

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
АЛТАЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Жилин Сергей Иванович

**НЕСТАТИСТИЧЕСКИЕ МОДЕЛИ И МЕТОДЫ
ПОСТРОЕНИЯ И АНАЛИЗА ЗАВИСИМОСТЕЙ**

05.13.01 – системный анализ, управление и обработка информации
(физико-математические науки)

Диссертация
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.т.н., профессор Ю.А. Поляков

Барнаул – 2004

СОДЕРЖАНИЕ

Введение.....	3
1. Теоретические аспекты обработки информации: нестатистический подход.....	9
1.1 Проблемы построения и анализа эмпирических зависимостей.....	9
1.2 Обоснование постановок информационных задач, возникающих при построении и анализе эмпирических зависимостей.....	16
1.2.1 Формализация схемы эмпирического моделирования при нестатистическом подходе.....	16
1.2.2 Логическая схема построения и анализа зависимостей при нестатистическом подходе.....	27
1.2.3 Математические методы и модели решения информационных задач.....	31
1.3 Направления дальнейших исследований.....	39
2. Построение и анализ зависимостей методом центра неопределенности.....	42
2.1 Метод центра неопределенности: предпосылки, варианты и основные результаты.....	42
2.2 Обработка информации методом центра неопределенности при правильных наблюдениях.....	51
2.2.1. Задачи регрессионного анализа.....	51
2.2.2. Задачи дисперсионного и ковариационного анализов.....	57
2.2.3. Временные тренды в данных.....	60
2.3 Обработка информации в случае неправильных наблюдений.....	66
2.4 Метод центра неопределенности и статистические методы оценивания: сравнительный анализ.....	70
2.4.1. Метод центра неопределенности и метод максимального правдоподобия.....	73
2.4.2. Метод центра неопределенности и метод наименьших квадратов.....	79
3. Использование метода центра неопределенности при решении задач обработки пространственных данных.....	86
3.1 Геометрические преобразования и привязка изображений.....	86
3.2 Совместная обработка неравноточных инженерно-геодезических измерений.....	89
Заключение.....	97
Литература.....	99
Приложения.....	114

ВВЕДЕНИЕ

Актуальность темы. Проблема построения и анализа функциональной зависимости по эмпирическим данным встает перед многими исследователями в различных отраслях науки. С помощью конструируемых математических моделей реальных объектов, явлений и процессов в природе и обществе решаются задачи выяснения внутренних характеристик изучаемого процесса, сжатия массивов эмпирической информации, прогнозирования, управления и т.п. Задача анализа зависимости включает в себя выявление и устранение противоречий в исходных данных, проверку гипотез о виде искомой зависимости, ее свойствах и оценку степени работоспособности найденной функции, то есть ответ на вопрос о пригодности эмпирической зависимости на практике.

В настоящее время построение и анализ зависимостей по эмпирическим данным развивается чаще всего в рамках вероятностно-статистического подхода, основными инструментами которого является метод наименьших квадратов и статистические процедуры исследования и анализа полученных с его помощью оценок. Указанные методы обладают экстремальными свойствами при выполнении определенных предположений. Однако на практике обеспечение или проверка исходных гипотез по имеющейся совокупности эмпирических данных затруднена. В то же время имеются случаи, в которых предположения об ограниченности ошибки наблюдений являются более естественными. Эти обстоятельства привели к появлению нестатистического подхода, который развивался как самостоятельно, так и в рамках интервального анализа в работах отечественных исследователей А.Ф. Бочкова, А.П. Воцинина, Л.В. Канторовича, А.В. Максимова, Н.М. Оскорбина, А.Е. Померанцева, О.Е. Родионовой, В.А. Суханова и др., а также зарубежных авторов таких, как Г. Бельфорте, Е. Вальтер, В. Крейнович, М. Миланезе, Р.Е. Мур, Дж. Нортон, Л. Пронцато, Г.Р. Сотиров и других. Основное предположение нестатистического подхода состоит в том, что границы ошибок наблюдения переменных и невязок зависимости являются известными.

Такие задачи, в частности, возникают при геометрической коррекции цифровых изображений и при обработке инженерно-геодезических измерений.

Вместе с тем, в литературе отсутствуют исследования, устанавливающие связь статистического и нестатистического подходов и описывающие концептуальные основы моделирования с использованием нестатистического подхода. Недостаточное внимание уделено в рамках нестатистического подхода постановке и решению ряда значимых задач таких, как определение ценности наблюдений в общей совокупности, выделение активных наблюдений, устранение выбросов, моделирование процесса старения информации и др. Решение указанных проблем позволило бы расширить сферу применения нестатистического подхода, заимствовать ряд приемов из статистического подхода и разграничить области применения статистического и нестатистического подходов в прикладных исследованиях.

Цель исследования. Цель работы заключается в разработке концептуальных основ и методов построения и анализа зависимостей при обработке наблюдений с интервально заданными нестатистическими ошибками.

Задачи исследования.

1. Разработка концептуальной модели процесса эмпирического моделирования при нестатистическом подходе.
2. Разработка логической схемы процесса эмпирического моделирования.
3. Разработка способов построения зависимостей, включающих количественные и качественные факторы, с использованием метода внешней аппроксимации множества неопределенности гиперпараллелепипедом (метода центра неопределенности) по интервальным наблюдениям без выбросов.
4. Разработка метода обработки совокупности интервальных наблюдений с выбросами при построении и анализе зависимостей.
5. Экспериментальное сравнение метода центра неопределенности с основными статистическими методами построения зависимостей.

6. Апробация разработанных нестатистических методов построения и анализа зависимостей при решении задач обработки пространственно-распределенных интервальных измерений.

Объектом исследования является обработка наблюдений с интервально заданными нестатистическими ошибками.

Предметом исследования являются математические задачи построения и анализа зависимостей в рамках нестатистического подхода, сравнение статистических и нестатистических методов оценивания, нестатистические методы обработки наблюдений с выбросами, нестатистические методы построения зависимостей, включающих количественные и качественные факторы.

Методы исследования. При выполнении работы использовались методы системного анализа, математического программирования, математической статистики, имитационного моделирования, обработки изображений.

Научная новизна.

1. Предложена теоретико-множественная концептуальная модель процесса обработки информации при нестатистическом подходе к построению и анализу зависимостей, позволяющая формализовать информационные задачи сжатия данных, определения ценности наблюдений, учета старения данных, планирования одиночных наблюдений, и разработана логическая схема обработки совокупности наблюдений.

2. Разработана методика сравнения метода центра неопределенности с методами максимума правдоподобия и наименьших квадратов при обработке наблюдений без выбросов и по результатам вычислительного эксперимента определены области преимущественного использования сравниваемых методов.

3. Разработан нестатистический метод обработки интервальных наблюдений с выбросами, апробированный на реальных данных.

Теоретическая и практическая значимость результатов работы.

Разработанная концептуальная теоретико-множественная модель позволяет формализовать и выработать математические постановки задач обработки информации, возникающие при эмпирическом моделировании в рамках нестатистического подхода.

Выводы об областях преимущественного использования статистических и нестатистических методов оценивания могут использоваться для обоснованного выбора способов построения и анализа зависимостей по эмпирическим данным в условиях отсутствия достоверной информации о распределении ошибок.

Предложенные в работе методы обработки наблюдений с интервально заданной нестатистической ошибкой могут быть использованы при обработке данных физического эксперимента, геодезических измерений, данных дистанционного зондирования.

Основные положения, выносимые на защиту.

1. Формализация эмпирического моделирования, включающая концептуальную теоретико-множественную модель и логическую схему процесса построения и анализа зависимостей при нестатистическом подходе.

2. Результаты сравнения статистических и нестатистических методов построения и анализа зависимостей, проведенного на основе вычислительного эксперимента, и разграничение областей преимущественного применения указанных методов.

3. Методика построения и анализа зависимостей в рамках нестатистического подхода, апробированная при решении задач обработки пространственных данных.

Апробация работы. Основные положения и отдельные результаты исследования докладывались и обсуждались на Республиканской научно-технической конференции "Региональные проблемы информатизации", (Барнаул, 1995), третьей Международной конференции "Методы дистанционного зондирования и ГИС-технологии для контроля и диагностики состояния ок-

ружающей среды" (Москва, 1996), на второй и четвертой Международных конференциях Интеркарто (Иркутск, 1996; Барнаул, 1998), на Международной научно-практической конференции «Историческая и современная картография в развитии Алтайского региона» (Барнаул, 1997), первой и шестой краевых конференциях по математике (Барнаул, 1998; Барнаул, 2003).

Публикации. По теме диссертации опубликовано 13 печатных работ, в том числе 6 статей в журналах и сборниках статей и 7 тезисов докладов на конференциях.

Структура и объем работы. Диссертация состоит из введения, трёх глав, заключения, списка литературы из 143 источников, двух приложений. Общий объем работы составляет 119 страниц.

В первой главе рассматриваются проблемы построения и анализа эмпирических зависимостей с помощью статистических методов, излагаются основные идеи нестатистического подхода, в рамках которого проводится формализация процесса эмпирического моделирования и построение поэтапной логической схемы указанного процесса. При этом выделяются информационные задачи, которые возникают при реализации разработанной логической схемы, и указываются математические инструменты (модели и методы), необходимые для решения этих задач.

Вторая глава посвящена разработке нестатистических методов обработки интервальных наблюдений. В частности, предлагаются способы использования метода центра неопределенности для решения задачи построения зависимости по эмпирическим данным при наличии в ее структуре не только количественных, но и качественных факторов. Разрабатывается метод обработки совокупности наблюдений с выбросами. Методом статистических испытаний производится сравнение метода центра неопределенности с методами максимального правдоподобия и наименьших квадратов, и по результатам сравнения на модельной зависимости указываются области преимущественного использования сравниваемых методов.

В третьей главе приведены решения двух прикладных задач обработки пространственных данных с использованием МЦН и моделей, разработанных во второй главе.

В приложениях содержится ряд вспомогательных материалов, которые иллюстрируют результаты расчетов, выполненных в третьей главе.

