

76. Коломак, Е.А. Пространственная концентрация экономической активности в России / Е.А. Коломак // Пространственная экономика. 2014. № 4. С. 82–99.

77. Минакир, П.А. Новая восточная политика и экономические реалии / П.А. Минакир // Пространственная экономика. 2015. № 2. С. 7–11.

78. Суслов, В.И. Сценарии экономического развития: инновационные аспекты / В.И. Суслов // ЭКО. 2010. № 2. С. 2–14.

79. Горелова, Г.В. О когнитивном моделировании сложных систем, инструментарий исследования / Г.В. Горелова // Известия ЮФУ. Технические науки. 2012. № 6 (131). С. 236–240.

80. Корнилова, М.В. Социальное самочувствие: понятие и основные показатели / М.В. Корнилова // Евразийское Научное Объединение. 2015. Т. 2, № 3 (3). С. 135–137.

81. Римашевская, Н.М. Социальная политика сбережения народа / Н.М. Римашевская // Ученые записки Петрозаводского государственного университета. Сер.: Общественные и гуманитарные науки. 2010. № 5. С. 75–82.

82. Айвазян, С.А. Сравнительный анализ интегральных характеристик качества жизни населения субъектов Российской Федерации / С.А. Айвазян. – М.: ЦЭМИ РАН, 2001. – 64 с.

83. Гришина, И.В. Качество жизни населения регионов России: методология исследования и результаты комплексной оценки / И.В. Гришина, А.О. Полынев, С.А. Тимонин // Современные производительные силы. 2012. № 1. С. 70–83.

84. Martyshenko, S.N. Information technology for increasing qualitative information processing efficiency / S.N. Martyshenko, E.A. Egorov // Journal of Modern Applied Statistical Methods. 2011. Vol. 10, № 1. P. 207–213.

85. Авдеева, З.К. Когнитивное моделирование для решения задач управления слабоструктурированными системами (ситуациями) / З.К. Авдеева, С.В. Коврига, Д.И. Макаренко // Управление большими системами: сб. тр. 2007. № 16. С. 26–39.

86. Глушакова, О.В. Парадигма публичного управления: обеспечение качества жизни / О.В. Глушакова // Проблемы теории и практики управления. 2013. № 11. С. 58–65.

87. Сухарев, М.В. Распределенные когнитивные модели и социальное партнерство / М.В. Сухарев // Петрозаводск – 300: Карелия в процессе перемен. – Петрозаводск: КарНЦ РАН, 2004. С. 341–347.

88. Бялецкая, Е.М. О принципах когнитивного моделирования сложных систем / Е.М. Бялецкая, И.Ю. Квятковская // Вестник Астраханского государственного технического университета. 2006. № 1. С. 116–119.

89. Гарькина, И.А. Когнитивное моделирование сложных слабоструктурированных систем: пример реализации / И.А. Гарькина, А.М. Данилов, Е.В. Королев // Региональная архитектура и строительство. 2008. № 2. С. 16–21.

90. Максимов, В.И. Когнитивные технологии для поддержки принятия управленческих решений / В.И. Максимов, Е.К. Корноушенко, С.В. Качаев // Информационное общество. 1999. № 2. С. 50–54.

91. Гузаиров, М.Б. Когнитивная модель формирования показателя качества жизни / М.Б. Гузаиров, Б.Г. Ильясов, Е.Ш. Закиева, И.Б. Герасимова // Вестник Уфимского государственного авиационного технического университета. 2013. Т. 17, № 2 (55). С. 215–220.

92. Ильясов, Б.Г. Системный подход к построению когнитивной модели качества жизни / Б.Г. Ильясов, Е.Ш. Закиева, И.Б. Герасимова // Вопросы современной науки и практики. Университет им. В.И. Вернадского. 2013. № 3 (47). С. 214–221.

93. Свечкарев, В.П. Интегрированные когнитивные архитектуры моделей социальных систем / В.П. Свечкарев, К.С. Радько // Инженерный вестник Дона. 2012. Т. 23, № 4-2 (23). С. 109.

94. Орлова, Т.М. Управление знаниями в контексте когнитивного подхода / Т.М. Орлова // Проблемы теории и практики управления. 2015. № 4. С. 113–122.

95. Казанцев, С.В. Антироссийские санкции для субъектов Российской Федерации / С.В. Казанцев // Регион: Экономика и Социология. 2015. № 1 (85). С. 20–38.

96. Величковский, Б.Т. Социальный стресс, трудовая мотивация и здоровье / Б.Т. Величковский // Бюллетень сибирской медицины. 2005. № 3. С. 5–19.

97. Стукаленко, Е.А. Дифференциация доходов населения: причины и последствия / Е.А. Стукаленко // Вестник Омского университета. Сер.: Экономика. 2014. № 1. С. 183–187.

Научное издание

Мартышенко Сергей Николаевич
Мазелис Лев Соломонович
Солодухин Константин Сергеевич

АВТОМАТИЗАЦИЯ АНАЛИЗА ДАННЫХ В ИССЛЕДОВАНИИ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

Монография

Компьютерная верстка М.А. Портновой
Подготовлено к изданию М.А. Шкарубо

Подписано в печать 30.12.2019. Формат 60×84/16.
Бумага писчая. Печать офсетная. Усл. печ. л. 10,0.
Уч.-изд. л. 9,5. Тираж 600 экз. [1–100]. Заказ 88

Издательство Владивостокского государственного университета экономики и сервиса
690014, Владивосток, ул. Гоголя, 41
Отпечатано во множительном участке ВГУЭС
690014, Владивосток, ул. Гоголя, 41