КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

М.Г. Шендерюк

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИСТОЧНИКОВЕДЕНИИ

Калининград
1997

КАЛИНИНГРАДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

М.Г. Шендерюк

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИСТОЧНИКОВЕДЕНИИ

Учебное пособие

Калининград
1997

Шендерюк М.Г. Количественные методы в источниковедении: Учеб. пособие / Калинингр. ун-т. - Калининград, 1996. - 75 с. - ISBN 5-88874-043-8.

Посвящено вопросам применения количественных методов в источниковедении массовых и нарративных источников. Рассматриваются методологические проблемы применения количественных методов в исторических исследованиях; описываются основные методы математической статистики, используемые историками; анализируются источниковедческие задачи, решаемые с помощью современных методов.

Адресовано в первую очередь студентам-историкам. Может быть полезно и более широкой аудитории историков-профессионалов, занимающихся клиометрическими исследованиями.

Печатается по решению редакционно-издательского Совета Калининградского государственного университета.

Рецензенты: лаборатория исторической информатики им. акад. И.Д. Ковальченко кафедры источниковедения МГУ им. М.В. Ломоносова; чл.-кор. РАН Л.В. Милов.

ISBN 5-88874-043-8

Марина Геннадьевна Шендерюк

КОЛИЧЕСТВЕННЫЕ МЕТОДЫ В ИСТОЧНИКОВЕДЕНИИ

Учебное пособие

Лицензия №020345 от 14.01.1997 г.
Редактор Н.Н. Мартынюк.
Оригинал-макет подготовлен Д.В. Голубиным.
Подписано в печать 11.12.1997 г. Формат 60×90 $^{1}/_{16}$.
Бум. для множит. аппаратов. Ризограф. Усл. печ. л. 4,7.
Уч.-изд. л. 5,0. Тираж 150 экз. Заказ .

Калининградский государственный университет,
236041, Калининград обл., ул. А.Невского, 14.

Памяти моего учителя
Ивана Дмитриевича Ковальченко

ВВЕДЕНИЕ

Настоящее учебное пособие посвящено вопросам применения количественных методов в источниковедении массовых и нарративных источников. Сегодня, в конце 90-х гг. XX в., когда практика клиометрических исследований насчитывает более трех десятилетий и даже самые большие скептики-традиционалисты признали за количественными методами право на существование, методы эти остаются инструментарием лишь довольно узкого круга специалистов. Овладение математическими методами представляет по-прежнему значительные трудности. Российские историки не имеют специальных исторических журналов, на страницах которых рассматривались бы вопросы методики и техники исторических исследований, нет широкого доступа к современным компьютерам, практически отсутствуют пакеты специальных программ, обеспечивающих нужды историка. Не спасает положение и введение в университетский учебный план обязательного, но небольшого по объему, курса «Количественные методы в исторических исследованиях». Связано это как со слабым распространением в учебной практике гуманитариев компьютерной техники, так и с недостатком соответствующих учебных пособий. Вышедшее в 1984 г. учебное пособие «Количественные методы в исторических исследованиях», отражающее практику преподавания дисциплины на историческом факультете МГУ, где обучение студентов квантитативной истории ведется с середины 70-х гг. по двум самостоятельным курсам и поставлено на иную техническую базу, сразу стало библиографической редкостью. Кроме того, обширный раздел этого пособия, посвященный описанию основных методов математической статистики, довольно труден для неподготовленного студента-историка.

Предлагаемое учебное пособие имеет тройное предназначение, оно содержит материал по трем читаемым на историческом факультете Калининградского университета курсам («Источниковедение», «Количественные методы в исторических исследованиях» и вводимому курсу «Методы математической статистики»). Итогом изучения названных курсов в рамках представленных в учебном пособии проблем должно стать овладение сту-

дентами минимумом знаний, необходимым для квалифицированного изучения трудов отечественных клиометристов и проведения собственных исследований на основе количественных методов.

Структура учебного пособия диктовалась практикой чтения автором курса «Количественные методы в исторических исследованиях», традиционно делящегося на три части, в которых последовательно рассматриваются методологические проблемы применения количественных методов в исторических исследованиях; описываются основные методы математической статистики, используемые историками; характеризуются основные направления отечественных клиометрических исследований. Такая структура изложения удовлетворяет и задачам изучения проблем применения количественных методов в исторических исследованиях в двух других названных курсах.

Однако, поскольку целью настоящего пособия является изучение применения количественных методов в источниковедении, первый и третий разделы пособия имеют более узкую проблематику, чем курс «Количественные методы». Методологические проблемы, освещаемые в первом разделе, не включают вопросы моделирования исторических явлений и процессов. В третьем разделе анализируются лишь работы отечественных клиометристов, посвященные решению собственно источниковедческих задач.

Учебное пособие адресовано в первую очередь, студентам-историкам. Однако оно может быть полезно и более широкой аудитории историков-профессионалов, занимающихся клиометрическими исследованиями.

Раздел 1. МЕТОДОЛОГИЧЕСКИЕ ПРОБЛЕМЫ

1.1. Математизация и компьютеризация исторического знания

Современная наука характеризуется интенсивным процессом *математизации* и *компьютеризации* знания. Это обусловлено как успехами в развитии прикладной математики и вычислительной техники – «компьютерной революцией», приведшей к радикальному расширению возможностей интеллектуальной деятельности человека, так и состоянием самой науки, перед которой в последней четверти XX столетия остро встали проблемы систематизации, хранения и использования все увеличивающегося объема накопленной информации и совершенствования методов ее выявления, обработки и анализа. Информационный взрыв в науке усилил потребности обобщенного подхода в познании явлений объективного мира, что проявляется в тенденции к интеграции научного познания.

На всех этапах развития науки существует две тенденции: с одной стороны, к детальному, углубленному, т.е. дифференцированному изучению явлений и процессов действительности, с другой стороны, к их целостному, системному, т.е. интегральному изучению. В определенные периоды та или другая тенденция выступает на первый план. Говоря о превалировании в современную эпоху *тенденции к интеграции*, надо иметь в виду, что эта тенденция осуществляется как синтез обобщенного и специализированного подходов к изучению реальности и сам процесс дифференциации исследований, продолжающий активно развиваться, идет все больше интегральным путем.

Интегральные исследования имеют ряд уровней: 1) общенаучный, 2) междисциплинарный, 3) внутридисциплинарный. К числу проблем, решаемых на *общенаучном уровне*, относятся так называемые глобальные проблемы, требующие интеграции многих наук разного профиля: естественных, технических и общественно-гуманитарных (например, экологическая). *Междисциплинарные* исследования основываются на синтезе методов ряда наук, причем как наук одного типа (например, математическая физика, биохимия), так и разных типов (техническая эстетика, математическая лингвистика). *Внутридисциплинарные* исследования осуществляют межпроблемную интеграцию в рамках одной науки (в исторической науке к таким исследованиям относятся, например, этноисторические, социально-экономические, социально-политические).

Интеграция научного познания основывается на синтезе теорий, понятий и методов различных областей науки, но может приводить и к необходимости создания новых теорий и общенаучных и межпроблемных методов исследования, универсализации языка науки. Это становится возмож-

ным лишь путем абстрагирования и формализации, высшим выражением которых является математизация научного знания.

Математизация знания способствует осуществлению интегрального, целостного, *системного подхода* к изучению явлений действительности. Суть такого подхода в том, что исследуемый объект, явление или процесс рассматривается как некая целостная система с присущей ей структурой. Задача исследования при этом состоит в раскрытии строения и взаимосвязей, характеризующих эту структуру, и в выявлении их качественного своеобразия. Поскольку математика является наукой о структурах в их абстрактной форме, то решение подобных задач оказывается наиболее эффективным при использовании различных количественных и математических методов, составляющих структурный и функциональный анализы, в которых реализуются принципы системного подхода.

Таким образом, потребность в математизации и компьютеризации научного знания диктуется прежде всего внутренними тенденциями развития современной науки, в том числе и науки исторической.

Информационный взрыв, произошедший в науке в последние десятилетия, охватил и область исторического знания. Объем накопленных фактических данных неуклонно возрастает, ежегодно появляются тысячи трудов по различным вопросам отечественной и зарубежной истории. Расширение круга исследовательских задач ведет, в свою очередь, к необходимости привлечения новых конкретно-исторических данных и более эффективного использования накопленных знаний. Это ставит перед историками две задачи: 1) расширения источниковой базы исследований, 2) повышения информативной отдачи источников.

Расширение источниковой базы осуществляется на основе вовлечения в научный оборот обширных комплексов массовых источников, характеризующих массовые явления и процессы исторической действительности.

Массовые явления представляют собой совокупность исторических феноменов (объектов), с одной стороны, обладающих одинаковыми свойствами, а с другой - характеризующихся различной мерой этих свойств. Совокупности таких объектов составляют в большей или меньшей степени сложные системы с присущими им структурами, подверженными непрерывным колебаниям и изменениям. *Массовые источники* - источники, характеризующие такие объекты действительности, которые образуют определенные общественные системы с соответствующими структурами. Массовые источники отражают сущность и взаимодействие массовых объектов, составляющих эти системы, а следовательно, строение, свойства и состояние самих систем[1].

[1] См.: Массовые источники по социально-экономической истории России периода капитализма. М., 1979. С.6.

К массовым источникам относятся различные статистические материалы и обследования, делопроизводственная документация, систематизированные справочные материалы и др.

Сама природа массовых источников предопределяет возможность и необходимость системного подхода и структурного и функционального анализа при их изучении - математизации исторических исследований. Кроме того, опыт использования ЭВМ при работе историков с массовыми источниками вызывает потребность в создании баз массовых исторических данных, банков машиночитаемой информации - компьютеризации исторического знания.

С математизацией и компьютеризацией тесно связано и решение задачи повышения информативной отдачи источников, извлечения из них новой, скрытой информации, непосредственно не выраженной в конкретно-исторических данных. Анализ многообразных взаимосвязей, присущих явлениям исторической действительности, с помощью количественных и компьютерных методов обработки исходных данных позволяет перейти к более глубокому и точному раскрытию сущности исторических явлений и процессов, поднять уровень исторических исследований.

Таким образом, применение количественных методов в исторических исследованиях отражает современный этап в развитии исторической науки, характеризующийся углублением процесса математизации и компьютеризации научного знания.

1.2. Сфера применения количественных методов

Краеугольным камнем в вопросе применения количественных методов в исторических исследованиях является соотношение количественного и качественного анализа. Бытует ошибочное представление о том, что количественные методы противостоят качественным, при этом с последними нередко отождествляются традиционные, описательные методы исторического анализа. Это приводит либо к выводу о несостоятельности количественных методов как формальных в сравнении с качественными, либо к абсолютизации количественных методов, поскольку описательные несовершенны. В действительности же, чтобы раскрыть внутреннюю суть изучаемых явлений и процессов, необходим глубокий качественный, *сущностно-содержательный анализ*, он является ключевым в любом научном исследовании, какими бы методами оно не оперировало. Но раскрыть сущность исследуемых явлений и процессов можно лишь на основе имеющейся о них информации. Эта информация может выражаться и подвергаться обработке двумя способами: описательным и количественным.

Таким образом, описательный и количественный анализ характеризуют явление, а качественный - сущность. Следовательно, речь может идти либо о *сущностно-описательном*, либо о *сущностно-количественнном анализе*, и альтернативой количественному анализу является не качественный, а описательный.

Качественный анализ - это совокупность аналитических и синтетических процедур, имеющих целью выявление коренных свойств, основных закономерностей и особенностей возникновения и функционирования исследуемых объектов и явлений.

Количественный анализ - это выявление и формирование системы численных характеристик изучаемых объектов, явлений и процессов действительности, которые, будучи подвергнуты определенной математической обработке, создают основу для раскрытия количественной меры соответствующего качества.

Совершенно очевидно, что сущностно-количественный анализ обладает рядом преимуществ перед сущностно-описательным, поскольку он позволяет выявить количественную меру тех качеств, которые присущи изучаемым явлениям или процессам. Но в то же время не следует отрицать или недооценивать традиционные методы исследования. В зависимости от исследовательской задачи и характера конкретно-исторических данных те или иные из них могут быть определяющими, но чаще целесообразно их умелое сочетание.

Теоретически количественные методы могут применяться при изучении любых явлений и процессов исторической действительности, поскольку всякому качеству присуще определенное количество. Однако переход к количественным методам осуществляется лишь тогда, когда наука «готова» к математизации - когда становится возможным измерение исторических явлений, что требует определенного теоретического уровня познания исследуемых явлений и процессов. «Математизация, таким образом, будет эффективной только тогда, когда математизируемая наука будет достаточно зрелой, обладающей сложившимся концептуальным аппаратом, т.е. в ней должны быть установлены на качественном уровне наиболее важные понятия, гипотезы, обобщения и законы»[2].

Существует три формы математизации научного познания:

1) численное выражение изучаемой реальности для выявления количественной меры соответствующих качеств;

2) построение математических моделей исследуемых явлений и процессов;

3) построение новых и выражение и анализ существующих научных теорий, т.е. формализация основных итогов самого научного знания.

[2] Рузавин Г.И. Математизация научного знания. М., 1984. С.191.

Основной формой математизации научного познания на современном этапе развития науки является моделирование изучаемых явлений и процессов. Моделирование основано на системном подходе и структурном и функциональном анализе. При этом строятся модели различной сложности – от моделей отдельных явлений до моделей обширных процессов объективной действительности.

Применение количественных методов в отечественных исторических исследованиях связано с именем блестящего исследователя, крупного организатора науки академика Ивана Дмитриевича Ковальченко, основавшего отечественную школу *квантитативной истории*. Перу И.Д. Ковальченко принадлежит около 200 работ, включая 6 монографий, в последней из которых[3] (удостоенной Государственной премии) он изложил целостную систему теоретико-методологических основ применения количественных методов исследования. Преждевременная смерть прервала его работу над очередной книгой, подводившей итог четырех десятилетий его размышлений над ключевыми вопросами аграрной истории России.

Ведущим направлением квантитативной истории с начала 60-х гг. стала область социально-экономических (прежде всего – аграрно-исторических) исследований. Однако сегодня количественные методы применяются практически во всех сферах исторических исследований: социально-политических, историко-культурных, демографических, археологических; при этом наиболее впечатляющие успехи достигнуты в источниковедении (создание баз и банков машиночитаемых данных, изучение происхождения и авторства источников, проверка достоверности, репрезентативности и сопоставимости источников). Проблема заключается не о том, где можно применить количественные методы, а в том, как применить, чтобы получить действительное приращение знания.

Основная цель исторического исследования, использующего количественные методы анализа, - получить новую, непосредственно не выраженную в исходных данных информацию. Историко-содержательный анализ этой информации должен дать новые знания об изучаемых явлениях и процессах. Успех здесь определяется тем, насколько логическая суть количественных методов соответствует внутренней логике самого явления или процесса. «Взаимопроникновение, синтез конкретно-содержательного, гуманитарного и формально-логического, математического подходов - вот тот узел, искусство завязывания которого при прочих равных условиях обеспечивает успех в применении математических методов в исторических исследованиях»[4].

---

[3] См.: Ковальченко И.Д. Методы исторического исследования. М., 1987.

[4] Ковальченко И.Д. Указ. соч. С.324.

1.3. Основные этапы клиометрического исследования

Историческое исследование, основанное на применении количественных методов, делится на несколько этапов:

1) постановка исследовательской задачи и формулировка содержательной гипотезы относительно ее разрешения;

2) отбор источников и формирование системы достоверных и репрезентативных конкретно-исторических данных;

3) выбор количественного метода, позволяющего формализовать содержательную гипотезу и дать четкую математическую постановку задачи;

4) математическая обработка и анализ количественных показателей;

5) интерпретация полученных результатов, подтверждение или опровержение выдвинутой гипотезы.

На каждом из этих этапов необходимо соблюдать определенные конкретно-методологические принципы.

Правильная *постановка исследовательской задачи* требует всестороннего подхода к изучаемым объектам, явлениям и процессам исторической действительности, рассмотрения их во всей сложности, взаимообусловленности и развитии. При этом важное значение имеет объективный анализ историографии, не допускающий проявления нигилизма или консерватизма по отношению к работам предшественников. Кроме того, постановка исследовательской задачи не должна исходить из стремления получить заранее заданный итог.

Последнее особенно важно для клиометрических исследований, поскольку нередко именно с помощью математических методов с их сложным и не всем понятным аппаратом пытаются сделать «сенсационные» открытия, в корне переворачивающие все ранее существовавшие представления об изучаемых явлениях или процессах.

После постановки исследовательской задачи происходит выявление исторических источников, содержащих данные, которые могут быть использованы для ее решения. Но прежде чем переходить к обработке и анализу этих данных, необходимо установить их достоверность и репрезентативность.

Достоверность - точность измерения соответствующих признаков изучаемых явлений и процессов. Эта точность может варьировать от весьма приблизительных количественных оценок до полного соответствия действительным размерам явлений.

Ошибка (погрешность) измерения - разница между величиной, полученной в результате измерения, и истинным значением признака.

Ошибки измерения имеют различную природу, классифицировать их можно следующим образом:

Ошибки измерения могут быть *качественными*, вызванными несостоятельностью или ограниченностью тех теоретико-методологических принципов, исходя из которых проводится измерение, и *количественными*, являющимися результатом неточности самих измерений, - собственно ошибками измерения.

Решая вопрос о достоверности конкретно-исторических данных, исследователь в первую очередь должен выявить те общеисторические и статистико-математические теоретико-методологические принципы, на основе которых проводилось измерение и получены результаты, зафиксированные в исторических источниках. При этом теоретико-методологические посылки определяют не только адекватность самого измерения, но и последующую сводку первичных данных. Так, например, общеизвестно, что ценнейшие первичные материалы земской статистики в России в конце XIX - начале XX века были в значительной степени испорчены и обеднены в процессе пообщинных сводок.

Собственно ошибки измерения делятся на ошибки регистрации количественных значений признаков и ошибки исчисления.

Ошибки регистрации могут быть систематическими и случайными. *Систематические ошибки* являются следствием проявления определенных причин, которые чаще всего могут быть установлены. При этом систематические ошибки измерения бывают *преднамеренными* (например, систематическое занижение прибыли и завышение расходов промышленниками) и *непреднамеренными*, связанными с округлениями и трудностью восстановления по памяти точных данных.

Случайные ошибки регистрации вызываются самыми различными причинами (небрежность и невнимательность регистраторов, неисправность измерительных приборов, несовершенство методов измерения), они имеют разнонаправленный характер (показатели то завышаются, то занижаются) и при большом числе наблюдений взаимопогашаются.

Ошибки исчисления возникают при обработке количественных данных в результате многократных вычислительных операций с неточными ис-

ходными показателями, замены точных расчетов приближенными, многократных округлений и т.д.

Таким образом, для выявления достоверности используемых количественных данных необходима проверка их точности путем определения ошибки измерения. Однако практически установление числового значения ошибки измерения представляется довольно сложным, а порой просто невозможным в силу отсутствия необходимых сведений. Тем не менее исследователь должен стремиться к выявлению всех возможных ошибок и их природы и хотя бы приблизительной оценке их величины.