Автор выражает глубокую благодарность к.ф.-м.н., доц., зав. кафедрой информатики Алтайского госуниверситета Максимову А.В. за научное консультирование при разработке теоретических аспектов нестатистического подхода в ходе работы над диссертацией, д.т.н., проф., зав. кафедрой теоретической кибернетики и прикладной математики Алтайского госуниверситета Оскорбину Н.М за обсуждение постановок ряда задач и своему научному руководителю д.т.н., проф. Полякову за ряд ценных замечаний, способствовавших как лучшему пониманию проблем обработки информации, так и их решению.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ОБРАБОТКИ ИНФОРМАЦИИ: НЕСТАТИСТИЧЕСКИЙ ПОДХОД

1.1. Проблемы построения и анализа эмпирических зависимостей

При изучении какого-либо явления принципиально не удастся учесть все бесконечное множество связей и факторов, от которых зависит протекающие моделируемого процесса. Наличие неучтенных связей приводит к необходимости появления в структуре модели составляющих, так или иначе отражающих элемент неопределенности, вносимый неучтенными связями в «закономерный» ход процесса. Довольно широкий класс конструируемых зависимостей выражается в форме, разделяющей детерминированную и неопределенную составляющие:

$$y = F(x, \beta) + \varepsilon_y, \quad x \in X, \quad (1.1)$$

где x – входной вектор; β – вектор параметров; X – множество значений x ; y – выходная переменная, которая ради простоты записей предполагается скалярной и, соответственно, F – скалярная функция, которая описывает детерминированную составляющую вариации y ; ε_y – величина, описывающая непредсказуемую составляющую исследуемого объекта.

Для описания факторов неопределенности могут использоваться различные модели: вероятностно-статистическая, интервальная, нечеткая [10, 31].

Статистическое описание неконтролируемых факторов предполагает возможность приписывания величине ε_y вероятностного, случайного характера. Случайные факторы полностью стохастически описаны, если задана их плотность распределения вероятностей. В большинстве ситуаций в условиях ограниченного эксперимента удастся получить лишь выборочные оценки параметров плотности, точность которых определяется планом эксперимента, числом опытов, дисперсией помехи, методом оценивания и т.д.

Вероятностно-статистический подход описания неопределенности и основанные на нем методы построения и анализа зависимостей имеют наибольшее распространение среди прочих методов восстановления зависимостей [6, 20, 26]. В частности, повсеместное применение находят классические методы параметрической статистики: регрессионный, дисперсионный, ковариационный анализы [2-4, 24, 36, 37, 52, 54, 64, 65, 70-71].

Согласно [78], регрессия – это зависимость среднего значения случайной величины от некоторой другой как случайной, так и неслучайной одной или нескольких величин, т.е. условное математическое ожидание:

$$M[y | x] = f(x).$$

Таким образом, функция регрессии описывает вероятностное соотношение между объясняющими переменными (регрессорами, факторами, входными, независимыми переменными) и зависимой переменной (откликом выходной, результирующей переменной). Естественным первым приближением для функции регрессии является ее линеаризация, и соответствующая модель носит название линейной регрессии. Для заданной совокупности N наблюдений предлагается следующее функциональное соотношение между реализовавшимся значением зависимой переменной и регрессорами:

$$y = X\beta + \varepsilon, \quad (1.2)$$

где $y = (y_1, \dots, y_N)^T$ – зависимая переменная; в матрице $X = (x_{ij})_{N \times n}$ построчно записаны наблюдения $x_i \in R^n$, $i = 1, \dots, N$ за объясняющими переменными X_1, \dots, X_n ; $\beta = (\beta_1, \dots, \beta_n)^T$ – вектор параметров; $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)^T$ – вектор невязок модели объекта.

В классической модели линейной регрессии помимо функционального соотношения (1.1), формулируются дополнительные предположения о стохастической структуре модели [4, 54]:

$$M\varepsilon_i = 0, \quad (1.3)$$

$$M\varepsilon_i^2 = \sigma^2, \quad (1.4)$$

$$M\varepsilon_i\varepsilon_j = 0, \forall i \neq j, \quad (1.5)$$

$$\text{rk } X = n < N, \quad (1.6)$$

$$X_j \text{ детерминированы}, \quad (1.7)$$

$$\varepsilon_i \sim N(0, \sigma^2). \quad (1.8)$$

При подобных предположениях основным способом оценки параметров β является метод наименьших квадратов (МНК):

$$\hat{\beta}_{\text{МНК}} = \arg \min_{\beta} \sum_{i=1}^N (y_i - x_i \beta)^2 = (X^T X)^{-1} X^T y. \quad (1.9)$$

Теоретическим обоснованием метода наименьших квадратов служит Теорема (Гаусс, Марков) *МНК-оценки являются несмещенными и эффективными линейными оценками при выполнении условий (1.2)-(1.7), и имеют нормальное распределение при дополнительном предположении (1.8).*

При выполнении предположения (1.8) МНК-оценки совпадают с оценками максимального правдоподобия и обладают свойством состоятельности.

Именно благодаря наличию таких свойств, как несмещенность, эффективность и состоятельность, МНК нашел широкое применение в прикладной статистике. Однако, классическая модель линейной регрессии достаточно чувствительна к нарушениям базовых предположений, при которых полезные свойства МНК-оценок, устанавливаемые теоремой Гаусса-Маркова, могут потерять свою силу. Недостатком МНК-оценок является и то, что эти потери трудно контролировать на практике.

Например, нарушение условий на вторые моменты – условия гомоскедастичности (1.4) и условия независимости (1.5) – приводит к тому, что МНК-оценки перестают быть эффективными в своем классе. Кроме того, МНК-оценка ковариационной матрицы оценок коэффициентов оказывается смещенной и несостоятельной, из-за чего тесты на значимость коэффициентов показывают неверный уровень значимости. Как правило, теоретические оценки дисперсии оценок коэффициентов занижаются, т.е. оценки оказываются слишком «оптимистическими». На практике же часто имеет место

своеобразная неустойчивость оценки точности МНК-оценок. Она заключается в том, что при сколь угодно малых отклонениях от заданных значений математического ожидания и ковариационной матрицы ошибок исходных данных относительная погрешность оценки точности конечного результата неограниченно возрастает с увеличением числа используемых наблюдений. Это делает иллюзорным свойство состоятельности оценки, так как оно является асимптотическим, то есть при конечном числе измерений привлечение дополнительной информации в ряде случаев может ухудшать точность МНК-оценки. Всегда на практике имеющее место любое сколь угодно малое отклонение от заданных значений математического ожидания и ковариационной матрицы ошибок приводит к нарушению свойства состоятельности [33, 103]. Это позволяет автору работы [103] назвать свойство состоятельности МНК-оценок одним из «мифов XX века».

Идентификация нарушения условий на вторые моменты ошибок представляет собой нетривиальную проблему. Известен, однако, ряд задач, в которых эти условия априорно нарушены. В первую очередь, это задачи анализа временных рядов, а также анализ стратифицированных обследований. Гетероскедастичность можно обнаружить (тестами Гольдфельда-Куандта, Бройша-Пагана и т.п.) лишь в специальных случаях (например, при независимости наблюдений), сделав дополнительное предположение об определенной функциональной форме этой зависимости [36, 54]. В общем случае гетероскедастичность без дополнительных предположений выявить, учесть и нейтрализовать невозможно: ковариационная матрица ошибок содержит $N \cdot (N - 1) / 2$ неизвестных, оценить которые по N наблюдениям невозможно.

Указанные нарушения основных свойств метода наименьших квадратов становятся особенно существенными при большом числе используемых наблюдений. Именно поэтому на них было обращено внимание в последние десятилетия, когда число измерений, проводимых в процессе наиболее ответственных экспериментов, стало достигать тысяч и десятков тысяч. При этом некритическое использование результатов теории метода наименьших квад-

ратов приводило часто к грубым ошибкам при решении важных прикладных задач. Примеров неработоспособности метода наименьших квадратов накопилось довольно много [103].

Одним из самых сложных случаев для анализа чувствительности МНК-оценок является нарушение предположения о виде распределения ошибок. Как правило, нельзя указать каких-либо веских причин, по которым конкретное распределение результатов наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог приближается логарифмически нормальным распределением, и т.д. Однако в подавляющем большинстве реальных ситуаций подобных моделей нет, и приближение реального распределения с помощью некоторого теоретического – чисто формальная операция [68, 71].

Неоднократно публиковались экспериментальные данные, свидетельствующие о том, что распределения реально наблюдаемых случайных величин, в частности, ошибок измерения, в подавляющем большинстве случаев отличны от нормальных [66, 73, 103]. Тем не менее, теоретики продолжают строить и изучать статистические модели, основанные на нормальности, а практики – применять подобные методы и модели, то есть, по меткому выражению А.И. Орлова [73], "ищут под фонарем, а не там, где потеряли".

Наиболее остро (возможно, излишне) критика некорректных случаев применения регрессионного анализа и статистических методов вообще прозвучала в работах [8, 9]. В этих работах отмечается, что математическую статистику нельзя считать наукой в строгом смысле этого слова, так как невозможно проверить на практике достоверность полученных с ее помощью результатов. В качестве примера «неверифицируемых» характеристик приводятся понятие генеральной совокупности, доверительного интервала на неизвестное среднее случайной величины, ошибки первого и второго рода при

проверке гипотез и др., которые невозможно проверить на реальных экспериментальных установках. Более того, ставится под сомнение сама возможность воспроизведения статистического эксперимента в различных лабораториях. Полемика вокруг подобных утверждений привела к более четкому пониманию границ применимости вероятностно-статистических методов и выявила ограничения слишком жесткой системы исходных допущений регрессионного анализа и основанных на нем процедур обработки данных [4, 62, 90, 91].

Многие прикладные задачи не укладываются в рамки статистического подхода к интерпретации неопределенности, поскольку, во-первых, само понятие неопределенности часто не ограничивается рамками случайности, а может быть трактовано и как незнание, и как неединственность возможных исходов и т.п. [10, 27, 31], а, во-вторых, многие ситуации характеризуются принципиальным несоблюдением условий статистического ансамбля при сборе информации о моделируемом явлении [4].

Одним из наиболее естественных способов описания факторов неопределенности является их интервальное представление, когда относительно ε_y ничего неизвестно, за исключением свойства быть ограниченным, то есть принадлежащим некоторому интервалу $[\varepsilon_y, \bar{\varepsilon}_y]$. При этом полагается, что ε_y может принимать любые значения из этого интервала, то есть ему не приписывается никакой вероятностной меры. По этой причине такой подход часто именуют нестатистическим [22,30, 31].

По-видимому, первой отечественной работой, указывающей на перспективность развития методов обработки наблюдений, оперирующих ограниченными неопределенностями, является работа Л.В. Канторовича [51].

Впоследствии подобный подход нашел широкое применение при решении разнообразных прикладных задач, в частности, задач идентификации, анализа и управления системами в условиях ограниченных возмущений [48, 57, 58, 98, 110, 113, 120, 124], задач оптимизации в условиях неопределенно-

сти [10, 29, 31, 110], и, наконец, при оценивании параметров эмпирических зависимостей [22, 28, 30, 105, 109, 114, 120, 126, 127].

Интервальное описание неопределенности позволяет взглянуть на процесс эмпирического моделирования с иных, нестатистических позиций. В своем исследовании мы ставим задачу формализации процесса эмпирического моделирования, делая акцент на те возможности, которые открываются при нестатистическом подходе к интерпретации неопределенности и которые часто “исчезают” или связаны с существенными затруднениями при вероятностно-статическом подходе.

Одной из таких фундаментальных возможностей, которой должен располагать исследователь в процессе эмпирического моделирования, является возможность выделять наблюдения, содержащие выбросы. Совокупность наблюдений, в которой отсутствуют «выбросы» назовем *правильными наблюдениями*. Наблюдения, содержащие выбросы (*неправильные наблюдения*), существенно «сдвигают» оценки, и выявление таких наблюдений представляется достаточно сложной задачей [69]. Способы обработки данных с выбросами изучаются в теории робастных оценок [73, 80, 92, 93, 96], в которой, однако, для этого приходится либо формулировать дополнительные гипотезы, либо привлекать дополнительную информацию, например, о доле выбросов в используемой совокупности. Необходимо также выполнение гипотез непротиворечивости информационных совокупностей и, следовательно, нужны и методы их проверки.

Еще одним важным инструментом является возможность рекуррентного оценивания эмпирической зависимости. Действительно, в ходе эксплуатации эмпирической модели в большинстве случаев необходимо постоянно формировать новые наблюдения и уточнять оценки. Традиционной в данном случае является проблема сложности оценивания параметров. Кроме того, пополнение совокупности эмпирической информации сопровождается процессом старения некоторых частей информации, и, следовательно, при накоплении эмпирической информации необходимо «удалять» старые данные.

При этом особое внимание следует обратить на ценность каждой порции информации.

Таким образом, актуальны системные исследования методов эмпирического моделирования и, в частности, дополнительные исследования в следующих направлениях.

1. Формализация процесса эмпирического моделирования при интервальной форме описания неопределенности (нестатистическом подходе).
2. Выделение частных информационных задач и разработка логической схемы, увязывающей их в единую процедуру построения и анализа зависимостей при интервальной форме описания неопределенности (нестатистическом подходе).
3. Разработка математических инструментов, используемых при решении информационных задач, возникающих при построении и анализе эмпирических зависимостей в рамках нестатистического подхода.

1.2. Обоснование постановок информационных задач, возникающих при построении и анализе эмпирических зависимостей

1.2.1. Формализация схемы эмпирического моделирования при нестатистическом подходе

В данном разделе решается первая из выделенных в разделе 1.1 задач системного анализа проблемы эмпирического моделирования – проводится формализация схемы эмпирического моделирования, которая выступает новой методологической основой построения и анализа зависимостей. Конкретным объектом формализации является принципиальная схема обработки данных и знаний при эмпирическом моделировании, изображенная на рис. 1.1. Формализация должна коснуться предположений об обрабатываемых данных и знаниях; гипотез, выдвигаемых относительно исследуемого объекта; форм описания конечного результата. Главным результатом формализа-

ции должен стать перечень и математическое описание задач, возникающих в процессе обработки эмпирической информации.



Рис. 1.1. Принципиальная схема процесса обработки данных и знаний при эмпирическом моделировании.

При вероятностно-статистическом подходе к эмпирическому моделированию фундаментом для формализации является предположение о существовании генеральной совокупности исследуемых объектов, на которой задано некоторое распределение вероятностей. Математический аппарат вероятностно-статистических методов построения и анализа зависимостей направлен на исследование характеристик этого распределения по ограниченной совокупности эмпирических данных, полученных при соблюдении условий статистического ансамбля [4] (статистической устойчивости [9]) и построение вытекающих отсюда выводов о свойствах исследуемого объекта.

Поскольку нестатистический подход состоит в отказе от такого способа формализации ситуации, в которой получена ограниченная совокупность эмпирических данных и знаний, необходимо найти иной принцип, по своей роли совпадающий с предположением о существовании генеральной совокупности в вероятностно-статистическом подходе. Такой принцип должен служить основанием для формализации всего процесса эмпирического моделирования и определять соотношение реальных исследуемых объектов и их математических описаний, т.е. являться важным элементом логического обоснования возможности использования *формальных* математических методов и моделей при исследовании *реальных* объектов. Анализ литературы, посвященной нестатистическим методам построения зависимостей [22, 28, 30, 51, 105, 109, 114, 120, 126, 127], показывает, что эта проблема логическо-

го обоснования применимости математического аппарата нестатистического подхода не ставится, формализация проводится на уровне математической модели исследуемой зависимости, а особенности процессов сбора исходных данных и перенесения математических выводов в реальность не формализуются.

С целью обоснованного использования математического языка для описания процесса эмпирического моделирования обсудим основания соотношения эмпирических моделей с реальностью.

Основой эмпирического моделирования и эмпирических моделей исследуемых объектов являются наблюдения за совокупностью объектов, выступающих по отношению к исследуемому в качестве объектов-аналогов. Чаще всего объектом-аналогом выступает сам объект с его поведением в прошлом. При моделировании часто возникает задача отделения в общей информационной совокупности объектов-аналогов от объектов, которые таковыми не являются. Например, задача устранения выбросов в совокупности наблюдений является, по сути, задачей устранения из общей информационной совокупности объектов, которые не могут выступать в качестве объектов-аналогов. Процесс обработки стареющей информации также есть процесс разделения объектов-аналогов и не объектов-аналогов, которые возникают в результате понижения их относительной информативности в общей совокупности.

Основанием переноса полученных по объектам-аналогам знаний на исследуемый выступает базовый принцип эмпирического познания «так было – так будет» (принцип эмпирической индукции) [9]. Хорошо известно, что этот принцип далек от безупречности. Для эмпирического моделирования в основе подхода заложен риск, что этот принцип в каждом конкретном случае может не выполняться, то есть выводы, полученные по объектам-аналогам, будут неверными. Этот риск должен, так или иначе, учитываться при построении эмпирической зависимости. Кроме того, должны учитываться ограниченность числа объектов-аналогов, неидеальность наблюдателей объектов-

аналогов и степень «усвоения знаний» при выбранном методе обработки данных. Таким образом, основными причинами снижения точности описания моделируемого объекта выступают:

1. Возможность нарушения основного принципа эмпирического моделирования (принципа «так было – так будет»).
2. Ограниченность числа объектов-аналогов и их неоптимальное сочетание.
3. Неидеальность наблюдателей за объектами-аналогами.
4. Низкий уровень «усвоения знаний» используемого метода обработки данных.

Отметим, что принцип «так было – так будет» является базовым и в мизесовском подходе в теории математической статистики [8, 9], однако формализация самой схемы эмпирического моделирования в этих работах не проводится.

Таким образом, несмотря на недостатки, присущие принципу «так было – так будет», он согласуется с естественнонаучными традициями и может быть принят в качестве основы формализации процесса эмпирического моделирования при нестатистическом подходе. Что касается развиваемого нами математического аппарата, то он тесно связан с подходом, изложенным в [112], где, однако, не ставится вопрос о соотношении предлагаемых математических конструкций с реальными исследуемыми объектами.

Итак, определим вначале возможные формы описания эмпирических зависимостей. Предлагается использовать две формы: параметрическую и интервальную. При описании детерминированной части модели объекта (1.1) в параметрической форме

$$y_0 = F(x, \beta), \quad \beta \in B(x), \quad x \in X, \quad (1.10)$$

ключевым элементом является множество $B(x)$ значений параметров β , определяющее семейство допустимых функциональных зависимостей. Интервальная форма описания семейства допустимых зависимостей состоит в указании нижней ($F^h(x)$) и верхней ($F^g(x)$) границ для $F(x, \beta)$:

$$F^h(x) \leq F(x, \beta) \leq F^g(x). \quad (1.11)$$

Очевидно, что описание (1.11) может быть получено по известному описанию (1.10):

$$\begin{aligned} F^h(x) &= F(x, \beta^h) = \min_{\beta \in B(x)} F(x, \beta); \\ F^g(x) &= F(x, \beta^g) = \max_{\beta \in B(x)} F(x, \beta). \end{aligned} \quad (1.12)$$

Следует, однако, заметить, что интервальная форма описания позволяет охватить более широкое семейство зависимостей по сравнению с параметрической формой и может использоваться в случаях, когда неизвестна структура восстанавливаемой функции $F(x, \beta)$. Обобщая интервальную форму описания зависимостей, можно говорить о множестве \mathbf{F} всех функций, построенных по заданной совокупности объектов-аналогов, предполагая при этом, что для искомой функции $F(x, \beta)$ выполняется условие принадлежности:

$$F(x, \beta) \in \mathbf{F}. \quad (1.13)$$

Обобщенную форму описания зависимости целесообразно использовать в случаях, когда вид связи входных и выходных переменных моделируемого объекта неизвестен и при эмпирическом моделировании используются те или иные рабочие гипотезы. При этом в реальной практике моделирования могут возникать случаи, когда для выбранной функции условие (1.13) не выполняется.

Далее, если не оговорено особо, мы предполагаем, что $F(x, \beta)$ скалярная функция и известна с точностью до параметров; наблюдения за переменными x, y являются правильными; наблюдаемые значения входных переменных $x_i, i = 1, \dots, N$ для всех объектов-аналогов известны точно (измеряются с пренебрежимо малой погрешностью), а соответствующие им значения выходной переменной y_i измеряются с ошибками ε_{y_i} , для которых известны оценки

$$|\varepsilon_{y_i}| \leq \bar{\varepsilon}_{y_i}, \quad i = 1, \dots, N, \quad (1.14)$$

Определение 1.1. *Порцией информации назовем группу наблюдений (данных) и знаний, которая позволяет сформировать самостоятельное ограничение на параметры β*

$$\underline{y}(x) \leq F(x, \beta) \leq \bar{y}(x), \quad x \in X' \subseteq X. \quad (1.15)$$

Следует подчеркнуть важность введенного понятия порции информации как отражения существенных сторон реальных информационных процессов при эмпирическом моделировании. Понятие «порция информации» соединяет в себе такие важные информационные понятия как «данные» и «знания». При сборе, хранении, обработке эмпирической информации эти понятия не разделимы. Выражение (1.3) можно воспринимать как некоторый предикат, т.е. один из способов выражения знаний, а сужение области действия этого предиката производится за счет подстановки в него конкретных данных. Как будет показано ниже (примеры 1.1 и 1.2), иногда порции информации могут образовываться исключительно знаниями (априорной информацией). При этом знания могут отражать не только факты, но и являются способом выражения гипотез. Введение понятия «порция информации» позволяет перевести словесные формулировки знаний и данных в теоретико-множественные соотношения.