Степень точности используемых количественных показателей должна давать возможность проведения сопоставлений и выявления различий в количественной мере изучаемых признаков.

Так, например, сравнивая доходы от земледелия двух крестьянских дворов x=100 р. и y=110 р., предположим возможность ошибки в сведениях в пределах 10%. Тогда доходы первого двора будут находиться в пределах $90<x<110$ р., а второго двора: $99<y<121$ р. Очевидно, что для установления различий в доходах этих двух дворов такая ошибка слишком велика. Лишь при ошибке в 4% и менее можно утверждать, что доходы второго двора превышают доходы первого ($96<x<104$ р., $105{,}6<y<114{,}4$ р.).

Для определения достоверности конкретно-исторических данных, зафиксированных в источниках, необходимо установить:

- каковы были представления о сущности изучаемых явлений в период создания исторических источников;
- кто (учреждение или лицо) и с какой целью проводил сбор;
- по какой программе проводился сбор;
- как был организован сбор;
- откуда поступали сведения;
- кто непосредственно собирал их;
- как обрабатывались и обобщались первичные данные;
- какова была система проверки данных и т.д.

Ясно представляя себе недостатки данных, следует искать пути повышения их информативной отдачи. При этом важно иметь в виду, что многие количественные показатели, отличающиеся значительными погрешностями, которые нельзя использовать для характеристики абсолютных значений изучаемых признаков, могут быть основой для получения весьма точных относительных сравнительных данных.

Кроме выявления достоверности конкретно-исторических данных, необходимо решить вопрос об их репрезентативности, представительности.

Качественная репрезентативность определяется тем, в какой мере показатели, на основе которых изучаются соответствующие явления и процессы, отражают именно те черты и свойства, которые характеризуют

внутреннюю суть этих явлений и процессов. Поэтому важно на основе содержательной интерпретации исследовательской задачи отобрать именно такие, представительные, показатели, без которых нельзя правильно раскрыть суть исследуемых явлений и процессов. При этом в исследовании, основанном на привлечении источников, содержащих большое число данных, эффективным является выявление наиболее существенных из них путем предварительной экспериментальной обработки небольшой выборочной совокупности объектов.

Количественная репрезентативность выражается в том, что показателей должно быть достаточно для получения надежных, т.е. имеющих необходимую точность, численных значений признаков, характеризующих изучаемые явления и процессы.

Когда исследователь имеет данные, характеризующие все объекты изучаемой совокупности, и подвергает их сплошной обработке, проблемы количественной репрезентативности показателей не существует. Однако, как правило, историк имеет дело либо со слишком большим объемом данных, которые трудно подвергнуть сплошной обработке, либо с немногими сохранившимися сведениями. В том и другом случае он работает с выборочными данными: либо с собственно выборкой, сформированной самим исследователем, либо с так называемой «*естественной выборкой*».

Вопрос репрезентативности выборочных данных решается с помощью хорошо разработанного в математической статистике выборочного метода. В основе его лежит положение о том, что репрезентативными являются *случайные выборки,* т.е. такие выборки, при формировании которых каждый объект изучаемой совокупности имеет одинаковый шанс попасть в выборку. Выборочный метод представляет различные способы формирования случайных выборок, однако пока не существует достаточно эффективных математических способов проверки случайности «естественных выборок». Историк в этом случае должен, опираясь на традиционные исторические методы анализа, выяснить историю возникновения и судьбу данных естественной выборки, условия их хранения, причины утраты части сведений, равномерность охвата сохранившимися данными исследуемой совокупности объектов в пространстве и во времени и т.д.

Некоторые методы проверки случайности «естественных выборок» можно найти в фундаментальном труде И.Д.Ковальченко «Русское крепостное крестьянство в первой половине XIX века» (М., 1967).

В России, до появления сплошных обследований крестьянских хозяйств в конце XIX в., описания крестьянских хозяйств охватывали лишь отдельные помещичьи имения. Так, по русской крепостной деревне первой половины XIX в. в вотчинных фондах содержатся подворные описи, которые охватывают всего несколько сот помещичьих имений из более чем 50

тыс. имений. В большинстве имений подворные описания не составлялись вообще, а из составленных описей сохранилась лишь небольшая часть. Для решения вопроса о репрезентативности этой «естественной выборки» И.Д.Ковальченко изучает историю происхождения и судьбу описей. Он устанавливает, что составление подворных описей крестьянских хозяйств чаще всего (даже в тех имениях, где они велись более или менее регулярно) было связано с различными случайными обстоятельствами (переход имения к новому владельцу, составление обзоров владений, проверка деятельности вотчинной администрации, изменение форм эксплуатации крестьян и размеров их повинностей и т.п.). Поскольку подворные описи имели практическую ценность в течение всего нескольких лет, т.к. менялись положение крестьян и состояние их хозяйств, то вся последующая их судьба зависела от множества случайных факторов. Кроме того, описи достаточно равномерно охватывают всю совокупность помещичьих имений в пространстве и во времени - на каждую губернию, входившую в зону размещения русского крепостного крестьянства, приходится по несколько описей, и они охватывают начало и середину века (10-20-е и 40-50-е годы XIX в.). Все это позволяет автору на содержательном уровне делать вывод о репрезентативности исследуемой «естественной выборки».

После формирования системы достоверных и репрезентативных конкретно-исторических данных выбирается количественный метод, позволяющий решить поставленную задачу. Главным требованием, предъявляемым к выбранному методу, является его *адекватность* сущности изучаемых явлений и процессов. Для решения вопроса об адекватности необходимо, ясно представляя себе логическую суть количественного метода, соотнести ее с логикой самого явления. Для этого исследователь должен сначала теоретически определить содержательную суть изучаемого явления и пути решения поставленной задачи, а затем выявить те количественные методы, которые наиболее пригодны для их реализации.

Так, например, изучая внутреннюю структуру помещичьего хозяйства Европейской России в пореформенную эпоху[5], исследователи сначала дали четкое теоретическое описание двух крайних вариантов организации помещичьего хозяйства – капиталистического и отработочного. Структура капиталистически организованного помещичьего хозяйства должна была характеризоваться взаимозависимостью и тесной сбалансированностью его компонентов, тогда как структура отработочного хозяйства такой сбалансированности иметь не должна. В соответствии с этой гипотезой выбирался метод исследования – корреляционный анализ, поскольку он позволяет

[5] См.: Ковальченко И.Д., Селунская Н.Б., Литваков Б.М. Социально-экономический строй помещичьего хозяйства Европейской России в эпоху капитализма: Источники и методы изучения. М., 1982.

выявлять тесноту взаимосвязи признаков. Построение корреляционной модели помещичьего хозяйства и ее анализ должны были ответить на вопрос о господстве капиталистической или отработочной системы в помещичьем хозяйстве России. Методика эта оказалась достаточно эффективной для раскрытия сути внутреннего социально-экономического строя помещичьего хозяйства.

Успех математической обработки и анализа количественных показателей на основе выбранного метода определяется *корректностью* применения математического аппарата. Поскольку математические методы имеют свой диапазон применения, необходимо учитывать те условия и ограничения, которые они предполагают. Так, например, многие методы математической статистики требуют проверки нормальности распределения количественных признаков, выявления случайности выборочных данных, определения вида функциональной зависимости между признаками и т.п.

Наконец, заключительной стадией клиометрического, как и любого другого, исследования является интерпретация полученных результатов. Определяющее значение здесь имеет уровень качественного, сущностно-содержательного анализа. От общей исторической эрудиции исследователя зависит корректность и глубина выводов, подтверждающих или опровергающих выдвинутую содержательную гипотезу, и определение дальнейших перспектив исследования.

Раздел 2. МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ

2.1. Первоначальные понятия статистики

Преимущественное положение в системе количественных методов, используемых историками, занимают методы математико-статистического анализа.

Термин «*статистика*» происходит от латинского слова «статус» (status) – положение, состояние явлений. Этот термин неоднозначен. Под статистикой понимают совокупность итоговых показателей, количественно характеризующих различные стороны общественной жизни, – экономику, политику, культуру. Под статистикой понимают также практическую деятельность по сбору и обобщению соответствующих данных. Статистикой называют и особую общественную науку.

Наука статистика, как и всякая иная наука, возникла из практических потребностей людей. Она имеет богатую историю. Примером совершенствования статистических приемов может служить изменение единицы наблюдения, связанной с налогообложением крестьян в России: в XV-XVI вв. – «соха» (крестьянская община, объединявшая до нескольких десятков дворов), в XVII в. – «двор», в XVIII в. – «ревизская душа» (крепостной крестьянин мужского пола).

Предметом статистики выступает количественная сторона массовых общественных явлений, взятая в неразрывной связи с их качественной стороной и отображаемая посредством статистических показателей.

Статистический показатель - это число, характеризующее ту или иную особенность, сторону общественных явлений.

Все общественные науки объектом своего изучения имеют общество. Объект изучения статистики выступает в виде особых множеств массовых общественных явлений – статистических совокупностей.

Статистической совокупностью называется множество объективно существующих во времени и пространстве явлений, однокачественных в определенной связи. Отдельные первичные неделимые элементы, или индивидуальные явления, составляющие статистическую совокупность, называются *единицами* совокупности, а число элементов совокупности – *объемом* совокупности.

С категорией статистической совокупности тесно связан широко известный закон больших чисел. *Законом больших чисел* называется весьма широкий принцип взаимопогашения (уравновешивания) случайных факторов (колебаний), наблюдающихся у индивидуальных явлений, в результате которого могут отчетливее проявиться внутренние необходимые связи явлений.

Закон больших чисел является одним из выражений диалектической связи между случайностью и необходимостью, он помогает выявлять необходимое там, где на поверхности выступает игра случайностей. С помощью закона больших чисел в статистических совокупностях устанавливаются имеющиеся в явлениях необходимые закономерные уровни и соотношения – *статистические закономерности*. Статистическая закономерность по своей природе близка к закону. Она так же, как и закон, отражает необходимые причинно-следственные связи. Однако эти связи здесь менее устойчивы, не всеобщи, как в законе, а относятся к определенному пространству и времени, справедливы лишь для данных условий развития конкретных явлений.

Связь и различие между статистикой и математикой заключается в том, что обе эти науки исследуют количественную сторону явлений, но математика исследует количественную сторону всех явлений (природы и общества) безотносительно к качеству, а статистика – количественную сторону лишь общественных явлений и всегда определенного качества.

В статистике применяется математика различных уровней. Длительное время статистики обходились в своей работе простейшими приемами элементарной математики (правилами арифметики, алгебраическими выражениями и т.п.). Но необходимость познания массовых случайных процессов вызвала к жизни и призвала на помощь статистикам специальный раздел высшей математики – математическую статистику. Исследованием случайных процессов занимается теория вероятностей.

Математическая статистика – раздел математики, посвященный математическим методам систематизации, обработки и использования статистических данных для научных и практических выводов.

Метод исследования, опирающийся на рассмотрении статистических данных о тех или иных совокупностях объектов, называется *статистическим*. Статистический метод применяется в самых различных областях знания. Однако черты статистического метода в применении к объектам различной природы столь своеобразны, что было бы бессмысленно объединять, например, социально-экономическую статистику, физическую статистику, звездную статистику и т.п. в одну науку. Общие черты статистического метода в различных областях знания сводятся к подсчету числа объектов, входящих в те или иные группы, рассмотрению распределения количественных признаков, применению выборочного метода, использованию теории вероятностей при оценке достаточности числа наблюдений для тех или иных выводов и т.п. Эта формальная математическая сторона статистических методов исследования, безразличная к специфической природе изучаемых объектов, и составляет *предмет математической статистики*.

Связь математической статистики с теорией вероятностей имеет в разных случаях различный характер. Теория вероятностей изучает не любые массовые явления, а явления случайные и именно «вероятностно случайные», т.е. такие, для которых имеет смысл говорить о соответствующих им распределениях вероятностей. *Случайное событие* – это событие, которое может наступить, в тех же условиях – не наступить или происходить иначе.

Теория вероятностей – математическая наука, позволяющая по вероятностям одних случайных событий находить вероятности других случайных событий, связанных каким-либо образом с первыми.

Тем не менее теория вероятностей играет определенную роль и при статистическом изучении массовых явлений любой природы, которые могут не относиться к категории вероятностно случайных. Это осуществляется через основанные на теории вероятностей теорию выборочного метода и теорию ошибок. В этих случаях вероятностным закономерностям подчинены не сами изучаемые явления, а приемы их исследования.

Методы математической статистики позволяют решать несколько типов исследовательских задач:

1) задачи статистического описания совокупности объектов;

2) задачи статистического оценивания параметров генеральной совокупности по выборочным данным;

3) задачи статистического анализа взаимосвязей;

4) задачи классификации объектов или признаков;

5) задачи сжатия информации.

Рассмотрим, как решаются эти задачи в исторических исследованиях с помощью основных математико-статистических методов.

2.2. Методы дескриптивной (описательной) статистики

Для анализа статистической совокупности прежде всего используются обобщающие количественные показатели, которые позволяют описать изучаемое явление или процесс в целом, показывая тенденцию его развития. Основными описательными характеристиками статистической совокупности являются средняя арифметическая, дисперсия и среднее квадратическое отклонение.

Прежде чем приступить к изучению статистической совокупности, необходимо на содержательном уровне выявить, является ли она качественно однородной. Широко известно, например, что земские статистики абсолютизировали однородность российского крестьянства, поэтому опубликованные сводные данные земско-статистических обследований преврати-

лись в тома средних цифр, нивелирующих существенные различия в экономическом состоянии разных типов крестьянских хозяйств.

Для анализа статистической совокупности удобно ее упорядочить в возрастающем или убывающем порядке, такая совокупность называется *вариационным (ранжированным) рядом*, а единицы совокупности – *вариантами* (обозначаются x_i, где i – номер варианты). Изменение (вариация) признака, по которому обследуются объекты, может быть дискретным или непрерывным. При дискретной вариации значения варианты отличаются на некоторую конечную величину и вариационный ряд называется *дискретным*. При непрерывной вариации отдельные значения признака могут отличаться на сколь угодно малую величину и вариационный ряд называется *интервальным*.

Существуют две группы характеристик вариационного ряда: средние величины и меры вариации (рассеяния) признака. *Средняя* представляет собой количественную характеристику качественно однородной совокупности. Наиболее распространенными средними являются средняя арифметическая, мода и медиана.

Средняя арифметическая ($\bar{x}$) – обобщающий показатель, выражающий типичные размеры количественных признаков качественно однородных явлений, определяется по формуле:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (2.2.1)$$

где x_i - варианта с порядковым номером i (i=1,…n); n – объем совокупности.

Мода (*Mo*) – варианта, которая чаще всего встречается в данном вариационном ряду.

Медиана (*Me*) – варианта, находящаяся в середине вариационного ряда:

Me=x_{m+1}, если число вариант нечетно (n=2m+1);

Me=$\frac{x_m + x_{m+1}}{2}$, если число вариант четно (n=2m).

Медиана используется, когда изучаемая совокупность неоднородна. Особое значение она приобретает при анализе ассиметричных рядов (рядов, у которых нагружены крайние значения вариант). Медиана дает более верное представление о среднем значении признака, т.к. она не столь чувствительна к крайним (нетипичным в плане постановки задачи) значениям как средняя арифметическая.

Средние позволяют охарактеризовать статистическую совокупность одним числом, однако не содержат информации о том, насколько хорошо они представляют эту совокупность. Для определения того, насколько сильно варьируются значения признака, используются такие характери-

стики, как размах вариации, дисперсия и среднее квадратическое отклонение.

Размах вариации (R) – это разность между наибольшим и наименьшим значениями признака:

$$R = x_{\max} - x_{\min}. \tag{2.2.2}$$

Показатель этот достаточно просто рассчитывается, однако является наиболее грубым из всех мер рассеяния, поскольку при его определении используются лишь крайние значения признака, а все другие просто не учитываются.

При расчете двух других характеристик меры вариации признака используются отклонения всех вариант от средней арифметической. Эти характеристики (дисперсия и среднее квадратическое отклонение) нашли самое широкое применение почти во всех разделах математической статистики.

Дисперсия (σ^2) – абсолютная мера вариации (колеблемости) признака в статистическом ряду - средний квадрат отклонения всех значений признака ряда от средней арифметической этого ряда:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \tag{2.2.3}$$

где x_i - варианта с порядковым номером i; $\bar{x}$ - средняя арифметическая;

n – объем совокупности.

Для представления меры вариации в тех же единицах, что и варианты, используется среднее квадратическое отклонение.

Среднее квадратическое отклонение (σ) – это квадратный корень из дисперсии:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}. \tag{2.2.4}$$

Рассмотренные меры рассеяния – абсолютные величины. Однако часто бывает необходимо сравнить вариацию одного и того же признака у разных групп объектов, выявить степень различия одного и того же признака у одной и той же группы объектов в разное время, сопоставить вариацию разных признаков у одних и тех же групп объектов. Для решения этих задач необходимо использовать относительные показатели. Таким показателем является коэффициент вариации.

Коэффициент вариации (V) – это отношение среднего квадратического отклонения к средней арифметической, выраженное в процентах:

$$V = \frac{\sigma}{\bar{x}} 100\%. \tag{2.2.5}$$

Пример 1.

Даны две группы людей, возраст которых (в годах):

1 группа: 27; 29; 30; 31; 31; 32;

2 группа: 13; 14; 14; 15; 61; 63.

Вычислим средний возраст для каждой группы. Получим, что и в первой, и во второй группе средний возраст одинаков и равен 30 годам. Тогда как очевидно, что для первой группы эта величина представительна, в ней действительно собраны 30-летние, а вторую группу она абсолютно не характеризует, т.к. в ней – подростки и пенсионеры. Тогда обратимся к характеристикам меры вариации признака.

Вычислим среднее квадратическое отклонение и коэффициент вариации для обеих групп по формулам (2.2.4) и (2.2.5). Получим $\sigma_1 \approx 1{,}63$, $\sigma_2 \approx 22{,}64$, $V_1 \approx 5\%$, $V_2 \approx 75{,}5\%$.

Таким образом, сравнение коэффициентов вариации позволяет говорить о значительных различиях рассматриваемых групп: первая группа представляет собой достаточно однородную совокупность, а вторая группа таковой не является.

Важную роль в изучении вариационных рядов играет их графическое изображение (термин «дескриптивный» переводится не только как «описательный», но и как «изобразительный», «наглядный»). Существует несколько способов графического изображения рядов (диаграмма, гистограмма, полигон, кумулята и др.), выбор которых зависит от вида вариационного ряда и цели исследования. Однако общим для всех типов графиков является то, что они показывают частоту встречаемости различных значений данного признака - *распределение* значений признака.

Пример 2.

В архивных фондах ГАКО выявлено 288 анкет-заявлений глав переселенческих семей, прибывших в колхозы и совхозы Калининградской области в 1947 году согласно правительственной программе заселения и освоения сельских районов нового края[6]. Анализ содержания анкет-заявлений позволил выделить основные признаки, которые служат хорошей иллюстрацией социального облика переселенца. Рассмотрим, например, признак «стаж работы в колхозе». Средний стаж работы в колхозе составлял 10 лет (при среднем возрасте 36 лет). Однако это число нивелирует имевшие место существенные различия в стаже. Рассмотрим диаграмму и гистограмму распределения переселенцев по стажу работы в колхозе.

[6] ГАКО. Ф.183, оп. 5, ед. хр. 38, 39, 42, 44, 46, 50, 54, 64.

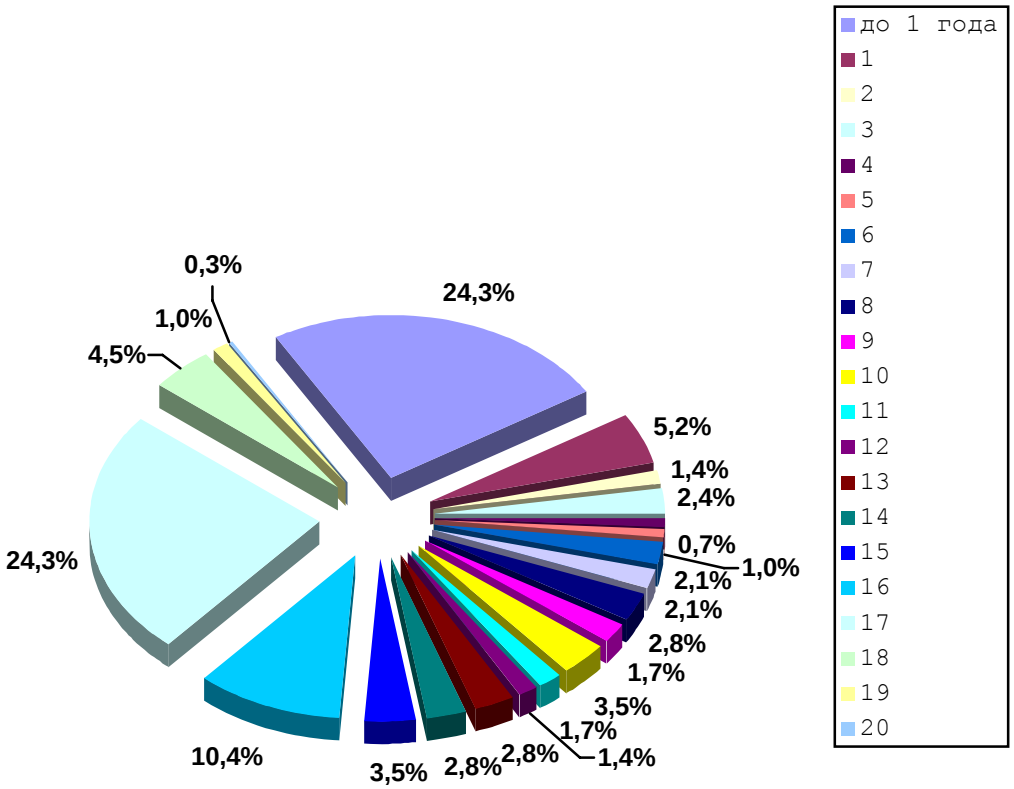

Рис. 1. Диаграмма распределения переселенцев по стажу работы в колхозе

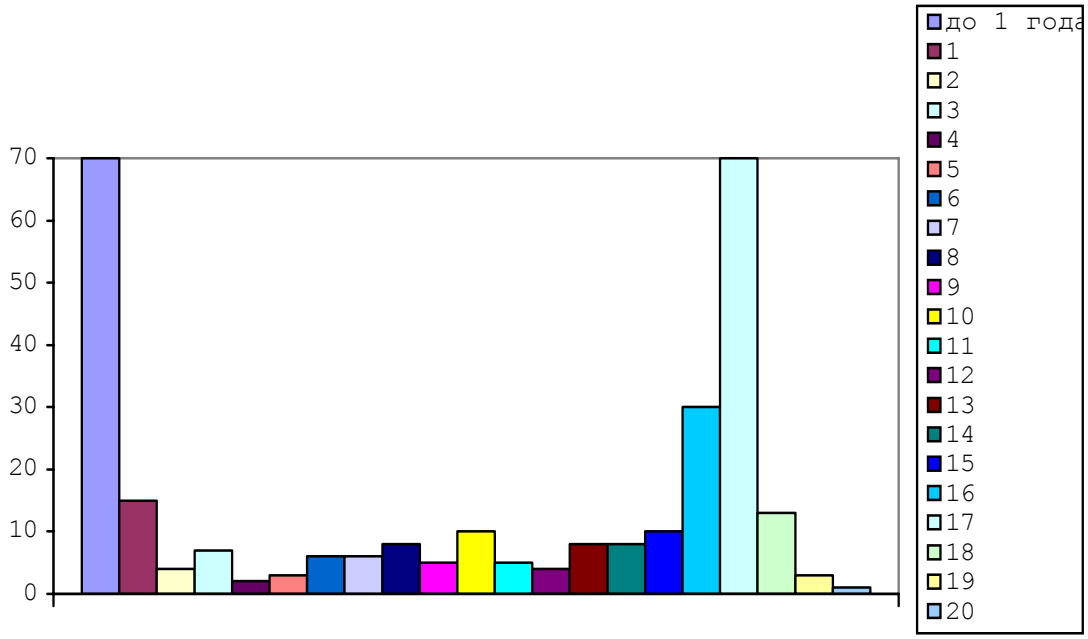

Рис. 2. Гистограмма распределения переселенцев по стажу работы в колхозе

Как показывают графики, выделилось две крупные группы (по 70 человек каждая, т.е. по 24,3%), охватив около половины всех переселенцев, со стажем менее года и со стажем 17 лет. Это свидетельствует о том, что население сельских районов новой области в первую очередь формировалось как теми, кто работал в колхозах страны с начала коллективизации (с 1930 г.), так и людьми, еще вчера не имевшими отношения к сельскому хозяйству (значительную часть последней категории составляли демобилизованные из Советской Армии).

Часто графическое изображение распределения значений признака используется для его сопоставления с *нормальным*, т.е. для проверки гипотезы о том, что значения данного признака *распределены по нормальному закону*. Нормальное распределение играет особую роль в теоретико-прикладном плане, поскольку нормальность является существенным условием корректности применения статистических методов.

Графически нормальное распределение изображается в виде симметричной одновершинной кривой, напоминающей по форме колокол. Высота (ордината) каждой точки этой кривой показывает, как часто встречается соответствующее значение. Форма нормальной кривой и положение ее на оси абсцисс полностью определяются двумя параметрами: средним арифметическим значением $\bar{x}$ и средним квадратическим отклонением σ. Вершина кривой соответствует среднему арифметическому значению, т.е. наиболее часто встречаются значения, близкие к среднему, а по мере удаления от него частота падает.

Каждому значению признака *x* соответствует определенное значение так называемой функции распределения $F(x)$, показывающее, какова вероятность существования значений, меньших данного значения *x*. Геометрически вероятность значений, меньших данного *x*, изображается площадью под кривой распределения слева от этого значения. Площадь под всей кривой равна 1, что соответствует полной достоверности, т.е. вероятности того, что признак вообще принимает какое-то (любое) значение.

В силу своей важности для практических приложений функция нормального распределения табулирована, т.е. существуют специальные таблицы, в которых каждому значению *x* ставится в соответствие вероятность *F(x)* существования значений, меньших *x*. Для удобства табулирования в качестве значений признака берутся не сами величины *x*, а так называемые нормированные отклонения их от среднего значения *t*, где $t = \frac{x - \bar{x}}{\sigma}$.

При замене *x* на *t* центр распределения смещается в точку 0, а единицей измерения становится величина среднего квадратического отклонения σ, но вид кривой распределения не изменяется. Среднее значение норми-

рованного отклонения t равно 0, а его среднее квадратическое отклонение равно 1. Нормированная функция нормального распределения обладает следующими свойствами: $F(-\infty)=0;\ F(\infty)=1;\ F(0)=\frac{1}{2};\ F(-t)=1-F(t)$.

2.3. Выборочный метод

Множество всех единиц статистической совокупности называется *генеральной совокупностью*.

На практике по тем или иным причинам не всегда возможно или же нецелесообразно рассматривать всю генеральную совокупность. Одна из двух проблем очень часто стоит перед историком: как по немногим сохранившимся данным получить широкую и достоверную историческую картину и как из многочисленных сведений отобрать минимальное количество данных, по которым можно было бы судить обо всем явлении в целом. Обе проблемы удовлетворительно решаются с помощью хорошо разработанного в математической статистике выборочного метода.

Из генеральной совокупности особым образом отбирается часть элементов - формируется *выборка*, и результаты обработки выборочных данных распространяются на всю генеральную совокупность. Теоретической основой выборочного метода является закон больших чисел.

Однако для характеристики всей генеральной совокупности могут служить лишь *репрезентативные* (представительные) выборки, т.е. выборки, которые правильно отражают свойства генеральной совокупности. В статистике доказано: чтобы выборка была репрезентативной, она должна быть *случайной*, т.е. каждая единица генеральной совокупности должна иметь равный шанс попасть в выборку.

Таким образом, задачей исследователя, в распоряжении которого имеются сплошные данные, является организация выборочного изучения этих данных путем формирования репрезентативной выборки. Если же он имеет дело с данными ранее проведенных выборочных обследований, необходимо проверить, как были организованы эти обследования, не нарушались ли принципы случайного отбора. Сложнее решить вопрос о репрезентативности так называемых «естественных выборок», поскольку надежных математических методов проверки их репрезентативности не существует. Здесь на первый план выступает изучение истории происхождения данных и их содержательный анализ.

Существует несколько видов выборочного изучения, позволяющих формировать репрезентативные выборки: случайный, механический, типический и серийный отбор.

Случайным является такой отбор, при котором все элементы генеральной совокупности имеют равную возможность быть отобранными. На практике случайный отбор производится с помощью жеребьевки или использования разработанных в статистике таблиц случайных чисел. При жеребьевке может осуществляться бесповторный отбор (когда выбранный элемент больше не участвует в выборке) или повторный (когда ему предоставляется шанс еще раз быть выбранным). При большом объеме генеральной совокупности проведение жеребьевки или использование таблиц случайных чисел становятся затруднительными, тогда применяют другие виды выборочного изучения.

Механический отбор сводится к тому, что генеральная совокупность разбивается на равные части и из каждой части берется одна единица. Например, 7, 17, 27, 37 и т.д.

Однако механическим отбором следует пользоваться очень осторожно, поскольку элементы исходной совокупности могут быть упорядочены, что может привести к возникновению систематических ошибок. Необходимо проанализировать изучаемую совокупность и применять механический отбор лишь в том случае, если элементы генеральной совокупности расположены случайным образом.

Механический отбор достаточно широко использовался в русской статистике. Например, механический отбор применялся земскими статистиками для обследований части крестьянских хозяйств не по обычной подворной карточке, а по особой расширенной программе. С помощью механического отбора изучалось состояние 25 млн. крестьянских хозяйств и накануне сплошной коллективизации, когда они были подвергнуты 10%-ному весеннему опросу и 5%-ному осеннему опросу.

Типический отбор заключается в том, что генеральная совокупность разбивается на типические группы, образованные по какому-либо признаку. Затем из каждой выделенной группы отбираются единицы либо случайно, либо механически. Например, территория, подлежащая обследованию, разделяется на районы, отличающиеся социально-экономическими или географическими условиями, и из каждого района производят отбор единиц в выборку. При этом допускается как отбор, пропорциональный численности отдельных типических групп, так и непропорциональный. Понятно, что более предпочтительным является пропорциональный отбор, поскольку он дает более точные результаты.

Серийный отбор предусматривает разбиение всей генеральной совокупности на группы (серии), из которых путем случайного или механического отбора выделяется их определенная часть, которая и подвергается сплошной обработке. Фактически, серийный отбор представляет собой

случайный или механический отбор, произведенный для укрупненных элементов исходной совокупности. Например, обследуются не единичные крестьянские хозяйства, а целые деревни или имения.

Итак, выборочный метод позволяет экстраполировать результаты обследования выборки на всю генеральную совокупность. При этом надо иметь в виду, что всегда будет возникать некоторая ошибка, показывающая, насколько хорошо характеристики выборки отражают соответствующие характеристики генеральной совокупности.

Ошибки, возникающие при использовании выборочных данных для суждения обо всей генеральной совокупности, называются *ошибками репрезентативности*. Они бывают систематическими и случайными.

Систематические ошибки – ошибки, возникающие при использовании выборочных данных, если не выполняются условия случайного отбора. *Случайные ошибки* – ошибки, возникающие при использовании выборочных данных за счет того, что для анализа всей совокупности используется только ее часть. Величина *ошибки выборки* – это разность между генеральной и выборочной средними.

В математической статистике существуют формулы для вычисления средней ошибки выборки на основе данных той выборки, с которой работает исследователь. Для различных видов выборочного изучения средняя ошибка выборки определяется по-разному. Рассмотрим формулы вычисления средней ошибки выборки при случайном отборе.

Средняя ошибка выборки (μ) *при случайном повторном отборе* определяется формулой:

$$\mu = \frac{\sigma}{\sqrt{n}}, \qquad (2.3.1)$$

где σ - оценка среднего квадратического отклонения в генеральной совокупности по выборке; n – объем выборки.

Средняя ошибка выборки при случайном бесповторном отборе:

$$\mu = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}, \qquad (2.3.2)$$

где N – объем генеральной совокупности.

Средняя ошибка малой выборки, т.е. выборки, объем которой не превышает 30 единиц, вычисляется по формуле:

$$\mu = \frac{\sigma}{\sqrt{n-1}}. \qquad (2.3.3)$$

Средняя ошибка выборки позволяет по выборочной средней судить о значении генеральной средней. Однако в конкретном выборочном исследовании ошибка может существенно отличаться от средней ошибки, превышая ее. Поэтому более эффективным является определение тех границ, в

которых «практически наверняка» находится действительная ошибка, допущенная в данной конкретной выборке. Эти границы определяются *предельной ошибкой выборки* (Δ) по формуле:

$$\Delta = t\mu, \qquad (2.3.4)$$

где t – коэффициент, вычисляемый по специальной таблице; μ - средняя ошибка выборки.

Коэффициент t определяется задаваемой исследователем вероятностью P ($0 \le P \le 1$). Для значений P, приближающихся к единице, практически исключается возможность того, что генеральная средняя будет отличаться от вычисленной выборочной средней больше, чем на Δ. Со своей стороны Δ указывает точность, гарантируемую заданным уровнем надежности (вероятности P). При этом, чем выше уровень вероятности (используются, например, значения 0,90; 0,95; 0,99 и др.), тем выше коэффициент t, а следовательно, и значение предельной ошибки Δ. Поэтому на практике приходится довольствоваться некоторым компромиссом между противоречивыми требованиями максимальной надежности и максимальной точности.

Таким образом, разность между генеральной и выборочной средними не будет превышать по модулю значения предельной ошибки выборки:

$$\left|\bar{x}_{ген} - \bar{x}_{выб}\right| \le \Delta, \qquad (2.3.5)$$

тогда можно определить интервал, в котором практически наверняка находится генеральная средняя, – *доверительный интервал*:

$$\bar{x}_{выб} - \Delta \le \bar{x}_{ген} \le \bar{x}_{выб} + \Delta, \qquad (2.3.6)$$

при этом всегда указывается надежность этого результата (значение P, которое использовалось при вычислении Δ).

Для малой выборки предельная ошибка выборки вычисляется по формуле:

$$\Delta = t(\kappa)\mu, \qquad (2.3.7)$$

где t рассчитывается исходя из так называемого закона распределения Стьюдента с κ степенями свободы (в отличие от больших выборок, где t вычисляется на основе нормального закона распределения), $\kappa = n - 1$.

Связь между коэффициентом t и вероятностью P в распределении Стьюдента сложнее, чем в нормальном распределении и определяется с учетом объема выборки.

Пример 3.

По урожайности зерновых культур 10 колхозов определить среднюю и предельную ошибки выборки и оценить пределы для генеральной средней.

Исходные данные (x_i , i = 1,…10 - урожайность зерновых в центнерах с гектара) и промежуточные вычисления можно записать в таблице:

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	6,5	-0,2	0,04
2	6,2	-0,5	0,25
3	5,4	-1,3	1,69
4	9,3	2,6	6,76
5	7,2	0,5	0,25
6	8,4	1,7	2,89
7	4,3	-2,4	5,76
8	6,0	-0,7	0,49
9	6,3	-0,4	0,16
10	7,4	0,7	0,49

Получим:

$$\bar{x} = 6{,}7;\ \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = 1{,}878;\ \sigma \approx 1{,}37;\ \mu = \frac{\sigma}{\sqrt{n-1}} \approx 0{,}46.$$

Для P=0,95 t=2,26 $\Rightarrow$ Δ=t(κ)μ $\approx$1,04 $\Rightarrow$ $5.66 \le \bar{x}_{ген} \le 7.74$

Очевидно, что полученная предельная ошибка (15%) слишком велика и объем выборки в 10 единиц не достаточен для суждения о реальной средней урожайности зерновых.

Важным вопросом в выборочном методе является определение необходимого объема выборки. Как правило, объем выборки определяется на основе содержательного анализа данных, например, в 10% или 20%. Обычно выборки такого объема бывает достаточно для получения надежных результатов. Однако можно определить объем выборки по специальной формуле. Для этого необходимо:

1) провести пробную 1 %-ную выборку и вычислить для нее выборочную среднюю и дисперсию;

2) задать необходимую предельную ошибку выборки Δ и уровень надежности P;

3) найти *объем выборки* по формуле:

$$n = \frac{t^2 \sigma^2}{\Delta^2}, \qquad (2.3.8)$$

где σ^2 - дисперсия признака, вычисленная по пробной выборке; Δ - заданная точность результатов выборочного исследования (заданная предельная ошибка выборки); t - табличный коэффициент, соответствующий заданной надежности результатов выборочного изучения (вероятности P). Если пробная выборка мала (n<30), то при определении коэффициента t учитывается также объем пробной выборки.

Пример 4.

Для рассмотренных в примере 3 данных об урожайности зерновых культур в колхозах определим требуемый объем выборки.

Зададим предельную ошибку выборки, равную 5%, она будет равна Δ=0,34, тогда, подставляя в формулу (6.8) значения t=2,26; σ=1,37 и Δ, получим n=86. Таким образом, для определения средней урожайности зерновых в колхозах с вероятностью 95% и точностью 5% необходимо произвести выборку, объемом 86 единиц.

2.4. Корреляционный анализ

В реальной исторической действительности существует диалектическое взаимодействие и взаимообусловленность во всех явлениях и процессах. При этом часто воздействие одних признаков на другие осуществляется столь скрыто и опосредованно, что уловить его без специального методического инструментария практически невозможно. Решить эту задачу позволяют хорошо разработанные в статистике методы корреляционного и регрессионного анализа.