Определение 1.2. *Множество B значений параметров β , удовлетворяющих (1.15), назовем информационным множеством.*

Определение 1.3. *Множество B_i значений параметров β , удовлетворяющих неравенству*

$$y_i - \bar{\varepsilon}_{y_i} \leq F(x_i, \beta) \leq y_i + \bar{\varepsilon}_{y_i}, \quad (1.16)$$

назовем информационным множеством объекта-аналога i

Определение 1.4. *Априорным назовем информационное множество, сформированное только на основе знаний о свойствах моделируемого объекта (знаний теоретического характера).*

Приведем примеры, иллюстрирующие определение 1.4.

Пример 1.1. Пусть известно, что $F(x, \beta)$ – неотрицательная функция. Тогда $B(x) = \{\beta \in R^n \mid F(x, \beta) \geq 0\}$ – информационное множество, сформированное только на основе этих знаний, т.е. представляет собой априорную информацию.

Пример 1.2. Пусть относительно параметров функции $F(x, \beta)$ известно, что первая компонента вектора $\beta = (\beta_1, \dots, \beta_n)$ неотрицательна. Тогда априорной информацией, отражающей эти знания, является информационное множество $B = \{\beta \in R^n \mid \beta_1 \geq 0\}$.

Использование введенных понятий порции информации и информационного множества позволяет основным инструментом процесса обработки информации при эмпирическом моделировании сделать теоретико-множественные операции. Суть обработки информации в этом случае сводится к выяснению того, каким образом соотносятся между собой, а также со свойствами моделируемого объекта информационные множества B_i , $i = 1, \dots, N$, сформированные на основе эмпирической информации, априорных знаний и т.д. Основным принцип при этом дает следующее

Определение 1.4. *Информационное множество B , не противоречит свойствам моделируемого объекта, если выполнено включение*

$$\beta \in B, \quad (1.17)$$

где β – «истинное» значение параметров.

При эмпирическом моделировании проверить условие (1.17) не представляется возможным. В связи с этим расширим принцип эмпирического моделирования «так было – так будет», включив в него условие: *если пересечение информационных множеств, имеющих при построении модели, не пусто, то есть*

$$B(N) = \bigcap_{i=1}^N B_i(x) \neq \emptyset, \quad (1.18)$$

то условие (1.17) выполнено.

Условие (1.18) можно усилить введением эмпирически формируемого множества B_{\min}^N , которое можно трактовать как «минимальное» допустимое. Этот вариант условия имеет вид:

$$B(N) = \bigcap_{i=1}^N B_i(x) \supseteq B_{\min}^N. \quad (1.18')$$

Замена условия (1.17) на (1.18) или (1.18'), несущественная при эмпирическом моделировании, требует при построении модели определенной научной добросовестности в процессе формирования множеств B_i . Легко привести примеры, когда (1.17) выполнено, а (1.18) не выполнено. В частности, в случае, когда неверно выбран вид функции $F(x, \beta)$. Развитие и формализация этого случая в эмпирическом моделировании представляются перспективными, однако, в данной работе не проводятся.

Ниже мы предполагаем, что (1.17) выполнено, а (1.18) используется лишь для контроля данного факта. Тогда задача эмпирического моделирования зависимости $y = F(x, \beta)$ решается в предположении (1.18).

Приведем графическую иллюстрацию к описанному принципу эмпирического моделирования на примере построения модели, которая является линейной по параметрам:

$$y = \sum_{j=1}^n \beta_j \cdot \varphi_j(x), \quad (1.19)$$

где $\varphi_j(x)$, $j = 1, \dots, n$ – заданная система базисных функций.

Графически проиллюстрируем случай $n = 2$, $\varphi_1(x) = 1$, $\varphi_2(x) = x$, т.е. задача состоит в определении коэффициентов β_1 , β_2 прямой $y = \beta_1 + \beta_2 x$ по экспериментальным данным (y_i, x_i) , $i = 1, \dots, N$, если модуль ошибки ε_{y_i} оценен сверху величиной $\bar{\varepsilon}_{y_i}$.

В данном случае информационные множества B_1, B_2, \dots, B_N заданы следующими двусторонними неравенствами

$$y_i - \bar{\varepsilon}_{y_i} \leq \beta_1 + \beta_2 x_i \leq y_i + \bar{\varepsilon}_{y_i}, \quad i = 1, \dots, N. \quad (1.20)$$

Следуя работе [35], эти множества иногда будем называть “полосками”, не только на плоскости (рис. 1.2.б), но и в многомерном пространстве. Название объясняется тем, что в линейном случае (1.19) при $n = 2$ границы множеств B_i – две параллельные прямые (при $n > 2$ – гиперплоскости), что непосредственно следует из (1.20).

На рисунке 1.2.б) приведено множество $B(N)$, которое получается в результате пересечения (1.18) множеств B_1, \dots, B_N . Практически, это означает, что совокупность информационных множеств B_i мы рассматриваем как систему неравенств. Современные вычислительные средства позволяют эффективно «работать» с этой системой.

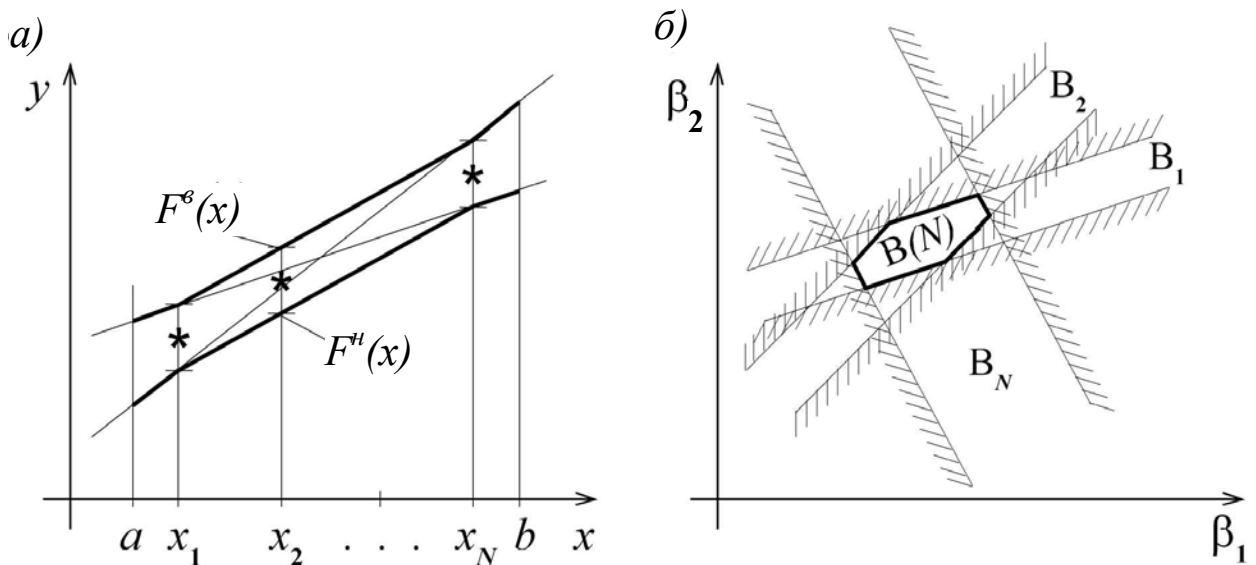


Рис. 1.2. Идентификации модели (1.19) а) в пространстве (x, y) ; б) в пространстве параметров (β_1, β_2) .

Между множеством параметров $B(N)$ и множеством функций $F(x, \beta)$, $(F^h(x) \leq F(x, \beta) \leq F^g(x))$ устанавливается взаимно однозначное соответствие. Каждой точке $\beta^0 \in B(N)$ соответствует единственная функция $y = F(x, \beta^0)$ и наоборот. Этот факт может быть использован в технологии обработки данных и знаний при эмпирическом моделировании.

Выделим два фундаментальных для эмпирического моделирования свойства множества $B(N)$. Первое связано с его «размером» и непротиворечивостью наблюдений и априорных знаний. Формальный признак непроти-

воречивости исходной информации выражается условием (1.18), которое можно на содержательном уровне выразить так. Если $B(N)$ пусто, то исходные данные и знания противоречивы, однако, нельзя утверждать, что если $B(N)$ не пусто, то противоречий нет. Более реальным для практики может оказаться индикатор непротиворечивости, выражающий *скорость уменьшения меры множества неопределенности* $B(N)$ при $N \rightarrow \infty$. Если для множества $B(N)$ известна или легко конструируется его внутренняя аппроксимация, то динамика объема аппроксимирующего множества и ее сравнение с нормативной (закономерной) при росте N могли бы быть индикатором непротиворечивости исходных предположений и наблюдений. Этот прием является еще одним способом усиления условия (1.18).

Второе фундаментальное свойство $B(N)$ связано с его ограниченностью. Можно утверждать, что если множество $B(N)$ содержит бесконечно удаленные точки, то исходных данных для построения модели недостаточно. Следует иметь в виду, что если $B(N)$ ограниченное множество, то может оказаться, что эмпирическая модель не является работоспособной, т.е. конечные пользователи не получают результаты с требуемой точностью. В каждой конкретной ситуации необходимо разработать критерии на «размер» множества неопределенности, при выполнении которых эмпирическая модель становится работоспособной.

Рассматриваемый подход к эмпирическому моделированию применим и в случае, когда структура функции $F(x, \beta)$ неизвестна (см. (1.13)). Без детального анализа это утверждение иллюстрируется рисунком 1.3 (см. [32]), на котором представлена задача оценки скалярной функции скалярного аргумента, удовлетворяющая условию Липшица с константой L

$$|F(x_1) - F(x_2)| \leq L \cdot |x_1 - x_2|, \quad \forall x_1, x_2 \in X.$$

Предполагается, что x_i на интервале $[a, b]$ измерены точно, а y_i — с некоторой ошибкой, ограниченной по модулю сверху величиной $\bar{\varepsilon}_{y_i}$.