Зависимости, которые присущи объективным явлениям природы и общества, делятся на функциональные и статистические.

Функциональная зависимость – это взаимосвязь между признаками, при которой каждому значению одного признака соответствует единственное значение другого признака.

Простейшей формой функциональной связи является линейная зависимость, которая характеризуется уравнением:

$$y = ax + b\,. \qquad (2.4.1)$$

Другими формами функциональной зависимости, применяемыми в статистическом анализе, являются парабола ($y = ax^2 + bx + c$), гипербола ($y = \frac{k}{ax + b}$), логарифмическая функция ($y = a \lg x$), экспонента ($y = ke^{ax},\ k > 0,\ a > 0$).

Функциональная зависимость предполагает изолированность взаимосвязанных признаков от воздействия других факторов. Но такая ситуация в явлениях общественной жизни практически не встречается. Здесь на связь между признаками влияет множество других факторов, и она проявляется лишь в тенденции, «в среднем». Такая зависимость называется статистической, или корреляционной.

Статистическая (корреляционная) зависимость – это взаимосвязь между признаками, при которой одному и тому же значению одного признака могут соответствовать различные значения другого признака.

Для выявления степени статистической зависимости между признаками используются методы корреляционного анализа.

Корреляционный анализ – совокупность методов математической статистики, позволяющих обнаружить корреляционную зависимость между случайными величинами или признаками и оценить значимость этой связи. Теснота связи определяется коэффициентом корреляции.

Основной мерой связи в корреляционном анализе является линейный коэффициент корреляции, который измеряет степень *линейной* зависимости между признаками.

Парный линейный коэффициент корреляции определяет тесноту связи между двумя признаками и рассчитывается по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad (2.4.2)$$

где x_i, y_i - значения признаков *x* и *y* для *i*-го объекта; *n* - число объектов; $\bar{x}$, $\bar{y}$ - средние арифметические значения признаков *x* и *y*.

Линейный коэффициент корреляции может принимать значения от -1 до +1. Чем ближе величина коэффициента корреляции к предельным значениям, тем теснее взаимосвязь между признаками. Равенство коэффициента нулю свидетельствует об отсутствии линейной связи между признаками. Если коэффициент корреляции равен +1 (или -1), то между признаками существует прямая (или обратная) функциональная зависимость.

При содержательном анализе взаимосвязей часто необходимо не только оценить тесноту связи между изучаемыми признаками, но и определить степень воздействия одного признака на другой. Для решения этой задачи используется коэффициент детерминации.

Коэффициент детерминации – показатель, определяющий долю (в процентах) изменений, обусловленных влиянием факторного признака, в общей изменчивости результативного признака:

$$D = r^2 100\%, \qquad (2.4.3)$$

где *r* - коэффициент корреляции.

Пример 5.

Определим степень корреляционной зависимости между доходом и размерами помещичьего хозяйства в России на рубеже XIX-XX вв. по сведениям о размерах (в десятинах) и доходах (в тыс. руб.) десяти помещичьих имений[7].

[7] Данные взяты из книги Миронова Б.Н. История в цифрах. Л., 1991. С.67.

Априори ясно, что доходность имения росла вместе с увеличением его размеров. Однако доходность имения, помимо его размеров, определялась еще качеством земли, состоянием хозяйства, деловыми способностями его владельца, близостью рынка, уровнем агротехники и другими факторами. Поэтому интересно узнать, насколько все-таки доходность определялась именно размерами имения.

Исходные данные (x_i - размеры имения в десятинах, y_i - доход имения в тыс. руб.) и промежуточные вычисления запишем в таблице:

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	240	1,50	-50	-0,10	2500	0,01	5,00
2	255	1,25	-35	-0,35	1225	0,1225	12,25
3	265	1,55	-25	-0,05	625	0,0025	1,25
4	270	1,40	-20	-0,20	400	0,04	4,00
5	285	1,45	-5	-0,15	25	0,0225	0,25
6	295	1,60	5	0	25	0	0
7	310	1,80	20	0,20	400	0,04	4,00
8	320	1,80	30	0,20	900	0,04	6,00
9	325	1,85	35	0,25	1225	0,0625	8,75
10	330	1,90	40	0,30	1600	0,09	12,00

Получим: $\bar{x} = 290$; $\bar{y} = 1{,}60$; $r = \dfrac{54{,}0}{\sqrt{8925 \cdot 0{,}43}} = \dfrac{54{,}0}{61{,}95} \approx 0{,}87$; $D = r^2 100\% = 76\%$.

Таким образом, доход имения примерно на 76% объясняется и обусловливается его размерами и на 24% - другими факторами.

Коэффициент корреляции рассчитывается, как правило, для выборочных данных, поэтому существуют приемы проверки значимости вычисленного коэффициента корреляции для всей генеральной совокупности.

Рассмотрим, как определяется значимость парного линейного коэффициента корреляции для случая малой выборки (практически для n<50):

1) вычисляется статистическая характеристика t_φ, подчиняющаяся закону распределения Стьюдента, по формуле:

$$t_\varphi = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad (2.4.4)$$

где r - вычисленный выборочный коэффициент корреляции; n - объем выборки.

2) t_ϕ сравнивается с табличной, или критической, величиной $t_{кр}$, зависящей от числа $k = n - 2$ и от заданной вероятности P:

а) если $t_ф \geq t_{кр}$, то можно сделать вывод о наличии связи;

б) если $t_ф < t_{кр}$, то гипотеза об отсутствии связи не отклоняется.

Пример 6.

Проверим значимость коэффициента корреляции, вычисленного в пятом примере. Вычислим $t_ф$ по формуле (2.4.4): $t_ф \approx 5{,}02$. Зададим вероятность $P = 0{,}99$, найдем для этой вероятности табличное значение $t_{кр} = 3{,}36$, получаем $t_ф > t_{кр}$.

Таким образом, с вероятностью 99% связь между доходностью и размерами помещичьих имений существует.

Зависимость между тремя и большим числом признаков изучается методами многомерного корреляционного анализа с помощью вычисления частных и множественных коэффициентов корреляции[8].

2.5. Регрессионный анализ

Анализ статистической зависимости предполагает не только оценку тесноты связи между признаками, но и выявление ее формы. Эта задача решается методами регрессионного анализа.

Регрессионный анализ – это совокупность методов математической статистики, позволяющих определить форму связи между результативным и факторным признаками, установленной корреляционным анализом. Корреляционная связь описывается с помощью уравнения регрессии.

Уравнение регрессии – это описание корреляционной связи с помощью подходящей функции.

Простейшее уравнение *линейной регрессии* имеет вид:

$$y = ax + b, \qquad (2.5.1)$$

где x - факторный признак; y - результативный признак; a и b - параметры уравнения, которые могут быть найдены *методом наименьших квадратов* по формулам:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}, \quad b = \bar{y} - a\bar{x}, \qquad (2.5.2)$$

где x_i, y_i - i-е значение признаков x и y соответственно; $\bar{x}$, $\bar{y}$ - средние арифметические признаков x и y; n - число значений признаков x и y.

[8] О множественной и нелинейной корреляции см.: Количественные методы в исторических исследованиях. М., 1984. Гл. 6. §2, 4.

Коэффициент *a* называется *коэффициентом регрессии*. Он показывает, на какую величину в среднем изменяется результативный признак *y* при изменении факторного признака *x* на единицу.

Если коэффициент регрессии положительный, то между результативным и факторным признаками наблюдается прямая зависимость: с ростом значения факторного признака значение результативного признака растет, и, наоборот, с уменьшением значения факторного признака значение результативного признака уменьшается. Если же коэффициент регрессии отрицательный, между признаками наблюдается обратная связь: с ростом значения факторного признака значение результативного признака уменьшается, и, наоборот, с уменьшением значения факторного признака значение результативного признака растет.

Метод наименьших квадратов позволяет выбрать «наилучшую» среди всех возможных прямых в том смысле, что она проходит «ближе всего» к точкам *диаграммы рассеяния* - изображения объектов как точек на плоскости двух признаков.

Пример 7.

Найдем уравнение линейной регрессии, описывающее корреляционную связь между размерами и доходом помещичьего имения по данным примера 5. Запишем промежуточные вычисления в таблице:

	x_i	y_i	$x_i\ y_i$	x_i^2
1	240	1,50	360,00	57600
2	255	1,25	318,75	65025
3	265	1,55	410,75	70225
4	270	1,40	378,00	72900
5	285	1,45	413,25	81225
6	295	1,60	475,00	87025
7	310	1,80	558,00	96100
8	320	1,80	576,00	102400
9	325	1,85	601,25	105625
10	330	1,90	627,00	108900
Σ	2895	16,1	4715,0	847025

Вычислим параметры *a* и *b* по формулам (2.5.2):

$a = \dfrac{10 \cdot 4715 - 2895 \cdot 16{,}1}{10 \cdot 847025 - 2895^2} = 0{,}00606$, b=1,61–0,00606·290= –0,1474.

Уравнение линейной регрессии примет вид: y=0,00606x–0,1474. Коэффициент регрессии в этом уравнении, равный 0,00606, означает, что при возрастании размеров имения на единицу, т.е. на 1 десятину, доход имения

возрастает на 0,00606 тыс. рублей, или на 6,06 рублей. С помощью уравнения регрессии можно предсказать примерный доход имения любых размеров.

Изобразим графически диаграмму рассеяния по данным десяти имений и прямую регрессии, описываемую полученным уравнением линейной регрессии (рис. 3).

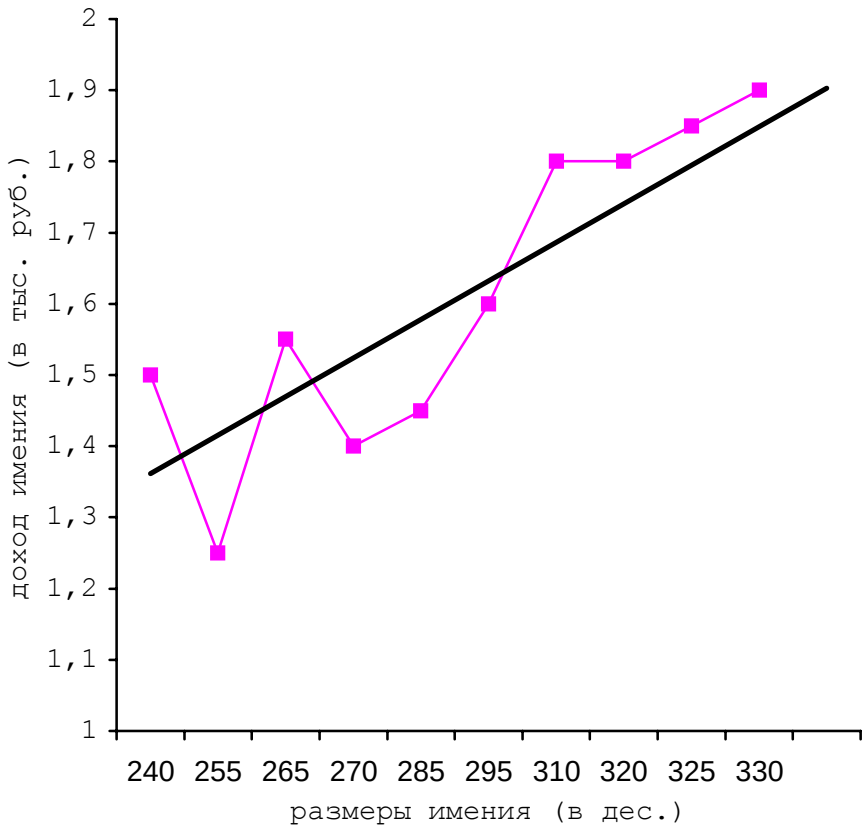

Рис. 3. График зависимости дохода помещечьего имения от его размеров

Прямая регрессии показывает тенденцию в изменении дохода имения в зависимости от его размеров.

Мы рассмотрели лишь наиболее простую форму связи между двумя признаками - линейную. Однако, во-первых, зависимости между признаками могут принимать самые разнообразные формы, а, во-вторых, при более полном анализе взаимосвязей необходимо учитывать, что на результативный признак обычно влияет не один фактор, а несколько. Выявить форму связи между результативным признаком и несколькими факторными признаками позволяет множественный регрессионный анализ[9].

[9] Подробнее о методах регрессионного анализа см.: Количественные методы в исторических исследованиях. Гл. 6. §3, 4.

2.6. Кластерный анализ

Важнейшей задачей исторической науки является классификация изучаемых объектов и явлений. Традиционно такая классификация сводится к группировке объектов на основе одного (двух-трех) признаков. Однако современные методы многомерного статистического анализа и компьютерные технологии позволяют учитывать при группировке все существенные структурно-типологические признаки (их может быть несколько десятков). Методы, на основе которых все схожие объекты можно собрать в одну группу, и при этом объекты из разных групп будут существенно отличаться, составляют совокупность *методов автоматической классификации* (кластерного анализа, таксономии).

Кластерный анализ – совокупность методов, составляющих раздел многомерного статистического анализа, с помощью которых осуществляется построение многомерной классификации объектов. Основная идея кластерного анализа заключается в последовательном объединении группируемых объектов по принципу наибольшей близости – схожести свойств. Процедура построения классификации состоит из последовательности шагов, на каждом из которых производится объединение двух ближайших групп объектов (кластеров[10]).

Рассмотрим *агломеративно-иерархический метод* кластерного анализа.

Пусть существует *n* объектов, каждый из которых характеризуется набором из *m* признаков. Каждый из этих объектов может быть представлен точкой в *m*-мерном пространстве признаков. О сходстве объектов можно судить по расстоянию между соответствующими точками: чем ближе точки расположены друг к другу, тем более схожи их свойства. Евклидово расстояние между точками определяется формулой:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \quad (\text{i, j} = 1,2,\ldots,\text{n}), \qquad (2.6.1)$$

где d_{ij} - евклидово расстояние между *i*-м и *j*-м объектами; x_{ik} - значение *k*-го признака для *i*-го объекта.

Подсчитав значения расстояний для всех пар объектов, получим квадратную симметричную матрицу *D* размером $n \times n$ ($d_{ij} = d_{ji}$, d_{ii}=0). На основе матрицы *D* можно вычислить расстояния между кластерами. Близость двух кластеров определяется как среднее значение расстояния между всеми такими парами объектов, где один объект пары принадлежит к одному кластеру, а другой - к другому:

[10] Кластер (англ. cluster - гроздь, скопление) – группа объектов, характеризующихся общими свойствами.

$$D_{pq}^2 = \sum_{i \in X_p} \sum_{j \in X_q} \frac{d_{ij}^2}{n_p n_q}, \qquad (2.6.2)$$

где D_{pq}^2 - мера близости между *p*-м и *q*-м кластерами; X_p - *p*–й кластер; X_q - *q*-й кластер; n_p , n_q - число объектов в *p*-м и *q*-м кластерах соответственно.

На первом шаге процедуры построения классификации в матрице расстояний D выбирается минимальное расстояние между объектами и объекты, находящиеся друг от друга на этом расстоянии, объединяются в один кластер. В матрице вычеркиваются строка и столбец, соответствующие первому из этих объектов, а расстояния от полученного кластера до всех остальных объектов вычисляются по формуле (2.6.2) и заносятся в строку и столбец матрицы расстояний, соответствующие второму объекту из первого кластера.

На втором шаге в матрице, содержащей уже *n*-1 строк и столбцов, снова выбирается минимальное расстояние и формируется новый кластер. Этот кластер может быть построен в результате объединения либо двух объектов, либо одного объекта с первым кластером. В матрице вычеркиваются строка и столбец и пересчитываются расстояния до второго кластера, и т.д.

Таким образом, процедура агломеративно-иерархического метода кластерного анализа состоит из *n*-1 аналогичных шагов, на каждом из которых происходит объединение двух ближайших кластеров (на первых шагах – объектов). В конце этой процедуры, на (*n*-1)-м шаге, получается кластер, объединяющий все *n* объектов.

Результаты построения многомерной классификации обычно изображают в виде дерева иерархической структуры (*дендрограммы*), содержащего *n* уровней, каждый из которых соответствует одному из шагов последовательного укрупнения кластеров.

Существенным вопросом в кластерном анализе является установление необходимого и достаточного числа кластеров. Как правило, это число определяется из показателей однородности и близости кластеров – внутригрупповой вариации.

Пример 8.

Рассмотрим результаты кластерного анализа 10 уездов Новгородской губернии на основе земско-статистических данных, характеризующих крестьянское хозяйство Новгородской губернии на уездном уровне.

Исходя из содержательного анализа набора показателей поуездных сводок земских переписей, было выделено 19 относительных признаков

группировки. Результаты построения с помощью кластерного анализа классификации 10 объектов (уездов Новгородской губернии) в 19-мерном пространстве признаков отражены на рис. 4.

Представленная дендрограмма наглядно раскрывает структуру классификации уездов Новгородской губернии в системе показателей крестьянского хозяйства. Исследуемые объекты разделились на три кластера, в каждый из которых вошли наиболее сходные в аграрном отношении уезды. Близость их выражается межкластерным расстоянием. Образованные кластерами районы губернии можно условно именовать «северный» (I), «центральный» (II) и «южный» (III). В северный район входят три северных территориально смежных уезда – Белозерский, Тихвинский и Устюженский; в южный – два южных (Демянский и Валдайский); центральный район образуют три западных (Новгородский, Крестецкий и Старорусский) и два северо-восточных (Кирилловский и Череповецкий) уезда[11].