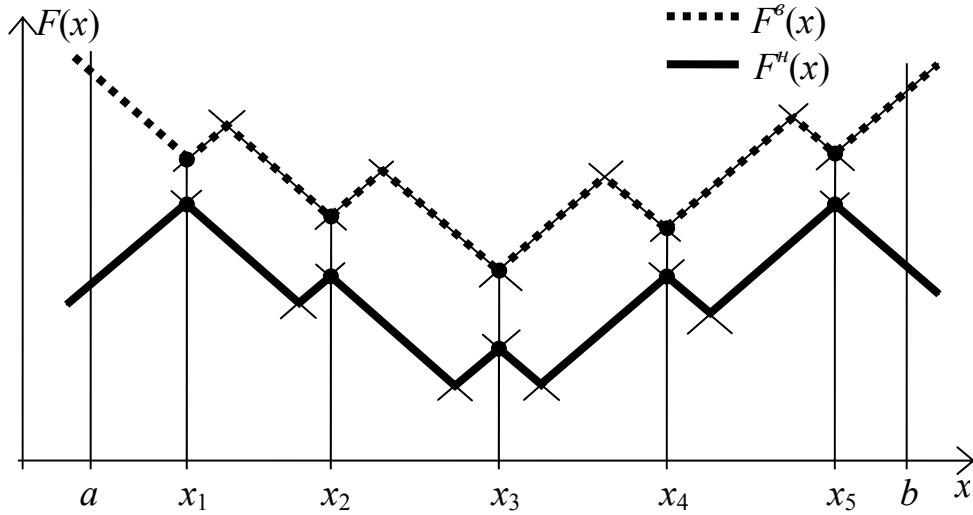


Рис. 1.3. Оценка функции $F(x)$ при неизвестной структуре, ($N = 5$).

Развитие методов построения и анализа эмпирических зависимостей в подобных ситуациях не входит в задачи нашего исследования. Поэтому ограничимся формальной записью функций $F^h(x)$, $F^6(x)$. Для этого определим Z_i – информационные множества в следующем виде:

$$Z_i = \{(y, x) \mid y_i - L(x - x_i) - \bar{\varepsilon}_{y_i} \leq y \leq y_i + L(x - x_i) + \bar{\varepsilon}_{y_i}\}, \quad i = 1, \dots, N.$$

Введем

$$Z = \bigcap_{i=1}^N Z_i.$$

Тогда

$$F^h(x) = \min \{y \mid (y, x) \in Z\}, \quad F^6(x) = \max \{y \mid (y, x) \in Z\}. \quad (1.21)$$

Таким образом, в данном случае мы имеем несколько иной способ формирования информационных множеств. Однако суть задач эмпирического моделирования при этом не меняется.

Итак, мы сформулировали основные понятия и принципы, используемые при эмпирическом моделировании в случае ограниченной неопределенности и ввели формализм, необходимый для математической постановки информационных задач, возникающих при эмпирическом моделировании. Следующий шаг в нашем исследовании состоит в выявлении таких информационных задач и построении принципиальной логической схемы, устанавливающей взаимосвязи между выделенными задачами.

1.2.2. Логическая схема построения и анализа зависимостей при нестатистическом подходе

Построение эмпирических зависимостей при нестатистическом подходе является многоэтапной процедурой, результатом которой является эмпирическая модель и оценки ее работоспособности. На каждом этапе этой процедуры приходится решать частные информационные задачи, логические взаимосвязи между которыми отражает схема, приведенная на рисунке 1.4. Дадим общую характеристику задач, решаемых на каждом этапе.

1. Принятие системно-технических решений по построению эмпирической зависимости. На первом этапе проводится изучение и анализ моделируемого объекта, объектов-аналогов, по результатам которых определяются входные и выходные переменные, и проводится обоснование возможности применения нестатистического подхода к построению и анализу зависимостей. На этом этапе формулируются требования к модели, определяется область пространства входных переменных, в рамках которой проводится моделирование, выбирается способ описания и, если возможно, вид эмпирической зависимости.

2. Сбор априорных знаний и данных. На этом этапе априорные знания о моделируемом объекте формулируются в виде порций информации, выбираются объекты-аналоги для наблюдений, формируется стартовая информационная совокупность B_0 (преимущественно на основе априорных знаний об объекте) и устанавливаются правила «поведения» множества неопределенности (минимальный допустимый «размер», темп сокращения и т.п.).

3. Планирование эксперимента и включение новых порций информации. Дополнительные к множеству B_0 порции информации B_1, \dots, B_N могут являться результатом формализации новых знаний о моделируемом объекте и сбора данных путем проведения «активного» или «пассивного» эксперимента [89]. «Активный» эксперимент позволяет собрать более качественные данные и дает возможность управлять размерами множества неоп-

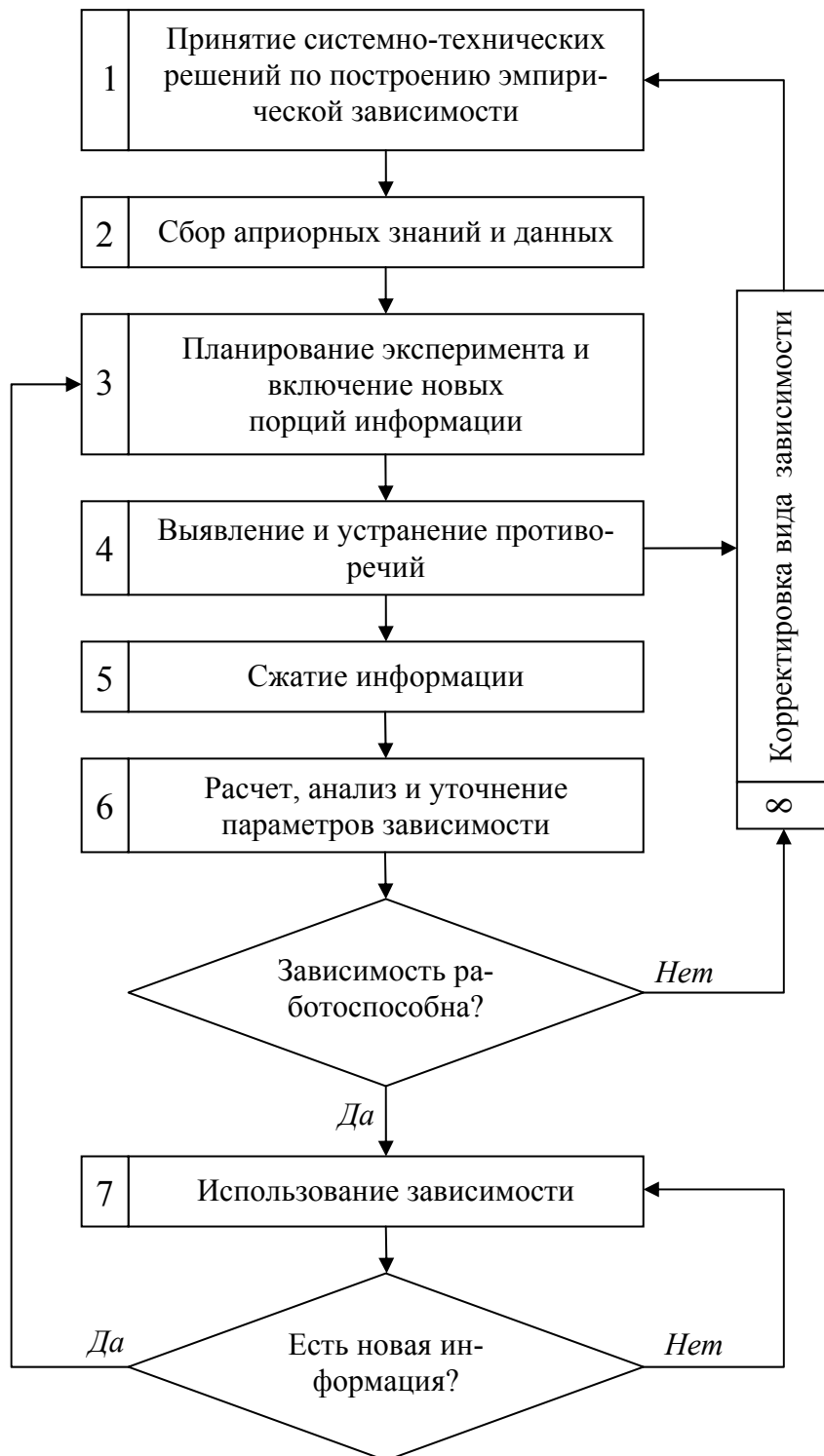


Рис. 1.4 Логическая схема построения и анализа эмпирических зависимостей при нестатистическом подходе.

ределенности путем указания точек факторного пространства (y, x) , для которых необходимо подобрать объекты-аналоги, однако его проведение требует дополнительных затрат. Затраты на «пассивный» эксперимент существенно ниже, однако именно в этом случае чаще всего приходится сталкиваться

ся с проблемой неправильных измерений, что порождает дополнительные затраты на их выявление и устранение.

4. Выявление и устранение противоречий. Противоречия во вновь образованной информационной совокупности могут быть обусловлены двумя причинами: неправильными измерениями, поставляющими порции информации, несовместимые с общей массой наблюдений, и несоответствием гипотетического вида зависимости, выбранного на этапе принятия системно-технических решений, накопленной информационной совокупности. В первом случае перед переходом к следующему этапу можно ограничиться поиском и устранением неправильных порций информации. Во второй ситуации приходится корректировать вид зависимости, что в общем случае требует повторного прохождения всех предыдущих этапов.

5. Сжатие информационных совокупностей. Включение каждой порции информации в общую информационную совокупность на этапах 2 и 3 сопровождается определением ее ценности. Содержание настоящего этапа состоит в выделении активных порций информации, то есть порций, имеющих ненулевую ценность. Порции информации, имеющие нулевую ценность, могут быть устранены, что позволяет сократить размерность задачи описания множества неопределенности.

6. Расчет, анализ и уточнение параметров зависимости. На этом этапе производится построение множества неопределенности или его приемлемых аппроксимаций, позволяющих выяснить интервалы изменения параметров конструируемой зависимости и, если необходимо, получить точечные оценки, оптимальные в некотором смысле. Анализ полученных интервалов значений параметров зависимости позволяет выдвигать гипотезы о их свойствах, например, о их значимости или незначимости. Проверка этих гипотез может проводиться непосредственным включением порций информации, описывающих те или иные предположения, в общую информационную совокупность. Наличие или отсутствие противоречий во вновь полученной ин-

формационной совокупности может служить основанием для отклонения или принятия гипотез соответственно.

7. Использование зависимости. После проверки полноты и работоспособности построенной зависимости модель может приниматься в эксплуатацию. Использование модельной зависимости сопровождается процессами старения имеющейся информации и накопления новых данных и знаний о моделируемом объекте, что делает необходимым повторение шагов 3-6 через определенные промежутки времени и подчеркивает актуальность разработки рекуррентных алгоритмов оценивания.

8. Корректировка вида зависимости. Изменение структуры зависимости, возможно сопровождаемое изменением состава входных переменных, требует дополнительных исследований моделируемого объекта и его объектов-аналогов. Переход к этому шагу производится в случае, когда на стадии выявления противоречий в информационной совокупности завершена процедура удаления выбросов, но множество неопределенности по-прежнему остается пустым, или, несмотря на непустоту множества неопределенности, модель оказывается неработоспособной.

Для реализации описанной логической схемы необходимы следующие математические инструменты (модели и методы).

1. Методы определения меры ценности порции информации B_i и всей совокупности B_1, \dots, B_N по отношению к исходному множеству B_0 .
2. Методы внешней и внутренней аппроксимации множества $B(N)$.
3. Методы точечного и интервального оценивания параметров зависимости и выходной переменной.
4. Методы сжатия информационной совокупности, то есть выделения активных порций информации.
5. Модель процесса старения информационных совокупностей в случае, когда свойства моделируемых объектов меняются во времени.
6. Методы обработки потока наблюдений, то есть бесконечной последовательности информационных множеств B_1, B_2, \dots

7. Методы определения оптимального плана эксперимента, т.е. точек (x_i, y_i) , $i = 1, \dots, N$, которые позволяют получить множество неопределенности $B(N)$ «минимальных размеров».
8. Методы обработки наблюдений с выбросами, то есть обработки данных и знаний в случае их искажения.

Ниже рассмотрим лишь проблемные вопросы, возникающие при решении задач с помощью этих инструментов.

1.2.3. Математические методы и модели решения информационных задач

Ценность информационных порций

Пусть известны информационные множества B_1, \dots, B_N , определяющие замкнутое и ограниченное множество неопределенности $B(N)$ параметров β и необходимо определить ценность очередной порции информации, выраженную в информационном множестве B_{N+1} . Для $N = 2$ и однофакторной линейно параметризованной зависимости пример этой ситуации представлен графически на рисунке 1.5, где множество $B(2)$ образовано пересечением порций B_1 и B_2 – многоугольник $ABDF$. При добавлении нового информационного множества B_3 получается множество неопределенности $B(3)$ – многоугольник $ABCEF$.

Мерой ценности порции B_{N+1} может служить отношение мер множеств $B(N) \setminus B(N+1)$ и $B(N+1)$, т.е. треугольника CDE и многоугольника $ABFD$ в нашем примере. Порция, определяющая B_{N+1} , имеет нулевую ценность по отношению к $B(N)$, если с ее получением множество $B(N+1)$ не изменяется по сравнению с множеством $B(N)$.

Информационная ценность порции, определяющей B_{N+1} , по отношению к $B(N)$ максимальна, если $B(N+1)$ содержит лишь одну точку β^* – истинное значение искомого параметра (при условии, что имеется включение $\beta^* \in B(N+1)$, то есть условие (1.17) выполнено).

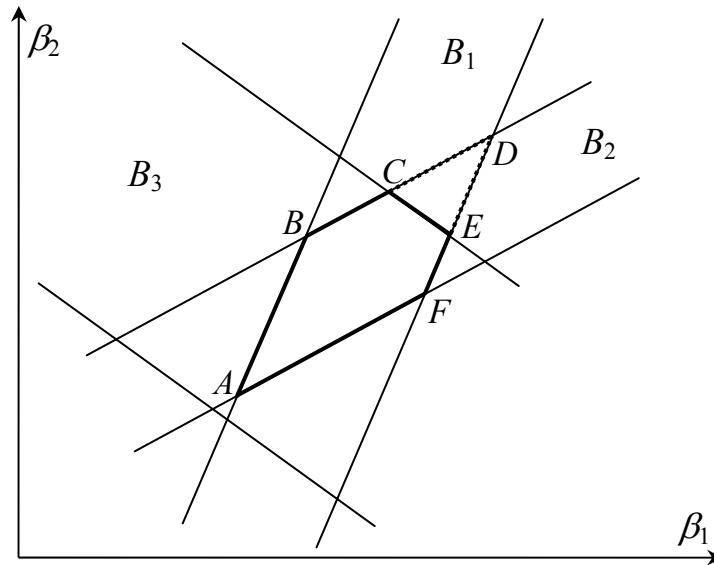


Рис. 1.5. Иллюстрация к задаче определения информационной ценности данных и знаний. Многоугольник $ABDF$ – множество неопределенности $B(2) = B_1 \cap B_2$. Многоугольник $ABCEF$ – множество неопределенности $B(3) = B(2) \cap B_3$.

Аналогично можно определять ценности произвольных парных, тройных и другие сочетаний порций информации и множеств неопределенности. Например, если $B(N|1)$ – множество неопределенности, полученное пересечением множеств B_2, \dots, B_N , то ценность сочетания порций, определяющих B_2, \dots, B_N , по отношению к $B(1)$ зависит от соотношения мер множеств $B(1)$ и $B(1) \setminus (B(1) \cap B(N|1))$.

При реализации описанной схемы определения ценности информации необходимо допускать, что рассматриваемая порция информации, определяющая B_1 , может быть сформирована на базе неправильных наблюдений или знаний, то есть может быть ложной ($\beta^* \notin B_1$). Так, если ценность этой порции максимальна, то это событие, вообще говоря, мало вероятно. На рисунке 1.5 в этом случае полоска B_1 расположится так, что B_1 будет содержать лишь точку A . Необходимо для каждой схемы определять нормативную меру ценности информационных порций и для порции B_1 либо проверять ее уникальность (то есть менять схему формирования данных), либо «загруб-

лать» порцию B_1 до нормативной ценности (в данном случае расширить порцию B_1). Эти рекомендации соответствуют условию (1.18').

Внешняя и внутренняя аппроксимация множества $B(N)$

Множество неопределенности $B(N)$ в общем случае эмпирического моделирования имеет сложную форму, не является выпуклым даже в случае построения линейной по параметрам модели. Исследование его свойств для некоторых классов моделей проведено в рамках интервальной математики (см., например, работу [100] и библиографию в ней). В связи с этим используются различные аппроксимации этого множества [18, 27, 28, 30, 100, 120, 135-137], среди которых наиболее простой является аппроксимация гиперпараллелепипедом. С информационной точки зрения допустима внешняя аппроксимация множества $B(N)$, которая, ухудшая в целом качество эмпирической модели, не вносит дезинформации (не нарушает условие (1.17)), однако во многих случаях упрощает обработку данных, анализ информационного процесса эмпирического моделирования, представление результатов и их визуализацию. Внешняя аппроксимация гиперпараллелепипедом множества $B(N)$ имеет вид:

$$B^{\Pi} = \{\beta \in R^n \mid \underline{\beta}_j \leq \beta_j \leq \bar{\beta}_j, \quad j = 1, \dots, n\}, \quad (1.22)$$

где
$$\underline{\beta}_j = \min\{\beta_j \mid \beta \in B(N)\}; \quad \bar{\beta}_j = \max\{\beta_j \mid \beta \in B(N)\}. \quad (1.23)$$

Таким образом, при внешней аппроксимации необходимо решить $2n$ экстремальных задач (1.23). Заметим, что замена множества $B(N)$ на множество B^{Π} связана с потерей информации, которая, вообще говоря, зависит от «формы» множества $B(N)$, его ориентации относительно выбранных систем координат и от размерности вектора β .

Оценку качества внешней аппроксимации можно выполнить, проводя внутреннюю аппроксимацию множества $B(N)$. Один из подходов и алгоритмы для ее определения предложен в работе [99] для множества решений переопределенной интервальной системы линейных уравнений. Сравнивая ин-

формационную ценность внутренней аппроксимации по отношению к внешней, можно сделать заключение о допустимости использования внешней аппроксимации множества $B(N)$. Кроме того, при этом исследовании можно выделить активные информационные множества и использовать их при сжатии информации.

В том случае, когда внешняя аппроксимация множества $B(N)$ допустима вне зависимости от числа N информационных совокупностей, достаточно сохранить для описания эмпирической модели n двусторонних неравенств (1.22). При этом, число хранимых информационных порций заметно сокращается.

Сжатие информации

Основная идея сжатия данных описана выше. Пусть $B(N)$ задано и пусть по отношению к этому множеству выделены из B_1, \dots, B_N информационные совокупности B_i , $i \in I_c$, имеющие ненулевую ценность. Оставшиеся порции информации B_i , $i \in \{1, \dots, N\} \setminus I_c$ имеют нулевую ценность. Точное сжатие информации состоит в отбрасывании нулевых по ценности информационных совокупностей.

Приближенное сжатие-замена $B(N)$ на B^H может быть произведено с ухудшением качества эмпирической модели.

Практическая задача состоит в разработке алгоритмов и программ для выполнения только точного сжатия данных. Процедуры приближенного сжатия описаны выше и принципиальных трудностей при их реализации не встречается.

Старение информации

Содержание процесса старения информации состоит в потере ее ценности в зависимости от времени хранения. Введенные выше понятия информационной порции, ее ценности, позволяют формализовать этот процесс.

Прежде всего, понятие старения данных возникает в ситуации, когда новые данные и знания имеют значительную информационную ценность. Эта

ситуация определяет процесс обновления данных, который сопутствует процессу старения информации и является противоположным ему.

В собственном понимании, процесс старения информации связан с изменением свойств объекта описания, для которого ранее полученные порции информации теряют информационную ценность во времени. В рамках рассматриваемого подхода эмпирического моделирования данный процесс можно описать так. Пусть структура функции $y(t) = F(x, \beta(t))$ не зависит от времени, а ее параметры $\beta(t)$ меняются с течением времени. Рассмотрим информационную совокупность B_i в момент времени $t = 0$ – в момент ее формирования. При правильных измерениях имеет место включение (1.17)

$$\beta(0) \in B_i, \quad (1.24)$$

т.е. выполнено двухстороннее неравенство (см. выражение (1.19)):

$$y_i - \bar{\varepsilon}_{y_i} \leq F(x_i, \beta(0)) \leq y_i + \bar{\varepsilon}_{y_i}. \quad (1.25)$$

Предположим, что $F(x, \beta(t))$ удовлетворяет условию Липшица по параметру t с константой L . Тогда для любого времени t_i с момента измерения, справедливо неравенство

$$y_i - \bar{\varepsilon}_{y_i} - L \cdot t_i \leq F(x_i, \beta(t)) \leq y_i + \bar{\varepsilon}_{y_i} + L \cdot t_i. \quad (1.26)$$

Выражение (1.26) задает порцию информации B_i , которая стареет (расширяется) с течением времени t_i формирования i -й порции информации. Если для объекта, свойства которого «дрейфуют» во времени, известен параметр L , то информационная порция $B_i(t_i)$, будучи активной в момент времени $t_i = 0$, постепенно теряет информационную ценность и через некоторый интервал времени будет исключена при сжатии информационной совокупности.