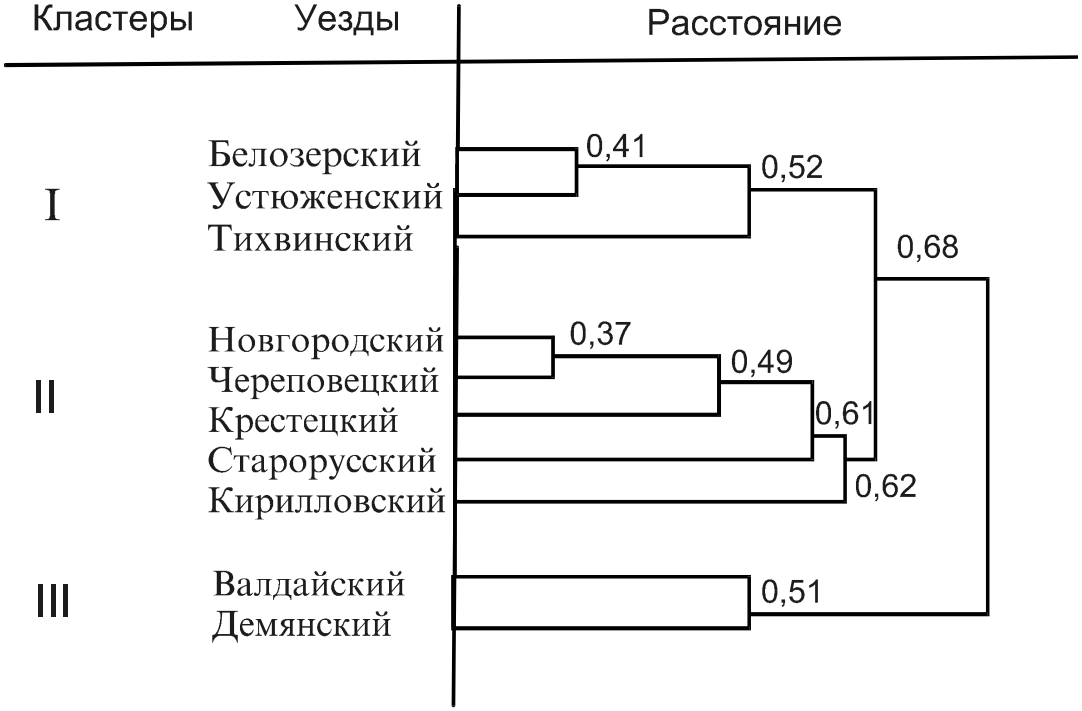

Рис. 4. Структура многомерной классификации уездов Новгородской губернии (дендрограмма)

[11] Результаты кластерного анализа уездов Новгородской губернии подробно обсуждаются в статье: Шендерюк М.Г. Опыт многомерной группировки уездов Новгородской губернии // Северо-Запад в аграрной истории России: Межвуз. темат. сб. науч. тр. / Калинингр. ун-т. Калининград, 1994. С.100-109.

2.7. Факторный анализ

Методы корреляционного анализа позволяют выявить структуру взаимосвязей признаков, характеризующих изучаемое явление или процесс, но они не дают ответа на вопрос: чем обусловлена именно такая структура связей? Известно, что связь между признаками может объясняться не только их взаимозависимостью, но и воздействием на рассматриваемые признаки неких общих, скрытых, глубинных причин – *общих факторов*, измерить которые непосредственно невозможно. Определить причины, обусловившие данную структуру взаимосвязей признаков, можно с помощью методов факторного анализа.

Факторный анализ – раздел многомерного статистического анализа, объединяющий методы анализа структуры множества признаков, характеризующих изучаемые явления и процессы, и выявления обобщенных факторов. Основное предположение факторного анализа заключается в том, что корреляционные связи между большим числом наблюдаемых показателей определяются существованием меньшего числа гипотетически наблюдаемых показателей или факторов.

Объясняя множество исходных признаков через небольшое число общих факторов, факторный анализ осуществляет *сжатие информации*, содержащейся в исходных коррелированных признаках.

Основными характеристиками факторного анализа являются факторные нагрузки и факторные веса.

Факторные нагрузки - это значения коэффициентов корреляции каждого из исходных признаков с каждым из выявленных факторов. Чем теснее связь данного признака с рассматриваемым фактором, тем выше значения соответствующих факторных нагрузок. Положительный знак факторной нагрузки указывает на прямую (а отрицательный знак – на обратную) связь данного признака с фактором. Значение факторной нагрузки, близкое к нулю, говорит о том, что этот фактор практически не влияет на данный признак.

Таблица факторных нагрузок (табл. 1) содержит m строк (по числу признаков) и k столбцов (по числу факторов).

Данные о факторных нагрузках позволяют судить о выборе исходных признаков, отражающих тот или иной фактор, и об относительной доле отдельных признаков в структуре каждого фактора.

Факторные веса – это количественные значения (мера проявления) выделенных факторов для каждого из n имеющихся объектов. Объектам с большими значениями факторных весов свойственна большая степень проявления свойств, присущих данному фактору, т.е. большая степень их развития в соответствующем фактору аспекте. В большинстве методов

факторного анализа (например, в центроидном, в методе главных компонент, в методе экстремальной группировки параметров и др.) факторы определяются как стандартизированные показатели со средним арифметическим значением 0 и средним квадратическим отклонением 1. Поэтому положительные факторные веса соответствуют тем объектам, которые характеризуются степенью проявления свойств больше средней, а отрицательные факторные веса соответствуют тем объектам, в которых степень проявления свойств меньше средней.

Таблица 1

Факторные нагрузки

№ признаков	№ факторов					
	1	2	…	j	…	k
1	a_{11}	a_{12}	…	a_{1j}	…	a_{1k}
2	a_{21}	a_{22}	…	a_{2j}	…	a_{2k}
.	.....					
i	a_{i1}	a_{i2}	…	a_{ij}	…	a_{ik}
.	.....					
m	a_{m1}	a_{m2}	…	a_{mj}	…	a_{mk}
Вклады факторов	V_1^2	V_2^2	...	V_j^2	...	V_k^2

Таблица факторных весов (табл. 2) содержит n строк (по числу объектов) и k столбцов (по числу факторов).

Таблица 2

Факторные веса

№ объектов	№ факторов					
	1	2	…	j	…	k
1	b_{11}	b_{12}	…	b_{1j}	…	b_{1k}
2	b_{21}	b_{22}	…	b_{2j}	…	b_{2k}
.	.....					
i	b_{i1}	b_{i2}	…	b_{ij}	…	b_{ik}
.	.....					
n	b_{n1}	b_{n2}	…	b_{nj}	…	b_{nk}

Данные о факторных весах определяют ранжировку объектов по каждому фактору. Значения факторных весов можно рассматривать как значе-

ния индекса, характеризующего уровень развития объектов в рассматриваемом аспекте.

Факторные веса могут быть основой для классификации исследуемых объектов. Создание многомерной типологии на основе факторного анализа оказывается особенно эффективным, когда имеется большое число признаков, характеризующих совокупность объектов, а их содержательный отбор представляет значительные трудности – выбрать наиболее информативные критерии группировки бывает далеко не просто. В такой ситуации необходимо начать со «сжатия» информации, а затем проводить классификацию по любому из выделенных факторов. При этом даже если группировка осуществляется на основе лишь одного фактора, она будет многомерной, поскольку даже в этом случае учитываются несколько исходных показателей.

Примером эффективного использования факторного анализа в историческом исследовании служит работа И.Д.Ковальченко и Л.И.Бородкина, посвященная изучению аграрной структуры районов Европейской России на рубеже XIX-XX веков[12]. Факторный анализ аграрного развития губерний Европейской России позволил исследователям не только охарактеризовать основные компоненты аграрной структуры и определить их сравнительные доли, но и получить обобщенные характеристики общего уровня аграрного развития отдельных районов и губерний страны.

Надо отметить, что область аграрно-исторических исследований является наиболее широким полем применения факторного анализа. Так, например, интересны результаты многомерной классификации 290 общин Симбирской губернии по данным 34 исходных показателей земских подворных переписей, осуществленной К.Б.Литваком на основе метода экстремальной группировки параметров факторного анализа[13]. С целью получить модели хозяйства зажиточного, беднейшего и среднего крестьянства автор объединил 34 исходных показателя в один фактор хозяйственной состоятельности крестьянского хозяйства, затем всю совокупность из 290 общин разбил на три группы. По мнению К.Б.Литвака, такая методика значительно эффективнее традиционных методов классификации, поскольку в данном случае отпадает проблема выбора критериев группировки, а образовавшиеся группы селений более однородны.

---

[12] См.: Ковальченко И.Д., Бородкин Л.И. Структура и уровень развития районов Европейской России на рубеже XIX-XX веков (Опыт многомерного анализа) // История СССР. 1981. №1.

[13] См.: Литвак К.Б. О пределах информативности пообщинных сводок земских переписей при изучении типов крестьянских хозяйств // Математические методы и ЭВМ в исторических исследованиях. М., 1985.

В данном разделе были рассмотрены основные методы математической статистики, нашедшие самое широкое применение в исторических исследованиях. При этом за пределами изложения остались такие важные сюжеты, как статистический анализ динамических рядов, анализ взаимосвязей качественных признаков, дисперсионный анализ и др. Для освоения этих методов рекомендуется обращение к специальной литературе и пакетам статистических программ (например, к пакету STATISTICA).

Раздел 3. ИСТОЧНИКОВЕДЧЕСКИЕ ЗАДАЧИ

3.1. Компьютерное источниковедение

В клиометрических исследованиях трудно отделить этап исторического построения от собственно источниковедческого анализа, поскольку все они нацелены на решение конкретных исторических проблем путем освоения новых комплексов массовых источников или извлечения из источника новой, скрытой, информации, т.е. так или иначе носят источниковедческий характер. В связи с этим в центре внимания клиометристов всегда стояли задачи адекватной формализации и репрезентации информации источника, создания баз данных, учитывающих специфику исторических источников.

«Микрокомпьютерная революция» конца 80-х – начала 90-х годов привела к тому, что из квантитативной истории выделилось особое направление, ориентированное на компьютерные технологии анализа исторических источников, - историческая информатика. Предмет и содержание новой дисциплины определены в первом в нашей стране учебнике по исторической информатике, созданном сотрудниками лаборатории исторической информатики им. академика И.Д. Ковальченко кафедры источниковедения Московского государственного университета им. М.В. Ломоносова[14].

Историческая информатика – это научная дисциплина, изучающая закономерности процесса информатизации исторической науки и образования; в основе исторической информатики лежит совокупность теоретических и прикладных знаний, необходимых для создания и использования в исследовательской практике машиночитаемых версий исторических источников всех видов.

Теоретической основой исторической информатики является современная концепция информации (включая социальную информацию) и теоретическое источниковедение, а прикладной – информационные (компьютерные) технологии.

Область интересов исторической информатики включает разработку общих подходов к применению информационных технологий в исторических исследованиях (в том числе – специализированного программного обеспечения); создание исторических баз и банков данных/знаний; применение информационных технологий представления данных и анализа структурированных, текстовых, изобразительных и др. источников; компьютерное моделирование исторических процессов; использование информационных сетей (Internet и др.); развитие и применение мультимедиа

[14] См.: Историческая информатика / Под ред. Л.И.Бородкина, И.М.Гарсковой. М., 1996. С.31.

и других новых направлений информатизации исторической науки; а также применение информационных технологий в историческом образовании.

Новые информационные технологии позволяют реализовывать источнико-ориентированный и проблемно-ориентированный подходы в исследовании, поэтому органическими составляющими исторической информатики являются «источниковедческая» (компьютерное источниковедение) и «аналитическая» компоненты. Обратимся к проблемам компьютерного источниковедения.

Компьютерное источниковедение – это совокупность методов и технологий создания машиночитаемых исторических источников. *Машиночитаемые источники* – это источники, переведенные в «электронную» форму. Однако, поскольку в машиночитаемую часть переводится только часть информации, потенциально содержащейся в источнике, то более корректным и часто употребляемым является термин «*машиночитаемые данные*» (МЧД). Вместе с тем машиночитаемые версии источников могут рассматриваться и как новые источники – машиночитаемые источники.

Создание и использование машиночитаемых данных началось в квантитативной истории еще в эпоху больших ЭВМ, когда исследователи не преследовали цель полного перевода источников в машиночитаемую форму и МЧД являлись не только информационной базой, но и результатом исследования. Крупные университеты и исследовательские центры стали коллекционировать машиночитаемые данные. Рост их числа привел к необходимости создания банков и архивов МЧД. С другой стороны, уже с 60-х годов официальные учреждения во многих странах стали производить машиночитаемую информацию, а к 80-м годам в США и Западной Европе около 80% правительственной документации создавалось в машиночитаемой форме. Машиночитаемые данные появились во многих архивах, библиотеках и музеях. Актуальными в связи с этим стали задачи разработки и совершенствования приемов создания и использования коллекций машиночитаемых данных. Микрокомпьютерная революция 80-х гг. открыла для решения этих задач новые перспективы.

Современные компьютерные технологии позволяют создавать машиночитаемые копии источников, максимально приближенные к оригиналу. Это расширяет возможности обработки и анализа данных источников, проведения историко-сравнительных исследований, обращения к архивам данных, созданным другими исследователями.

Коллекции машиночитаемых данных получили название баз данных. В широком смысле *база данных* – это массив данных, хранимый в вычислительной системе. Однако не всякий информационный массив является базой данных в строгом смысле этого понятия, поскольку согласно технологии баз данных организация информации в базе данных должна быть под-

чинена определенным требованиям. Более корректным в этой связи является следующее определение базы данных[15]:

База данных – это совокупность структурированных взаимосвязанных данных при такой минимальной избыточности, которая допускает их использование для различных приложений в определенной предметной области.

Стандартные требования к организации базы данных:

- *Интегрированность* (централизованное хранение информации). Неинтегрированные базы данных по одной и той же проблеме (созданные, например, в разное время и с разными целями) почти неизбежно обладают избыточностью и не являются непротиворечивыми.

- *Взаимосвязанность и структурированность*, отражающие существенные свойства объектов реального мира.

- *Независимость* описания данных от прикладных программ (логическая и физическая независимость), т.е. изменения, касающиеся логической структуры данных, не должны влиять на их расположение в памяти системы.

В современной технологии баз данных эти задачи решаются централизованно с помощью *систем управления базами данных* (СУБД). Главная роль СУБД состоит в обеспечении пользователя необходимыми инструментальными средствами *описания данных* и средствами *манипулирования данными* как на логическом, так и на физическом уровне, а также в обеспечении *защиты данных* (от несанкционированного доступа, от разрушения при сбоях оборудования) и их *целостности* (непротиворечивости).

Проблемы проектирования и работы с базами данных рассматриваются в специальной литературе. Помимо названного учебника по исторической информатике, основные принципы и концепции создания баз данных и их специфика для исторических исследований излагаются в монографии И.М.Гарсковой[16].

Информационные системы на больших ЭВМ, построенные с использованием технологии баз данных, получили название банков данных.

Банк данных – это система информационных, математических, программных, языковых, организационных и технических средств, предназначенных для централизованного накопления и коллективного многоаспектного использования данных для получения необходимой информации.

Основными компонентами банка данных как информационной системы являются (см. рис. 5)[17]:

---

[15] См.: Историческая информатика. С.145-146.

[16] См.: Гарскова И.М. Базы и банки данных в исторических исследованиях. М., 1994.

[17] См. там же. С.54.

1) база данных (БД);
2) система управления базой данных (СУБД);
3) администратор базы данных (АБД);
4) словарь-каталог данных;
5) вычислительная система;
6) обслуживающий персонал.

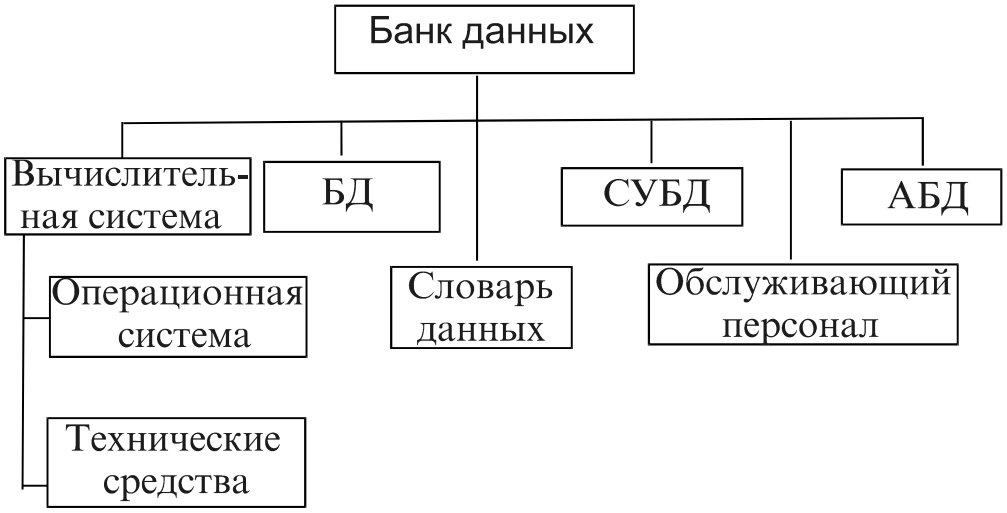

Рис. 5. Основные компоненты банка данных

Как уже отмечалось, появление и использование машиночитаемых данных привели к созданию во многих странах банков и архивов МЧД по различным гуманитарным исследованиям, а в последние годы возникли и специализированные архивы машиночитаемых исторических данных. Перечень архивов и банков данных, которые могут представлять интерес для историка, приводится в таблице 3[18]. Наиболее значительную коллекцию машиночитаемых данных в области социальных наук имеет крупнейший архив – Межуниверситетский Консорциум по политическим и социальным исследованиям (ICPSR) в Анн-Арборе (Мичиган, США), основанный в 1962 г. как сообщество Исследовательского Центра Мичиганского университета и 21 других университетов США. Сейчас в ICPSR входит более 350 колледжей, университетов и архивов, в том числе более 20 иностранных членов (архивов и университетов).

В нашей стране процесс создания банков и архивов машиночитаемых исторических данных находится на начальном этапе. Первые базы данных на материалах исторических источников в строгом понимании этого термина появились в начале 90-х гг. В это же время был создан Консорциум

[18] См.: Гарскова И.М. Указ. соч. С.17.

Таблица 3

Банки данных и архивы МЧД, представляющие интерес для историка

Название банка данных (архива МЧД)	Страна
Australian Social Science Data Archive	Австралия
Banque de Donnees Socio-Politiques	Франция
Belgian Archives for the Social Sciences (BASS)	Бельгия
Center for Machine-Readable Texts in the Humanities	США
Centre de Traitement Electronique des Documents (CETEDOC)	Бельгия
Danish Data Archives (DDA)	Дания
Data Clearing House for the Social Sciences	Канада
Data Library Computer Center (DLCC)	Канада
Data and Program Library Services (DPLS)	США
Demographic Data Base	Швеция
Duke Data Bank of Documentary Papyri	США
Economic and Social Research Council Data Archive (ESRCDA)	Англия
Ethnic Minority Data Archive	Англия
Hebrew University Faculty of Social Sciences	Израиль
Instituto di Linguistica Computazionale	Италия
Inter-University Consortium for Political and Social Research (ICPSR)	США
Indian Social Science Research Counsil	Индия
Literary and Linguistic Computing Center	Англия
Medieval and Early Modern Data Bank	США
Nederlands Historisch Data Archief (NHDA)	Голландия
Norsk Tekstarkiv	Норвегия
Norwegian Social Science Data Service (NSD)	Норвегия
Oxford Text Archives	Англия
Public Archives of Canada	Канада
Roper Center of Public Opinion Research	США
School of Oriental and African Studies	Англия
Social Science Data Archive (SSDA)	Австралия
Social Science Data Library	США
Social Science Data Library	Канада
Steinmetzarchief (STAR)	Голландия
Swedish Social Science Data Service (SSDS)	Швеция
Social Research Informatics Society (TARKI)	Венгрия
Thesaurus Linguae Graecae	США
Wiener Institut fur Sozialwissenschaftliche Documentation und Methodik (WISDOM)	Австрия
Zentral Archiv fur empirische Sozialforschung (ZA)	ФРГ
Zentrum fur Historische Sozialforschung (ZHS)	ФРГ
Банк данных по истории России	Россия

по базам данных в отечественной истории, который в 1992 г. преобразован в Банк машиночитаемых данных по истории России. И.М.Гарскова в своей монографии приводит описания некоторых коллекций МЧД, уже заявленных разработчиками[19]. Среди них, например, описываются реляционная база данных по аграрной истории России первой половины XVII в., созданная по материалам писцовых книг; просопографическая база данных по депутатам I Государственной Думы, составленная на основе справочных печатных изданий, посвященных депутатскому корпусу первой Думы, и др.