Основной проблемой применения модели старения порций информации (1.26) на практике является оценка параметра L для реальных данных и ее адекватность реальным информационным процессом. Возможны и технические сложности, если константа L зависит от переменных x и β .

Задача планирования эксперимента

В некоторых ситуациях пополнение информационной совокупности B_1, \dots, B_N можно осуществлять путем активного воздействия на моделируемые объекты. В этом случае необходимо найти точку $x_{N+1} \in X$, такую, что ожидаемая ценность информационной порции B_{N+1} по отношению к множеству неопределенности $B(N)$ была бы максимальной. В качестве иллюстрации подхода рассмотрим задачу идентификации параметров (β_1, β_2) функции

$$y = -\beta_1 \ln\left(1 - \frac{x}{\beta_2}\right), \quad 0 \leq x < \beta_2. \quad (1.27)$$

Пусть x_i измеряется точно, а y_i – с ошибкой ε_{y_i} , $i = 1, \dots, N$. Рисунок 1.6 иллюстрирует схему обработки данных в пространстве параметров.

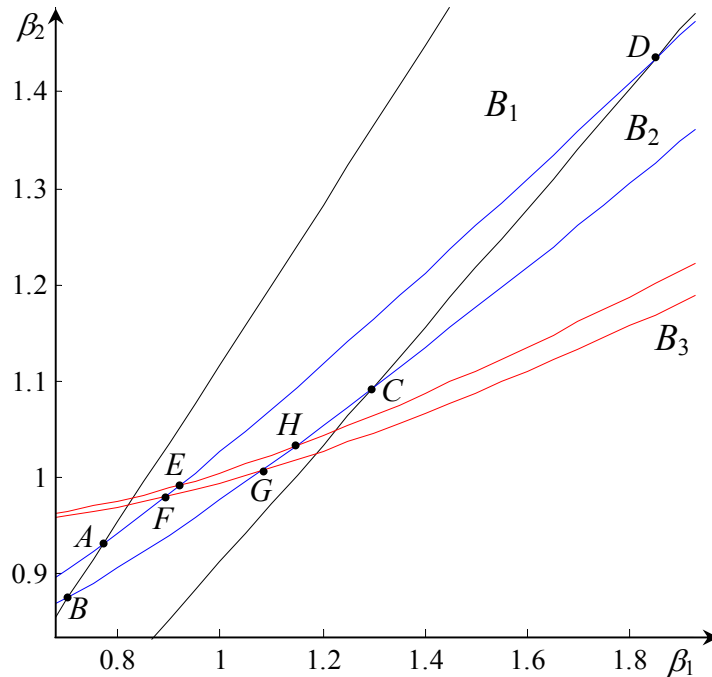


Рис. 1.6. Множества неопределенности параметров β_1 и β_2 зависимости (1.27) при различном выборе точек наблюдения: $ABCD = B_1 \cap B_2$; $EFGH = B_1 \cap B_2 \cap B_3$.

В расчетах было задано $\beta_1 = 1$, $\beta_2 = 1$, $\varepsilon_{y_i} = 0,1$, $i = 1, 2, 3$. Для точек $x = 0,5$, $x = 0,8$ множество неопределенности $B(2) = \{(\beta_1, \beta_2)\}$ приведено на

рисунке 1.6 (область $ABCD$). В силу «близости» полосок B_1 и B_2 область неопределенности является «бананообразной» и точность построения модели не велика. Выбор точки $x = 0,95$ и добавление полоски B_3 резко сокращает множество неопределенности $B(3)$ (область $EFGH$ на рисунке 1.6).

Данный пример показывает необходимость эффективного планирования эксперимента. Вместе с тем, строгая формализация схем планирования одиночных или групповых наблюдений требует дополнительных исследований, в частности, с использованием понятий априорных и апостериорных оценок информационной ценности наблюдений.

Задачи обработки потока данных

Часто при эксплуатации эмпирической модели новые наблюдения поступают непрерывно, «вытесняя» устаревшие. В рассматриваемом подходе эмпирического моделирования каждое новое наблюдение (x_{N+1}, y_{N+1}) приводит к увеличению состава информационной совокупности. При последовательной обработке данных возникают две проблемы. Во-первых, не ясно, можем ли мы «отбросить» указанную порцию информации, особенно в случае старения информации, если ее начальная ценность является нулевой. Основанием для этого сомнения выступает возможность ненулевой ценности порции B_{N+1} в ином составе информационных множеств. Во-вторых, может возникнуть проблема большой размерности множества наблюдений, влекущая дополнительные расход вычислительных ресурсов и рост сложности алгоритмов обработки. Как показано в [40, 118], при определенных условиях второй проблемы можно избежать. Таким образом, в реальных условиях мы без потери информационной ценности можем «забывать» результаты прошлых наблюдений, имеющих нулевую ценность.

Обработка данных, содержащих выбросы

Проблема выявления нарушений в исходных данных относится к числу самых сложных [69], в том числе и в предлагаемом подходе к эмпирическому моделированию. Она неоднократно обсуждалась выше и следует подчерк-

нать ее фундаментальный характер для эмпирического моделирования в целом, когда уровень априорной информации (знаний) об объекте является низким, а каждая точка наблюдений, возможно, содержит ошибки, превышающие их оценки. Мы рассмотрим случай, когда структура модели выбрана верной, оценки ошибок большинства измеренных значений правильны, а некоторое число измерений содержат выбросы. Данный случай является основным в робастной статистике. Описываемый ниже на примере построения линейно параметризованной функции подход способен обнаружить противоречия в исходных предположениях и в общем случае, однако, точно определить причину нарушения в этих случаях вряд ли возможно.

Пусть обработке подлежат информационные совокупности B_1, \dots, B_N и пусть, определенности ради, B_1 сформировано на основе данных, содержащих грубые ошибки. Для линейной по параметрам модели (1.19) запишем систему неравенств

$$y_i - w \cdot \bar{\varepsilon}_{y_i} \leq \sum_{j=1}^n \beta_j \cdot \varphi_j(x_i) \leq y_i + w \cdot \bar{\varepsilon}_{y_i}, i=1, \dots, N, \quad (1.28)$$

в которой множителем $w > 0$ границы информационных множеств можно искусственно сжимать ($w < 1$) или раздвигать ($w > 1$). При $w = 1$ информационные совокупности совпадают с (1.25):

Множество неопределенности B^N зависит от $w > 0$, то есть имеет вид $B(N, w)$. Проверку непротиворечивости исходных данных выполним путем решения задачи

$$w^* = \min \{w \mid B(N, w) \neq \emptyset\}. \quad (1.29)$$

В данном случае мы находим такое w^* , при котором система неравенств находится на границе совместимости, то есть $B(N, w^* - \varepsilon) = \emptyset$ при любом $\varepsilon > 0$. Очевидно, что при $w^* < 1$ мы не имеем оснований для утверждения о противоречивости исходных данных. При $w^* > 1$, мы определенно утверждаем, что исходные данные противоречивы.

В простейшем случае, когда в наборе наблюдений присутствует одиночный выброс, процедура устранения выбросов может состоять в следующем. Удаляем по одному из неравенств-строк в системе (1.28) и решаем задачу (1.29). При удалении из системы неравенств строки, соответствующей выбросу, получим $w^* \leq 1$. На этом процесс исключения выбросов завершается.

В случае наличия нескольких выбросов, исключение наблюдений необходимо проводить с использованием некоторых более сложных критериев. Один из возможных вариантов предлагается в разделе 2.3. Перспективным для выявления выбросов представляется использование теории несобственных задач математического программирования [38].

1.3. Направления дальнейших исследований

Проведенный анализ проблем эмпирического моделирования указывает на определенные трудности в практическом использовании процедур вероятностно-статистического подхода и ограниченность сферы их применимости при построении и анализе зависимостей. В результате формализации процесса эмпирического моделирования при нестатистическом подходе построена принципиально новая логическая схема, для реализации которой необходимо располагать рядом математических инструментов. Детальная разработка и обоснование всей совокупности этих математических моделей и методов, а также разработка способов их использования в практике моделирования представляют собой довольно обширную задачу, полное решение которой вряд ли возможно в рамках одной диссертационной работы.

Настоящая диссертационная работа посвящена разработке в рамках предложенной концептуальной схемы эмпирического моделирования методов обработки интервальных наблюдений. В частности, исследования проводятся в следующих направлениях.

1. Разработка методов обработки интервальных наблюдений в задачах, постановки которых являются традиционными для статистических процедур построения и анализа зависимостей по эмпирическим данным:

- задач дисперсионного анализа;
- задач ковариационного анализа;
- задач учета временных трендов при анализе временных рядов.

2. Разработка методов выявления выбросов в совокупности наблюдений.

3. Разработка методики сравнения и сравнение нестатистических методов оценивания с оценками метода максимального правдоподобия и оценками метода наименьших квадратов в условиях правильных наблюдений.

4. Апробация разработанных методов при решении задач обработки пространственных данных:

- геометрического преобразования и привязки изображений;
- обработки неравноточных инженерно-геодезических измерений.

Выводы по главе 1

1. На практике не всегда удается обеспечить выполнение или проверку условий применимости вероятностно-статистических процедур построения и анализа зависимостей, традиционно используемых для обработки эмпирических данных. Кроме того, принятая при вероятностно-статистическом подходе интерпретация неопределенных факторов как случайных не охватывает всех практически значимых ситуаций, в которых неопределенность может быть также связана с незнанием и неединственностью возможных исходов. Эти причины обуславливают актуальность развития нестатистического подхода к обработке эмпирических данных, система гипотез которого требует не знаний о виде распределения ошибки, а только факта ее ограниченности.

2. Проведено исследование нестатистического подхода к построению и анализу зависимостей, исходя из общей схемы эмпирического моделирования процессов, основанного на принципе «так было – так будет» и предло-

жена логическая схема процесса построения и анализа зависимостей. Указаны подлежащие разработке математические модели и методы, необходимые для реализации основных этапов предложенной схемы.