Информационные системы на больших ЭВМ создавались и обслуживались большим числом лиц. С внедрением в исследовательскую практику персональных компьютеров часто одно и то же лицо становится и разработчиком, и пользователем, и администратором, и программистом, а сам банк данных состоит лишь из двух компонент: БД и СУБД, т.е. из базы данных в соответствующей системе управления базой данных. Такие банки данных стали называться *персональными*.

Вопросы проектирования баз данных требуют отдельного рассмотрения, поэтому коснемся лишь сюжетов построения баз данных, связанных со спецификой разных исторических источников.

Определяющее значение для перевода источников в машиночитаемую форму имеет уровень их структурированности, в соответствии с этим источники можно разделить на статистические, структурированные, текстовые (нарративные) и графические.

Статистические источники представляют собой таблицы статистических показателей (количественных данных), собранных по всем объектам некоторой совокупности (хозяйствам, губерниям, отраслям промышленности, группам населения и т.п.). Важными свойствами статистических источников являются массовый характер первичных сведений и агрегирование первичной информации. Статистические данные обычно являются либо первичными, либо агрегированными. Структура организации данных на основе первичных данных статистических источников (на микроуровне) представляет собой обычную таблицу «объекты – признаки». Структура на макроуровне (на основе агрегированных данных) – это сложные многомерные группировки по иерархическому принципу или принципу таблиц сопряженности на основе некоторых критериев (тематических, пространственных или хронологических).

Формулярные источники, совсем недавно получившие название структурированных (highly structured historical sources), изначально имеют четкую структуру (формуляр), что делает их наиболее удобными для перевода в машиночитаемый вид. К структурированным источникам относятся ма-

[19] См.: Гарскова И.М. Указ. соч. Приложение 1. С.132-185.

териалы переписей, книг церковной или гражданской регистрации рождения, крещения, брака и смерти, личные дела и личные карточки, анкеты, справочники. Основными особенностями этих источников являются отсутствие агрегированной информации и соединение разнотипной информации (текстовой, числовой, логической) в одном формуляре. Формуляр источника часто представляет собой практически готовую структуру базы данных (надо только описать атрибуты объектов).

Текстовые (нарративные) источники являются наиболее трудными для формализации и перевода в машиночитаемую форму. Основная особенность этих источников - отражение в них структуры естественного языка. Хотя в тексте может присутствовать и формальная структура (разделы, параграфы, абзацы и т.п.), степень формализации текстовых источников невысока. Текст можно хранить в полном виде как линейную последовательность символов или в формализованном виде (с некоторой потерей информации), в последнем случае необходимо внести в текст специальные коды, поместить в нем нужные смысловые единицы.

Наконец, в последнее время создаются базы данных, содержащие, наряду с описательной, графическую информацию. Графическую информацию в исторических исследованиях представляют изобразительные источники, фотодокументы, географические карты и др. Однако и обычные тексты (особенно это касается средневековых текстов) могут быть представлены в виде графических изображений, если их вводить с помощью устройства оптического ввода – *сканера*.

Итак, при построении баз данных необходимо учитывать особенности структуры исторических источников, на основе которых они создаются. При этом исследователь, имеющий дело с менее структурированным источником, может не только вводить в память компьютера полный его текст, но и формировать некоторые структуры, внешние по отношению к тексту, которые позволяют извлекать из этого текста новую информацию в соответствии с задачами исследования.

Таким образом, современные компьютерные технологии создания баз и банков машиночитаемых данных открывают новые перспективы для исторических исследований, не только расширяя круг источников (как первичных, так и производных, ранее не существовавших), но и совершенствуя методический инструментарий историка.

Рассмотрим теперь, как с помощью количественных методов решаются задачи классического источниковедения.

3.2. Изучение происхождения источника

Многие древние памятники дошли до нас в десятках списков и редакций, поэтому их источниковедческий анализ предполагает прежде всего установление взаимоотношений редакций и списков, выявление генетической связи всех сохранившихся и утраченных текстов памятника и воссоздание истории текстов. Эти задачи решаются путем довольно сложного сравнительно-текстологического анализа, облегчить который можно с помощью компьютерного построения классификации списков.

Рассмотрим, как применяются количественные методы и компьютер в изучении происхождения нарративных источников на ставшем классическим примере построения «генеалогического древа» (стеммы) древнейшего юридического памятника славянского права IX века – «Закона Судного Людем»[20].

В основе построения классификации лежит *метод «групп»*, предложенный французским текстологом Д.Ж. Фроже. Главная идея метода заключается в следующем: если списки-«потомки» приобретают все особенности списков-«предков», то история копирования списков вполне определенным образом зашифрована в разночтениях списков. Тогда на основе анализа структуры разночтений можно построить генеалогическое древо списков.

Метод «групп» имеет довольно жесткие условия:

1) у каждого списка имеется только один протограф;

2) в каждом списке содержатся все ошибки его протографа;

3) одинаковые ошибки не содержатся в списках, имеющих в качестве своих протографов независимые списки.

Логическая схема метода «групп» легко формализуема с помощью языка теории множеств и теории графов. Однако модель Фроже упрощает реальный процесс копирования списков, что значительно сужает круг источников, к которым данный метод можно применить.

В качестве предмета исследования Л.В.Милов и Л.И.Бородкин выбрали один из древнейших памятников славянской юридической мысли «Закон Судный Людем» (ЗСЛ), исходя из того, что характер этого произведения (свод законов) налагает жесткие ограничения на процесс копирования, приближая его к модельному. ЗСЛ – раннехристианский юридический памятник, созданный в 60-х годах IX в. одним из славянских просветителей Кириллом-Константином в пределах Велико-Моравского княжества. Позже ЗСЛ нашел практическое применение в Болгарии конца IX – начала X

[20] См.: Бородкин Л.И., Милов Л.В. О некоторых аспектах автоматизации текстологического исследования (Закон Судный Людем) // Математические методы в историко-экономических и историко-культурных исследованиях. М., 1977.

века. Однако тексты этого памятника сохранились только на Руси в составе древнерусских юридических сборников XIII – XVII вв. Для анализа использовалось академическое издание краткой редакции ЗСЛ, содержащее 54 списка 4-х изводов.

Поскольку применение метода «групп» требует сличения всех списков с некоторым исходным экземпляром – «экземпляром ссылок», то в качестве исходного был взят наиболее древний датированный список – список ЗСЛ из Новгородской кормчей 1280 г. Все разночтения текста, полученные при сличении всех списков с «экземпляром ссылок», были закодированы и составили более 15 тысяч вариантов разночтений. Этот материал и послужил исходной информацией для реализации метода «групп».

В процессе компьютерной обработки информации выявились некоторые противоречия между реальной структурой вариантов разночтений и требованиями модели, которые были ликвидированы в результате экспертной оценки специалиста-историка. В целом анализ характера противоречий позволил сделать вывод о том, что реальный процесс копирования списков ЗСЛ можно описать моделью метода «групп».

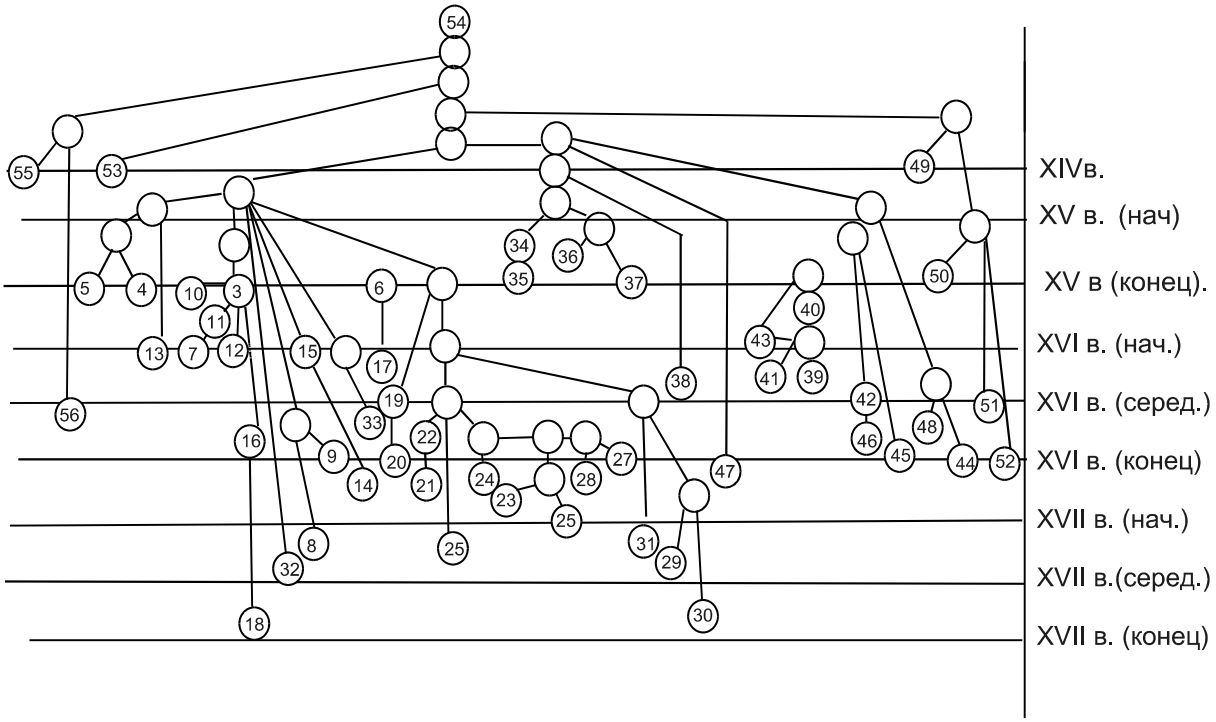

Рис. 6. Стемма «Закона Судного Людем»

Итогом работы стало построение генеалогического древа – стеммы (рис. 6), отражающей историю текста ЗСЛ и дающей новую, принципиально важную информацию. Каждый из четырех изводов выделился на стемме в отдельное «прадерево», т.е. все списки каждого извода имеют одного

общего «предка», при этом изводы не пересекаются. Интересным результатом, подтверждающим корректность проведенной классификации, стало то, что построенная стемма не содержит хронологических противоречий: все сохранившиеся списки выстроились в цепях генеалогического древа точно по фактической хронологии, хотя сведения о дате списков в компьютер не вводились (названия нумерованных списков памятника даны в таблице 4). Кроме того, компьютер «реконструировал» большое число несохранившихся списков (на стемме они обозначены пустыми кружочками). По отношению к известным 54 спискам они составляют примерно 60% (31 реконструкция). При этом каждая реконструкция поставлена на определенное место в генеалогии списков, что позволяет судить о том, что было, казалось, навсегда утрачено.

ЗСЛ краткой, т.е. древнейшей, редакции сохранился в большинстве списков в одних и тех же юридических сборниках с Русской Правдой – «Кормчих книгах». Из всех списков только в Устюжской Кормчей, Иосифовской Кормчей и Варсонофьевской Кормчей нет текста Русской Правды. При этом соседство в книгах является не просто механическим. Как доказал М.Н.Тихомиров, тексты каждой из кормчих книг носят следы единой редакции, поэтому можно предположить, что генеалогии списков Русской Правды и ЗСЛ будут совпадать.

Историками проведено фундаментальное текстологическое исследование ста с лишним списков Русской Правды Пространной редакции. Вся совокупность списков типологически была разделена на три вида:

1) Синодально-Троицкий вид, списки которого сохранились в составе Кормчих книг и Мерил Праведных;

2) Пушкинско-Археографический вид, списки которого сохранились в юридических сборниках особого состава;

3) Карамзинский вид, списки которого сохранились в позднейших сборниках XV-XVI вв.

В свою очередь виды подразделяются на изводы. Так, Синодально-Троицкий вид делится на Чудовский извод (или Розенкампфовский), Софийский извод (или Новгородско-Софийский), Ферапонтовский извод, извод Мерила Праведного и т.д.

Таким образом, если классификация списков Русской Правды Синодально-Троицкого вида, проведенная историками-текстологами традиционными методами, совпадает со стеммой, созданной исследователями с помощью компьютера на основе метода «групп», тогда можно говорить о корректности и эффективности компьютерной классификации.

Таблица 4

Перечень списков «Закона Судного Людем»

Номер списка	Наименование списка	Датировка
	Чудовский извод	
3	Чудовский	1499 г.
4	Розенкампфовский	конец XV в.
5	Троицкий V	конец XVI в.
6	Список Публичной библиотеки им. Салтыкова-Щедрина	конец XV в.
7	Соловецкий II	начало XVI в.
8	Троицкий II	начало XVI в.
9	Царского II список	вторая пол. XVI в.
10	Крестининский	конец XV в.
11	Овчинниковский I	конец XV в.
12	Академический II	начало XVI в.
13	Антониево-Сойский	начало XVI в.
14	Ионовский	конец XVI в.
15	Троицкий III	начало XVI в.
16	Музейский I	вторая пол. XVI в.
17	Возмицкий	1533 г.
18	Забелинский	конец XVII в.
19	Ферапонтовский	середина XVI в.
20	Толстовский II	вторая пол. XVI в.
21	Фроловский I	конец XVI в.
22	Соловецкий IV	вторая пол. XVI в.
23	Егоровский IV	конец XVI в.
24	Архивный II	конец XVI в.
25	Румянцевский II (Никоновский)	1620 г.
26	Рогожский II	начало XVII в.
27	Кирилло-Белозерский I	вторая пол. XVI в.
28	Кирилло-Белозерский II	1590 г.
29	Погодинский II	первая пол. XVII в.
30	Погодинский III	вторая пол. XVII в.
31	Царского III список	первая пол. XVII в.
32	Хлудовский	первая пол. XVII в.
33	Никифоровский	вторая пол. XVI в.

Окончание табл. 4

Номер списка	Наименование списка	Датировка
	Софийский извод	
34	Софийский	1470-1490 гг.
35	Румянцевский	конец XV в.
36	Вязниковский	вторая пол. XV в.
37	Ярославский	конец XV в.
38	Прилуцкий	1534 г.
39	Соловецкий III	1519 г.
40	Архивный I	кон. XV –нач. XVI в.
41	Егоровский I	начало XVI в.
42	Царского I список	середина XVI в.
43	Овчинниковский II	1518 г.
44	Хворостининский	конец XVI в.
45	Толстовский III	конец XVI в.
46	Егоровский II	вторая пол. XVI в.
47	Годуновский II	конец XVI в.
48	Фроловский-Браиловский	вторая пол. XV в.
	Извод Мерила Праведного	
49	Троицкий	XIV в.
50	Синодальный II	1467-1481 гг.
51	Кирилло-Белозерский	середина XVI в.
52	Синодальный III	1587 г.
	Древнейший извод	
53	Варсонофьевский	XIV в.
54	Новгородский	1280 г.
55	Устюжский	XIV в.
56	Иоасафовский	XVI в.

Сопоставление стеммы с классификацией списков Пространной Русской Правды, данной в академическом издании текстов Русской Правды, показало их полное совпадение. Основные группы ветвей генеалогического древа по составу списков оказались идентичны группам, составленным учеными «вручную». Кроме того, Л.И.Бородкину и Л.В.Милову удалось

решить ряд спорных вопросов[21]. Так, один из изводов Русской Правды В.П.Любимов назвал Розенкампфовским по Розенкампфовскому списку, а М.Н.Тихомиров считал, что наиболее важным для всего извода является Чудовский список. Стемма показала, что в основании извода лежит именно Чудовский список и извод должен называться Чудовским. Стемма внесла существенные коррективы в оценку целой ветви списков ЗСЛ (и Русской Правды) – Ферапонтовского извода. В.П.Любимов полагал, что этот извод очень позднего происхождения и является скорее свидетельством того, что Русская Правда была уже в XVI-XVII вв. литературным памятником, а не правовым кодексом. Итоги компьютерной классификации показали весьма древнее происхождение Ферапонтовского извода, сформировавшегося ранее Чудовского извода и являвшегося действующим источником права. Вместе с тем стало ясно, что до наших дней дошли лишь «обломки» извода, его самые позднейшие списки, а его костяк составляли несохранившиеся, но высчитанные компьютером протографы.

Как уже отмечалось, в качестве «экземпляра ссылок» исследователями был выбран древнейший список Новгородской кормчей 1280 г. Однако после построения стеммы возникает задача более правильного выбора исходного протографа – переориентации «корня» генеалогического древа на основе экспертной работы историка, при этом соотношения списков в стемме меняться не будут.

После содержательного анализа в качестве исходного протографа был выбран несохранившийся список, от которого ведут начало два списка (Устюжский и Иоасафовский), находящиеся в кормчих книгах Устюжского типа. При этом Устюжская кормчая, кроме ЗСЛ, содержит текст «Номоканона 50 титулов» византийского юриста Иоанна Схоластика в древнейшем славянском переводе, предпринятом, по мнению многих ученых, Мефодием в период его жизни в Моравии в 60-80-х годах IX в. Соседство этих двух памятников IX в. позволило исследователям предположить, что Устюжская кормчая ведет к древнейшему сборнику Кирилло-Мефодиевской эпохи, откуда и проник на Русь Закон Судный Людем.

Таким образом, построенное генеалогическое древо списков ЗСЛ имеет большое значение для исследования происхождения не только текста ЗСЛ, но и многих текстов, входящих в состав Кормчих книг и Мерил Праведных. Прежде всего, компьютерная классификация дает принципиально новые возможности для изучения истории происхождения Пространной Русской Правды.

---

[21] См.: Бородкин Л.И., Милов Л.В. Некоторые аспекты применения количественных методов и ЭВМ в изучении нарративных источников // Количественные методы в советской и американской историографии: Материалы советско-американских симпозиумов. Балтимор, 1979. Таллинн, 1981. М., 1983. С.386.

3.3. Атрибуция источника

Наряду с изучением происхождения нарративных источников важной источниковедческой задачей является их атрибуция. Установление авторства нарративных текстов является одной из самых сложных проблем в отечественном источниковедении. Связано это с тем, что до XVII в. включительно в большинстве своем литературные и исторические произведения были анонимны. Открытое объявление имени считалось нескромным и даже греховным. При этом анонимность выражается не только и не столько в отсутствии авторской подписи, сколько в слабости авторского начала, что отражает специфику средневекового литературного произведения. Во-первых, органическую часть средневекового памятника составляют многочисленные цитирования текстов религиозно-философского, литературно-религиозного и церковно-канонического содержания, лишающие текст его индивидуальных черт. Во-вторых, в силу существования так называемого «литературного этикета» литературные произведения разных жанров написаны согласно канонизированным жанровым и стилистическим приемам по определенному формуляру, т.е. личное начало поглощено традиционными книжными приемами. Наконец, средневековый авторский текст необычайно подвижен и изменчив, поскольку целые поколения русских книжников: авторов, редакторов и писцов традиционно вмешивались в текст, становясь практически его соавторами. Все это приводит к значительным трудностям атрибуции нарративных текстов традиционными методами.

Сложность определения авторского стиля средневекового памятника объясняет тот факт, что многие десятилетия вопросы атрибуции того или иного произведения являются в историографии либо спорными, либо не до конца решенными. К числу дискутирумых проблем относятся споры об авторе «Повести временных лет», о подлинности «Слова о полку Игореве», о принадлежности публицистических сочинений Ивана Пересветова перу Ивана Грозного и др.

Новые перспективы в решении спорных проблем атрибуции нарративных источников открылись с внедрением в исследовательскую практику количественных методов. Более двух десятилетий группа исследователей под руководством члена-корреспондента РАН, профессора Л.В.Милова ведет работу по атрибуции повествовательных (прежде всего древнерусских) текстов X-XVIII вв. на основе новой методики формализации авторского стиля. Итоги многолетних исследований нашли отражение в фундаментальной коллективной монографии[22].

[22] См.: От Нестора до Фонвизина. Новые методы определения авторства / Под ред. чл.-кор. РАН Л.В.Милова. М., 1994.

Традиционно выявление авторских особенностей осуществляется путем выявления деталей стиля, присущих тому или иному автору (излюбленных слов, оборотов и выражений), которые являются субъективно-осознанными или субъективно признанными существенными для стиля автора. Однако такая методика часто оказывается малоэффективной, поскольку лексика автора может диктоваться жанром или автор может подражать авторитету. Более полезным для атрибуции оказывается учет подсознательных особенностей стиля автора. Подсознательные особенности письменного языка какого-либо автора выражаются в специфике употребления и чередования им различных грамматических форм (существительных, прилагательных, глаголов и т.п.) безотносительно к их лексическому содержанию. Предложенная исследователями методика заключается в формализации конкретного авторского текста путем перевода его в систему грамматических форм (классов) и применении к формализованному таким образом тексту метода анализа парных встречаемостей грамматических классов слов. Система грамматических классов слов включает 150 наименований (кодов) модификаций частей речи русского языка, которая применяется в двух вариантах: а) с учетом всех грамматических классов слов, т.е. всех 150 кодов – первая система кодирования; б) с учетом лишь знаменательных слов, т.е. без союзов, предлогов и частиц – вторая система кодирования, включающая 135 кодов.

Суть *метода анализа парных встречаемостей грамматических классов слов* заключается в следующем.

1. Выявляются частоты парных встречаемостей тех или иных грамматических классов (форм) в авторском тексте из 1000 значимых слов (размер выборки в 1000 слов установлен экспериментально). При этом частоты парных встречаемостей у различных авторов будут различаться. Исключение составляют откровенные подражания и «плагиаты», но и в этих случаях обнаруживается минимум индивидуальных особенностей.

2. На основе лишь связи «слева - направо», т.е. в направлении развертывания текста, строится матрица частот парной встречаемости (статистических связей) и задается порог частоты встречаемости.

3. Полученная система статистических связей грамматических классов формализуется на языке теории графов построением графа сильных связей. Граф состоит из вершин и дуг, где вершина – это грамматический класс, а дуга – сильная связь (не ниже заданного порога). Анализ графа дает представление о стилевых особенностях автора.

Рассмотрим пример построения графа сильных связей по условной матрице парных встречаемостей (рис. 7)[23]. На матрице отражены частоты встречаемостей в тексте шести грамматических классов (1 – существи-

[23] См.: От Нестора до Фонвизина... С.344.

тельное, 2 – глагол, 3 – наречие, 4 – прилагательное, 5 – причастие, 6 – предлог). Возьмем порог встречаемости не ниже 6. На матрице видно, что частота 6 и больше встречается у пар в первой строке – 1:4 (8) и 1:5 (12); во второй строке – 2:6 (7); в третьей строке – 3:1 (8) и 3:4 (9); в четвертой строке – 4:4 (6) и 4:6 (6); в пятой строке - сильных связей нет; в шестой строке – 6:2 (6) и 6:3 (7). В соответствии с этим строятся вершины 1-6 классов, а дуги обозначают направления сильных связей. В построенном графе вершины несут разную «нагрузку» по числу дуг. Вершины с числом дуг не меньше трех называются узлами.

	1	2	3	4	5	6
1	5	4	3	8	12	1
2	3	2	1	2	3	7
3	8	1	3	9	2	3
4	3	4	5	6	4	6
5	1	3	2	3	1	4
6	5	6	7	2	1	3

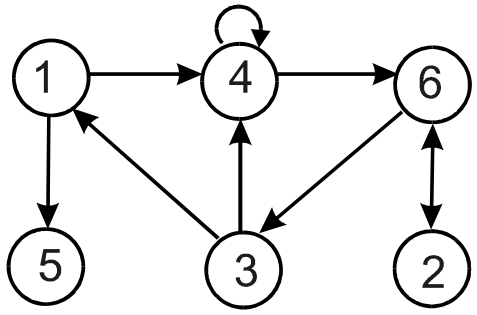

Рис. 7. Условные матрица и граф

Обратимся теперь к конкретным результатам атрибуции древних текстов на основе метода анализа парных встречаемостей грамматических классов слов.

Одной из наиболее интересных источниковедческих проблем, обсуждаемых в литературе уже два столетия, является проблема авторства «Повести временных лет» (ПВЛ). Хотя в историографии преобладала точка зрения о создании этого памятника монахом Киево-Печерского монастыря Нестором, против нее всегда выдвигались аргументы, основанные на анализе значительных расхождений между летописью и житийными произведениями, принадлежащими перу Нестора. Сопоставление текстов житийных произведений с некоторыми летописными статьями «Повести временных лет» (рассказами о Киево-Печерском монастыре и о гибели Бориса и

Глеба, помещенными под 1051, 1074 гг. и 1015 г.) на основе новой методики позволило Л.В.Милову сделать ряд принципиально важных выводов[24].

Графы текстов бесспорно несторовских произведений («Жития Феодосия» и «Чтения о Борисе и Глебе») оказались достаточно схожи по структуре не только друг с другом, но и с графами Печерской повести 1051, 1074 гг. и летописной статьи 1015 г. Сильное сходство графов «Чтения о Борисе и Глебе» и «Жития Феодосия», с одной стороны, и Печерской повести, с другой - подтверждают и построенные исследователем общие графы (рис. 8-9).

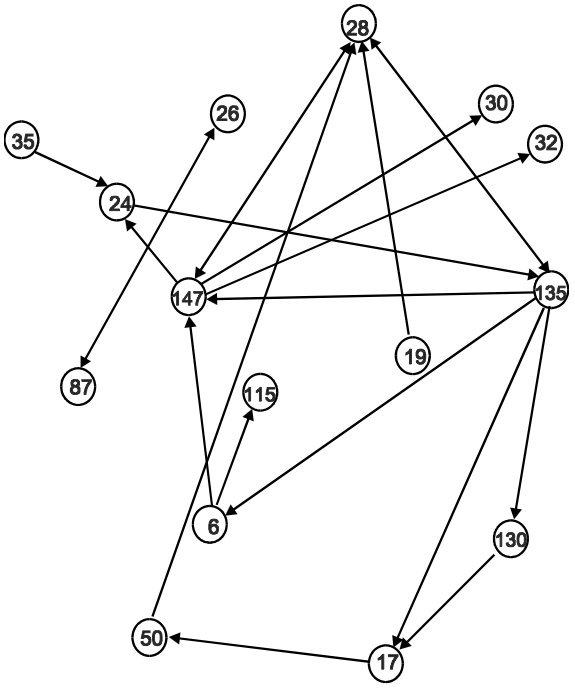

Рис. 8. Общий граф «Чтения о Борисе и Глебе» и Печерской повести

Общий граф «Чтения о Борисе и Глебе» и Печерской повести (рис. 8) имеет 21 связь, а коэффициент близости очень высок (0,38), что свойственно текстам одного автора. Общий граф «Жития Феодосия» и Печерской повести (рис. 9) также имеет 21 связь. Коэффициент близости равен 0,32, что тоже свидетельствует о том, что автором обоих текстов являлось одно лицо.

В качестве контрастного примера Л.В.Милов приводит общий граф Печерской повести и «Сказания о Борисе и Глебе» - произведения той же эпохи и того же жанра, но бесспорно признанного другого авторства

[24] См.: От Нестора до Фонвизина... Очерк I. Кто был автором «Повести временных лет»? С.40-69.

(рис. 10). В нем всего 14 связей и соответственно более низкий коэффициент близости (0,21).

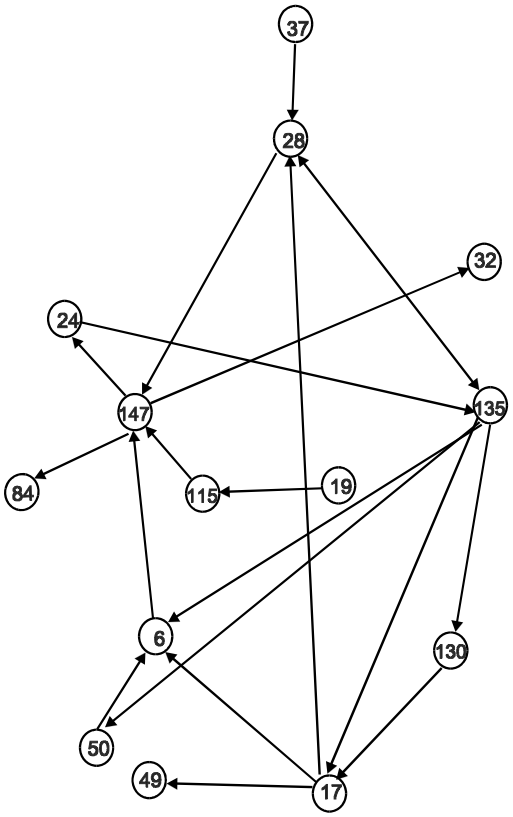

Рис. 9. Общий граф «Жития Феодосия» и Печерской повести

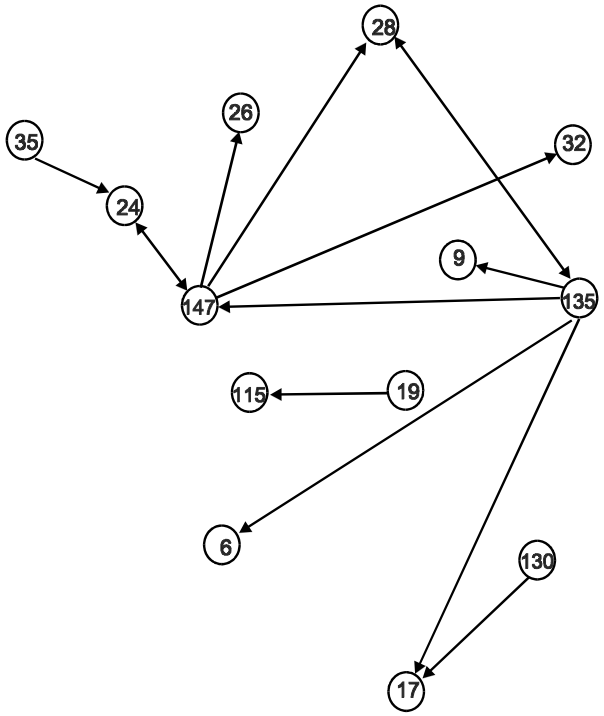

Рис. 10. Общий граф Печерской повести и «Сказания о Борисе и Глебе»

Аналогичным образом автор сравнивает несторовские житийные произведения с графом летописной статьи 1015 г. Анализ текстов с помощью применения новой методики позволяет исследователю отвергнуть господствующее в литературе мнение о том, что летописная статья 1015 г. написана вероятным создателем Начального свода на основе «Сказания о Борисе и Глебе». Л.В.Милов полагает, что статья была создана Нестором, а затем поздний редактор ПВЛ (вероятно, Сильвестр) внес фактические исправления в текст, опираясь уже на данные «Сказания о Борисе и Глебе».

Таким образом, подводя итоги изучения текстов древнейшего русского летописного свода, исследователь призывает прекратить более чем вековые споры об авторе «Повести временных лет»: «Создателем этого летописного свода, как выясняется, был все-таки знаменитый печерский черноризец Нестор»[25].

Метод анализа парных встречаемостей грамматических классов слов позволил историкам-клиометристам достаточно эффективно решить сложные проблемы атрибуции ряда других древних памятников:

- Убедительно доказано, что внутренняя структура знаменитого шедевра древнерусской словесности – «Слова о полку Игореве» - является структурой языка XII столетия, что исключает возможность позднейшей подделки.

- Опровергнуто распространенное мнение о составлении памятника XVI века Степенной книги митрополитом Афанасием и выявлен автор целой серии очерков в составе Степенной книги – писатель Василий-Варлаам, сотрудничавший в литературной дружине митрополита Макария при создании Великих Четьи-Миней.

- При обнаружении некоторого стилевого сходства сочинений Ивана Грозного, Андрея Курбского и Ивана Пересветова, объясняемого активным вмешательством переписчиков и канцеляристов в авторский текст, отвергнуты гипотезы, согласно которым Иван Пересветов – это псевдоним Ивана Грозного, а переписка Грозного и Курбского создана князем Шаховским в 20-30-х гг. XVII века.

- Определены авторы двух анонимных произведений конца XVIII в. – «Писем к Фалалею» (Денис Фонвизин) и «Деревенского зеркала или общенародной книги» (Андрей Болотов).

Перспективы развития данного направления квантитативной истории исследователи связывают с созданием банка данных всех атрибутированных и анонимных нарративных текстов. Тогда «история древнерусской и русской культуры обогатится огромным количеством новых действующих

[25] От Нестора до Фонвизина... С. 339.

лиц, а круг известных историографии авторов обогатится произведениями, созданными ими, но по тем или иным причинам ставшими анонимными»[26].

3.4. Определение достоверности и репрезентативности источника

До сих пор рассматривалось, как с помощью количественных методов решаются задачи источниковедческой критики нарративных источников. Обратимся теперь к массовым источникам.

Основной целью источниковедческого анализа массовых источников является установление достоверности и репрезентативности зафиксированных в них количественных данных. Как было показано в первом разделе настоящего пособия, решая вопрос о достоверности конкретно-исторических данных, исследователь должен изучить историю происхождения и судьбу источников. Однако, наряду с содержательным анализом, для определения достоверности можно использовать некоторые математико-статистические методы.

Часто в распоряжении историка имеется несколько источников, содержащих данные по изучаемой им проблеме, и необходимо решить вопрос об их сравнительной достоверности. Для решения этого вопроса полезно сравнить статистические характеристики (средние и меры вариации) одних и тех же признаков, полученные по данным разных источников. Но сравнения описательных характеристик бывает недостаточно для суждения о действительном сходстве данных, содержащихся в разных источниках. Для определения степени их сходства используется корреляционный анализ, позволяющий выявлять тесноту взаимосвязей между признаками.

При этом может возникнуть несколько ситуаций:

1. Данные источника, достоверность которого неизвестна, сравниваются с данными достоверного источника. Тогда высокие коэффициенты корреляции между соответствующими показателями источников будут свидетельствовать о достоверности исследуемого источника, а низкие – о его недостоверности.

2. Сравниваются несколько источников неизвестной достоверности. Если корреляционная взаимосвязь данных, извлеченных из этих источников, достаточно высока, то источники рисуют одинаковую картину и, следовательно, отличаются высокой достоверностью. Если же сопряженность показателей источников отсутствует, то либо все эти источники недостоверны, либо один из них достоверен, а другой (другие) – нет. Решить проблему достоверности в этом случае можно путем включения данных того или иного источника в совокупность других показателей, характеризующих систему, в которую входит и изучаемое явление. Если рассматривае-

[26] От Нестора до Фонвизина... С. 342.

мые данные вписываются в систему других показателей (сопряжены с ними), то они являются достоверными. В противном случае они недостоверны.

Рассмотрим варианты применения корреляционного анализа для определения достоверности источников на конкретных примерах.

Б.Н.Миронов сравнивает данные об урожаях ржи по сведениям губернаторских отчетов, достоверность которых признается некоторыми исследователями сомнительной, с признанными достоверными данными частных хозяйств за 1841-1850 гг. (табл. 5)[27].

Таблица 5

Урожаи ржи в Европейской России по губернаторским отчетам (I) и по записям частных хозяйств (II) в 1841-1850 гг. (в «самах»)

Сведения	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850
I	3,4	4,0	4,5	3,9	3,5	3,1	3,3	2,4	3,9	3,2
II	6,1	8,3	8,8	7,6	7,4	6,2	6,3	5,1	7,3	6,4

По приведенным данным можно вычислить средние арифметические, вариации признака и коэффициент корреляции между рядами урожаев по двум источникам. Получим: $\bar{x}_{\text{I}}=3{,}52$; $\bar{x}_{\text{II}}=6{,}95$; $V_{\text{I}}=15{,}8\%$; $V_{\text{II}}=15{,}4\%$; $r=0{,}953$.

Очевидно, что уровень урожаев источники отражают по-разному, а динамику урожаев – практически одинаково. Высокий коэффициент корреляции позволяет исследователю сделать вывод о достоверности сведений губернаторских отчетов в отношении синхронности и пропорциональности колебаний урожаев. При этом различия в уровне урожаев, зафиксированные в источниках, он объясняет тем, что данные губернаторов охватывали всю крестьянскую и помещичью пашню, а записи частных хозяйств – пашню отдельных помещиков.

Конечно, в вопросе о достоверности губернаторских отчетов еще рано ставить точку (слишком уж существенны различия в данных об урожайности этих двух источников). Однако результаты проведенного анализа свидетельствуют о том, что не следует пренебрегать сведениями губернаторских отчетов, считая их полностью недостоверными.

Анализ достоверности урожайной статистики по данным трех независимых источников (Центрального статистического комитета (ЦСК), Министерства земледелия и земств) с помощью методов математической стати-

[27] См.: Миронов Б.Н. Указ. соч. С.61-62.

стики проводился Д.Н. Иванцовым еще в начале XX века[28]. Общие итоги корреляции данных показали, что эти разные источники рисуют очень сходную картину динамики урожайности.

Так, корреляционная взаимосвязь динамики урожайности ржи по 50 губерниям Европейской России в 1885-1908 гг., по данным ЦСК и Министерства земледелия, равнялась у крестьян и у частных владельцев 0,92. При этом по отдельным губерниям коэффициенты корреляции превышали 0,90 у помещиков в 32 и у крестьян в 36 губерниях из 50, а были менее 0,75 соответственно в 4 и 1 губерниях. Что касается сведений ЦСК и земств, то в среднем по 18 губерниям эти данные дают коэффициенты корреляции 0,92 у крестьян и 0,89 у помещиков. По отдельным губерниям взаимосвязь погодных средних урожаев значительно превышает 0,90 (табл. 6).

Таким образом, динамику урожайности данные разных источников урожайной статистики отражают одинаково, что свидетельствует о достаточно высокой ее достоверности в рассмотренном аспекте.

Таблица 6

Взаимосвязь погодных средних урожаев по сведениям ЦСК и земств

Губерния	Годы	Коэффициент корреляции
Воронежская	1886-1908	1,00
Вятская	1892-1905	0,97
Московская	1885-1908	0,92
Нижегородская	1892-1901	0,98
Орловская	1896-1903	0,98
Полтавская	1886-1910	0,97
Саратовская	1899-1906	0,93
Херсонская	1887-1907	0,98
Ярославская	1903-1909	0,99

Проблему достоверности сведений разных источников, сопряженность которых отсутствует, изучал И.Д.Ковальченко по данным переписи 1897 г. и Комиссии 16 ноября 1901 г. о сельскохозяйственных наемных рабочих[29]. Для этого по данным двух источников им были вычислены коэффициенты корреляции между долей наемных сельскохозяйственных рабочих (наемные рабочие в процентах к общему числу работников) и другими показателями социально-экономического развития по 50 губерниям Европейской России (табл. 7).

[28] См.: Массовые источники по социально-экономической истории России периода капитализма. С.252-256.

[29] См.: Массовые источники... С.269-270.

Таблица 7

Корреляционная взаимосвязь обеспеченности сельскохозяйственными рабочими с другими факторами социально-экономического развития

Факторы	Перепись 1897 г.	Данные 1901 г.
Хозяйства с наймом (в % к общему числу)	0,75	-0,01
Грамотные (%)	0,81	-0,09
Посевы (дес. на душу сельского населения)	-0,01	0,57
Продуктивный скот (на душу населения)	0,33	0,20
Урожайность зерновых (пудов с дес.)	0,28	-0,31

Данные переписи 1897 г. о наемных рабочих тесно взаимосвязаны с долей хозяйств, применявших наемный труд (0,75), т.е. с признаком, отличавшимся высокой достоверностью. Взаимосвязь с размерами посевов отсутствует (-0,01), т.е. относительные размеры земледельческого производства не определяли степени применения наемного труда. Наблюдается слабая связь с обеспеченностью продуктивным скотом и урожайностью и тесная взаимосвязь с грамотностью населения (0,81), которая отражает общий уровень буржуазно-капиталистического развития. Все это свидетельствует о достоверности данных переписи 1897 г. о наемных рабочих как показателе сравнительного уровня применения наемного труда в сельском хозяйстве отдельных губерний.

Иная картина с данными Комиссии 16 ноября 1901 г. Здесь показатели применения наемного труда имеют связь лишь с размерами посевов (0,57), что объясняется тем, что данные 1901 г. о наемных рабочих исчислялись исходя из учета их потребности в земледелии. С долей хозяйств, применявших наемный труд, и степенью грамотности, т.е. с ведущими факторами, сведения 1901 г. совсем не связаны. Следовательно, сведения 1901 г. не отражают сравнительной степени применения наемного труда в сельском хозяйстве отдельных губерний.

Таким образом, достоверными данными об обеспеченности сельскохозяйственными наемными рабочими располагает перепись 1897 г., а сведения Комиссии 16 ноября 1901 г. в этом отношении недостоверны.

Другой важной задачей источниковедческого анализа массовых источников является определение репрезентативности (представительности) содержащихся в них конкретно-исторических данных. Как уже отмечалось в первом разделе работы, исследователь должен установить качественную и количественную репрезентативность данных. Качественная репрезентативность (достаточность данных для раскрытия внутренней сути изучаемого явления или процесса) определяется на основе их содержательного анализа. Проблема репрезентативности данных в количественном отношении решается с помощью выборочного метода математической статистики.

Применение выборочного метода оказывается наиболее эффективным, когда в распоряжении историка имеются большие объемы массовых источников, сплошная обработка которых весьма затруднительна, да и вряд ли целесообразна, поскольку на основе репрезентативных выборок можно получить достаточно надежные результаты.

Примером успешного использования выборочного метода для формирования репрезентативных данных является исследование В.З.Дробижевым, А.К.Соколовым и В.А.Устиновым социальной структуры рабочего класса по материалам профессиональной переписи 1918 года[30].

Профессиональная перепись рабочих и фабрично-заводских служащих России 1918 г. охватила территорию 31 губернии, на которые в то время распространялась Советская власть. Перепись коснулась 6973 фабрик и заводов с 1246343 рабочими и служащими, она явилась одной из самых массовых по охвату фабрично-заводского персонала и более полной по объему полученных сведений, чем многие последующие обследования рабочего класса. Первичный бланк переписи включал 37 важных с точки зрения социального анализа вопросов: национальность, место рождения, возраст, возраст первоначального поступления на предприятие, должность, профессия, стаж в должности и профессии, потомственность, уровень квалификации, наличие земли в деревне и характер связи с сельским хозяйством и т.д. Первичные материалы переписи составляют более миллиона личных карточек рабочих. Понятно, что такой огромный массив данных можно изучать лишь на основе выборочного метода.

При определении путей выборочной обработки переписи исследователи провели выборочный эксперимент на материалах Ярославской и Воронежской губерний, поскольку Ярославская принадлежала к промышленно развитым губерниям и имела значительное число рабочих, а материалы Воронежской губернии дают представление о рабочих сельскохозяйственного района. При этом было решено взять несколько выборок. Так, из первичных материалов Ярославской губернии механическим способом были отобраны каждые десятая, двадцатая и сотая анкеты (10, 5 и 1 %-ные выборки). Исходя из анализа распределений признаков в выборках различных объемов, определялся оптимальный вариант, обеспечивающий репрезентативность анализируемых данных. Расчеты показали, что достаточно точное (с вероятностью 95 %) представление обо всех параметрах социальных слоев рабочего класса дает 5 %-ная выборка. Вместе с тем обнаружилось, что сравнительно небольшая часть рабочих – рабочие, занятые в сфере общественного управления на производстве, - в 5 %-ную выборку практически не попадала. Тогда доля отбора для этой категории была увеличена

[30] См.: Дробижев В.З., Соколов А.К., Устинов В.А. Рабочий класс Советской России в первый год диктатуры пролетариата. М., 1975.

(по некоторым губерниям обрабатывались сведения обо всех рабочих данной группы).

Таким образом, сочетание содержательного анализа с корректным проведением выборочного обследования позволило исследователям сформировать систему репрезентативных данных, в полной мере характеризующих социальную структуру рабочего класса Советской России в 1918 году.

Более сложным вопросом источниковедческого анализа массовых источников является установление репрезентативности «естественных выборок». В этом случае историк должен доказать, что сохранившиеся сведения носят случайный характер, поскольку случайность данных является главным условием их представительности. Здесь преимущественную роль играет историко-содержательный анализ. Однако дополнительную проверку случайности естественной выборки можно осуществить с помощью метода «критерия знаков».

Применение *метода «критерия знаков»* сводится к следующему:

Сохранившиеся данные по какому-либо признаку записываются в той последовательности, в какой они встречаются в источнике. Затем из каждого последующего значения вычисляется каждое предыдущее, соответствующая разность оказывается либо положительной (+), либо отрицательной (-). В итоге получается определенное число плюсов и минусов. Если различия между значениями случайны, т.е. если выборка случайна, то число плюсов (или минусов) не выходит за рамки критических границ, определенных в специальных таблицах для каждого объема выборки.

Метод «критерия знаков» использует, например, Б.Н.Миронов для определения репрезентативности данных о ценах четверти ржи за 1708 г. по 36 уездам[31]. Проведенный содержательный анализ позволяет рассматривать сохранившиеся данные о хлебных ценах за 1708 г. как случайную выборку (никакой преднамеренности в сборе сведений о ценах и сохранении их в архивах не было). Чтобы убедиться в этом, исследователь обращается к методу «критерия знаков» (табл. 8).

Как видно из таблицы 8, число плюсов равно 15, а число минусов – 18. Критические границы для выборки в 36 единиц составляют 12-24 плюсов (или минусов). Следовательно, поскольку полученные плюсы и минусы не выходят за пределы критических границ, выборку можно считать случайной.

Таков основной круг источниковедческих задач, решение которых с помощью количественных методов дает эффективные результаты.

[31] См.: Миронов Б.Н. Указ. соч. С.47.

Таблица 8

Проверка случайности выборки методом «критерия знаков»

Уезд	Цена (коп.)	Знак разности	Уезд	Цена (коп.)	Знак разности
1	40		19	30	-
2	43	+	20	29	-
3	40	-	21	45	+
4	79	+	22	40	-
5	74	-	23	42	+
6	40	-	24	40	-
7	55	+	25	36	-
8	42	-	26	50	+
9	42		27	30	-
10	50	+	28	24	-
11	40	-	29	25	+
12	43	+	30	40	+
13	43		31	32	-
14	35	-	32	30	-
15	40	+	33	20	-
16	30	-	34	30	+
17	36	+	35	25	-
18	50	+	36	32	+

ЗАКЛЮЧЕНИЕ

Современные тенденции развития науки и компьютерная революция привели к потребности в математизации и компьютеризации научного знания, что в исторической науке связано с необходимостью решения задач расширения источниковой базы исследований и повышения информативной отдачи источника. Расширение источниковой базы осуществляется путем вовлечения в научный оборот обширных комплексов массовых источников, на основе которых создаются базы и банки машиночитаемых данных. Применение к историческим источникам математических методов, реализующих системный подход, позволяет извлекать из них новую, скрытую информацию.

В вопросе применения количественных методов в исторических исследованиях принципиальное значение имеет соотношение количественного и качественного анализа. Количественный анализ не противостоит качественному, а является составной частью исследования, качественный анализ в котором обязателен и имеет преимущественное значение. При этом, поскольку всякому качеству присуще определенное количество, сфера применения количественных методов практически не ограничена. Проблема заключается лишь в выявлении метода, адекватно отражающего суть изучаемого явления или процесса, и корректном его применении.

Самое широкое применение в исследовательской практике историков получили методы математической статистики. Эти методы особенно эффективны при изучении массовых исторических источников. Они позволяют решать задачи статистического описания совокупности объектов (методы дескриптивной статистики), статистического оценивания параметров генеральной совокупности по выборочным данным (выборочный метод), статистического анализа взаимосвязей (методы корреляционного и регрессионного анализа), классификации объектов или признаков (методы кластерного и факторного анализа), сжатия информации (методы факторного анализа). Применяя методы математической статистики, историк получает информацию, которая не может быть выявлена описательными методами. Это позволяет строить модели изучаемых явлений и процессов, адекватно отражающие их внутреннюю суть, итогом анализа которых является приращение знания.

Значительных успехов квантитативная история достигла в источниковедении массовых и нарративных источников. Введение в практику исторических исследований новых компьютерных технологий, связанных с созданием баз и банков машиночитаемых данных, открывает новые возможности хранения и использования исторических источников, изменяет информационную среду, совершенствует методику анализа.

Применение количественных методов дает возможность более плодотворно решать ряд источниковедческих задач: выявлять происхождение и авторство нарративных памятников, устанавливать достоверность и репрезентативность массовых источников.

Таким образом, квантификация исторических исследований позволяет значительно углубить изучение многих явлений и процессов действительности. При этом главным условием успешного применения количественных методов является глубина содержательного анализа, что должно учитываться на всех этапах клиометрического исследования: от постановки исследовательской задачи до интерпретации полученных результатов.

Контрольно-проверочные вопросы

1. Чем обусловлен процесс математизации и компьютеризации научного знания?
2. Каковы уровни интегральных исследований?
3. В чем суть системного подхода к изучению явлений действительности?
4. Что такое массовые источники?
5. Что является альтернативой количественному анализу?
6. Что такое качественный анализ?
7. Каковы формы математизации научного знания?
8. Какие этапы включает клиометрическое исследование?
9. Что означает правильная постановка исследовательской задачи?
10. Как классифицируются ошибки измерения?
11. Что необходимо установить для выявления достоверности источника?
12. В чем заключается репрезентативность конкретно-исторических данных?
13. Что такое статистическая совокупность?
14. Какие типы исследовательских задач позволяют решать методы математической статистики?
15. Что такое дескриптивная статистика?
16. Какие характеристики измеряют среднее значение признака?
17. Какие показатели характеризуют меру вариации признака?
18. Чем отличается коэффициент вариации от среднего квадратического отклонения?
19. Что показывает графическое изображение вариационного ряда?
20. Что такое нормальное распределение?
21. Что называется генеральной совокупностью?
22. На чем основано применение выборочного метода?
23. Какие выборки являются репрезентативными?
24. Каковы виды выборочного изучения?
25. Что такое средняя и предельная ошибки выборки?
26. Как можно определить объем выборки?
27. Что измеряет коэффициент корреляции?
28. Что показывает коэффициент детерминации?
29. Как проверить значимость коэффициента корреляции?
30. Что описывает уравнение регрессии?
31. Что вычисляется с помощью метода наименьших квадратов?
32. Что показывает коэффициент регрессии?
33. Для чего используется кластерный анализ?
34. В чем суть кластерного анализа?

35. Почему факторный анализ называется методом сжатия информации?
36. Что такое факторные нагрузки и факторные веса?
37. Что такое историческая информатика?
38. Что такое база данных?
39. Какие задачи решают СУБД?
40. Каковы основные компоненты банка данных?
41. Чем определяется специфика исторических источников, используемых для создания базы данных?
42. Какие задачи классического источниковедения решаются с помощью количественных методов?
43. Каковы ограничения «метода групп» Фроже?
44. Чем обусловлена сложность атрибуции древних текстов?
45. В чем суть метода анализа парных встречаемостей грамматических классов слов?
46. Какие проблемы атрибуции древнерусских памятников удалось решить исследователям с помощью новой методики?
47. С помощью какого метода можно определить сравнительную достоверность источников?
48. Какой метод позволяет сформировать систему репрезентативных данных источника?
49. Как устанавливается репрезентативность «естественных выборок»?
50. В чем суть метода «критерия знаков»?

Список рекомендуемой литературы

Барг М.А. Принцип системности в историческом исследовании // История СССР. 1981. № 2.

Бокарев Ю.П. Социалистическая промышленность и мелкое крестьянское хозяйство в СССР в 20-е годы: источники, методы исследования, этапы взаимоотношений. М., 1989.

Бородкин Л.И. Многомерный статистический анализ в исторических исследованиях. М., 1986.

Буховец О.Г. Социальные конфликты и крестьянская ментальность в Российской империи начала XX века: новые материалы, методы, результаты. М., 1996.

Воронкова С.В. Российская промышленность начала XX века: источники и методы изучения. М., 1996.

Гарскова И.М. Базы и банки данных в исторических исследованиях. М., 1994.

Дробижев В.З., Соколов А.К., Устинов В.А. Рабочий класс Советской России в первый год диктатуры пролетариата (Опыт структурного анализа). М., 1975.

Историческая информатика / Под ред. Л.И.Бородкина, И.М.Гарсковой. М., 1996.

Кахк Ю.Ю. Нужна ли новая историческая наука? // Вопросы истории. 1969. № 3.

Кахк Ю.Ю., Лиги Х.М. О связи между антифеодальными выступлениями крестьян и их положением // История СССР. 1976. № 2.

Кащенко С.Г. Реформа 19 февраля 1861 г. на Северо-Западе России (Количественный анализ массовых источников). М., 1995.

Ковальченко И.Д. Русское крепостное крестьянство в первой половине XIX века. М., 1967.

Ковальченко И.Д. Исторический источник в свете учения об информации (к постановке проблемы) // История СССР. 1982. № 3.

Ковальченко И.Д. Методы исторического исследования. М., 1987.

Ковальченко И.Д., Бородкин Л.И. Аграрная типология губерний Европейской России на рубеже XIX –XX веков (опыт многомерного количественного анализа) // История СССР. 1979. № 1.

Ковальченко И.Д., Бородкин Л.И. Структура и уровень аграрного развития районов Европейской России на рубеже XIX–XX веков (Опыт многомерного анализа) // История СССР. 1981. № 1.

Ковальченко И.Д., Милов Л.В. Всероссийский аграрный рынок. XVIII – начало XX в. (Опыт количественного анализа). М., 1974.

Ковальченко И.Д., Моисеенко Т.Л., Селунская Н.Б. Социально-экономический строй крестьянского хозяйства Европейской России в эпоху капитализма: (источники и методы исследования). М., 1988.

Ковальченко И.Д., Селунская Н.Б., Литваков Б.М. Социально-экономический строй помещичьего хозяйства Европейской России в эпоху капитализма: Источники и методы изучения. М., 1982.

Ковальченко И.Д., Сивачев Н.В. Структурализм и структурно-количественные методы в современной исторической науке // История СССР. 1976. № 5.

Количественные методы в гуманитарных науках. М., 1981.

Количественные методы в исторических исследованиях: Учеб. пособие / Под ред. И.Д. Ковальченко. М., 1984.

Количественные методы в советской и американской историографии: Материалы советско-американских симпозиумов. Балтимор, 1979, Таллинн, 1981. М., 1983.

Краткий клиометрический словарь / Сост. М.Г. Шендерюк. Калининград, 1994.

Массовые источники по социально-экономической истории России периода капитализма. М., 1979.

Массовые источники по социально-экономической истории советского общества. М., 1979.

Математические методы в исторических исследованиях. М., 1972.

Математические методы в исследованиях по социально-экономической истории. М., 1975.

Математические методы в историко-экономических и историко-культурных исследованиях. М., 1977.

Математические методы в социально-экономических и археологических исследованиях. М., 1981.

Математические методы и ЭВМ в исторических исследованиях. М., 1985.

Математические методы и ЭВМ в историко-типологических исследованиях. М., 1989.

Методы количественного анализа текстов нарративных источников. М., 1983.

Милов Л.В., Булгаков М.Б., Гарскова И.М. Тенденции аграрного развития России первой половины XVII столетия. Л., 1986.

Миронов Б.Н., Степанов З.В. Историк и математика. М., 1975.

Миронов Б.Н. История в цифрах. Л., 1991.

От Нестора до Фонвизина. Новые методы определения авторства / Под ред. чл.-кор. РАН Л.В. Милова. М., 1994.

Россия и США на рубеже XIX – XX столетий (Математические методы в исторических исследованиях). М., 1992.

Славко Т.И. Математико-статистические методы в исторических исследованиях. М., 1981.

Становление российского парламентаризма начала XX века / Под ред. Н.Б. Селунской. М., 1996.

Устинов В.А., Фелингер А.Ф. Историко-социальные исследования, ЭВМ и математика. М., 1973.

Хвостова К.В. Количественный подход в средневековой социально-экономической истории. М., 1980.

Хвостова К.В. Количественные методы в историческом познании // Вопросы истории. 1983. № 4.

СОДЕРЖАНИЕ