3. В реализации указанных математических инструментов, необходимых для применения логической схемы процесса эмпирического моделирования, основное внимание в диссертационной работе предложено сосредоточить на разработке:

- нестатистических методов, позволяющих решать задачи построения и анализа зависимостей, содержащих среди входных переменных как количественные, так и качественные факторы;
- методов обработки совокупности наблюдений с выбросами;
- методики сравнения статистических и нестатистических процедур оценивания в условиях правильных наблюдений;
- схем использования предложенных методов при решении задач обработки пространственных данных.

ГЛАВА 2. ПОСТРОЕНИЕ И АНАЛИЗ ЗАВИСИМОСТЕЙ МЕТОДОМ ЦЕНТРА НЕОПРЕДЕЛЕННОСТИ

2.1. Метод центра неопределенности: предпосылки, варианты и основные результаты

Идея исследования области допустимых значений параметров β (множества неопределенности) той или иной модели является довольно естественной и высказывалась многими исследователями не только в связи с решением задачи построения эмпирических зависимостей. Во многих ситуациях множества неопределенности довольно сложны как с точки зрения описания, так и с точки зрения восприятия и интерпретации исследователем. Поэтому довольно часто исследователь заинтересован не в полной характеристике множеств неопределенности, а в их оценивании (аппроксимации) с помощью некоторых простых объектов: гиперпараллелепипедов [17, 27, 28, 88, 99, 100], эллипсоидов [18, 22, 135-137] и т.п.

Спектр подходов к оцениванию возникающих в различных задачах множеств неопределенности и дальнейшего использования этих оценок довольно широк. Так, в 1962 году Л.В. Канторович в работе [51] предлагает в случае построения линейно параметризованной эмпирической зависимости использовать для этой цели методы линейного программирования [39].

Становление интервальной математики [7, 50, 100, 101] позволило многим задачам, неотъемлемой частью постановки которых являются ограниченные неопределенности, обрести свой специфичный язык. В настоящий момент язык интервальной математики представляет собою мощное средство описания такого рода задач в наиболее общей постановке. Арсенал методов решения интервальных задач довольно богат [7, 50, 100, 101, 121-123, 130]. В частности, создан ряд методов внешнего и внутреннего оценивания множеств неопределенности с помощью гиперпараллелепипедов для интервальных систем линейных уравнений, в том числе, и переопределенных [99, 100].

Однако в ряде случаев использование методов интервальной математики оказывается не вполне эффективным. Это связано со следующими обстоятельствами. Многие из задач интервального анализа в общей постановке оказываются принципиально трудно решаемыми [123]. Кроме того, получение оптимальных или хотя бы достаточно «узких» оценок множеств неопределенности во многих ситуациях затруднено из-за специфического для интервальных методов эффекта обертывания (wrapping effect).

Указанные моменты заставляют искать некоторые специфические подходы к решению некоторых интервальных задач. Одним из таких направлений является сведение интервальных задач к оптимизационным задачам с последующим использованием разного рода методов оптимизации [108, 129, 143]. Еще одной причиной, стимулирующей развитие этого направления, является тот факт, что некоторые задачи не требуют постановки и решения в самой общей интервальной форме (например, когда не все величины, фигурирующие в задаче обязаны быть интервальными). Это упрощает задачу и делает еще более оправданным и эффективным использование ее оптимизационной постановки и применение традиционных методов оптимизации и, в частности, методов математического программирования. Поэтому наряду с методами интервальной математики направление, обозначенное Л.В. Канторовичем, получило свое дальнейшее развитие.

Одним из методов, укладывающихся в это русло, является следующий метод, реализующий идеи эмпирического моделирования применительно к задаче определения параметров линейно параметризованной многофакторной зависимости

$$y = \sum_{i=0}^n \beta_i x_i \quad (2.1)$$

по эмпирической информации, главное место в которой занимает таблица экспериментальных данных, полученная в N наблюдениях:

$$T = \{(y_j, x_{0j}, x_{1j}, \dots, x_{nj}) \mid j = 1, \dots, N\}. \quad (2.2)$$

Предполагается, что значения входных переменных x_i известны эксперимен-

татору точно или, иначе говоря, погрешностью их измерения можно пренебречь, а выходная переменная y в j -м наблюдении измеряется с абсолютной погрешностью, оцениваемой по модулю величиной ε_j . Если точное значение выходной переменной в каждом j -м наблюдении обозначить y_j^0 , то имеют место неравенства

$$|y_j^0 - y_j| \leq \varepsilon_j. \quad (2.3)$$

Учитывая (2.1), получаем

$$\left| \sum_{i=0}^n \beta_i x_{ij} - y_j \right| \leq \varepsilon_j, \quad (2.4)$$

или, что то же самое,

$$y_j - \varepsilon_j \leq \sum_{i=0}^n \beta_i x_{ij} \leq y_j + \varepsilon_j. \quad (2.5)$$

Основным объектом дальнейшего анализа является область B допустимых значений параметров β_i , $i = 0, \dots, n$:

$$B = \left\{ \beta = (\beta_0, \dots, \beta_n) \left| y_j - \varepsilon_j \leq \sum_{i=0}^n \beta_i x_{ij} \leq y_j + \varepsilon_j, j = 1, \dots, N \right. \right\}. \quad (2.6)$$

Очевидно, что B – полиэдральное, а, следовательно, и выпуклое множество. При этом B ограничено тогда и только тогда, когда ранг матрицы наблюдений $X = (x_{ij})_{(n+1) \times N}$ равен $n + 1$. Содержательно неограниченность множества B может интерпретироваться как недостаток эмпирической информации. Пустота множества B говорит о противоречивости собранной информации.

Главным принципом нестатистической обработки наблюдений, определяющим все последующие алгоритмы обработки наблюдений и получаемые выводы, является отсутствие предпочтений для элементов множества B (их равная значимость при выборе в качестве оценок параметров).

В виду сложности полного описания множества B в ряде случаев ограничиваются некоторыми его аппроксимациями. В частности, в этой роли можно использовать брусы (гиперпараллелепипеды с гранями, параллельны-

ми координатным плоскостям), охватывающие множество неопределенности B . Наименьший из таких брусов отыскивается путем решения следующих задач линейного программирования:

$$\underline{\beta}_i = \min_{\beta \in B} \beta_i, \quad \bar{\beta}_i = \max_{\beta \in B} \beta_i, \quad i = 0, \dots, n. \quad (2.7)$$

Интервалы $[\underline{\beta}_i, \bar{\beta}_i]$, $i = 0, \dots, n$, определяющие этот брус, содержат в себе возможные точечные оценки параметров β_i , а их длины могут выступать в качестве меры точности точечных оценок.

В соответствии с главным принципом нестатистической обработки наблюдений точечной оценкой параметров β_i зависимости (2.1) в равной степени может служить любой из элементов множества B . В частности, одним из наиболее простых способов построения точечной оценки $\hat{\beta} = \left(\hat{\beta}_0, \dots, \hat{\beta}_n \right)$ является выбор в этом качестве срединной точки охватывающего бруса, отыскиваемого при решении задач (2.7):

$$\hat{\beta}_i = \frac{1}{2}(\underline{\beta}_i + \bar{\beta}_i), \quad \Delta\beta_i = \frac{\bar{\beta}_i - \underline{\beta}_i}{2}, \quad i = 0, \dots, n. \quad (2.8)$$

Помимо задачи точечного и интервального оценивания параметров зависимости (2.1) в отношении множества B может ставиться и задача интервального оценивания выходной переменной y зависимости (2.1) в точке x :

$$\underline{y}(x) = \min_{\beta \in B} \beta x, \quad \bar{y}(x) = \max_{\beta \in B} \beta x. \quad (2.9)$$

Интервал $[\underline{y}(x), \bar{y}(x)]$ содержит возможные значения выходной переменной y в точке x при различном выборе параметров зависимости. В качестве точечной оценки прогнозного значения зависимости (2.1) в точке x по аналогии с (2.8) может использоваться полусумма концов интервала:

$$\hat{y}(x) = \frac{1}{2}(\underline{y}(x) + \bar{y}(x)), \quad \Delta y(x) = \frac{1}{2}(\bar{y}(x) - \underline{y}(x)). \quad (2.10)$$

Приведенные постановка и методы решения задач интервального и точечного оценивания параметров линейно параметризованной зависимости и интервальной и точечной задач прогноза значений выходной переменной вы-

текают из идеи Л.В. Канторовича и были независимо предложены в разных формах сразу несколькими исследователями: Спиваком с соавторами в 1970 г., Миланезе и Бельфорте 1982 г., Оскорбиным в 1983 г, Воцининым и Сотировым в 1989 г. Базовые построения и выводы у всех названных авторов совпадают в своих ключевых идеях с точностью до терминологии.

Так С.И. Спивак с соавторами [23, 49, 53, 81-86] в рамках идеологии Канторовича при построении точечных оценок параметров используют метод выравнивания по Чебышеву, который сводится к решению экстремальной задачи

$$\hat{\beta} = \arg \min_{\beta \in B} \max_{j=1, \dots, N} \left| y_j - \sum_{i=0}^n \beta_i x_{ji} \right|. \quad (2.11)$$

Точка $\hat{\beta}$, являющаяся решением этой задачи, называется чебышевским центром множества B . Наряду с (2.8) поиск чебышевского центра множества неопределенности дает еще один способ построения точечной оценки параметров линейно параметризованной зависимости. В [23] рассматривается задача чебышевского выравнивания для модели с дробно-линейной зависимостью от параметров. Заметим, что на идее оптимального выбора оценок или управляющих воздействий в наихудшем случае основаны такие близкие по духу к нестатистическим методам оценивания подходы к обработке информации как минимаксная обработка информации [59], гарантированное оценивание [57, 58, 102, 103] и грубая идентификация [87].

В работе [126] М. Миланезе и Г. Бельфорте указывают на ряд сложностей в практическом применении метода максимума правдоподобия и указывают на возможность использования нестатистического подхода, позволяющего получить интервалы неопределенности параметров (2.7) и интервалы неопределенности оценок (2.9) в случае ограниченной ошибки неизвестной природы. В [109] предлагаются рекуррентные методы вычисления оценок (2.8) и (2.11), которые сравниваются по сходимости с рекуррентными оценками наименьших квадратов и рекуррентным методом Фогеля-Хуанга [118].

В работе Н.М. Оскорбина [75] наряду с понятием множества неопределенности вводится и термин «центр неопределенности» для обозначения выбираемой согласно (2.8) и (2.10) точечных и интервальной оценок параметров. В дальнейшем вся совокупность приемов построения и анализа эмпирических зависимостей в рамках идеологии Канторовича была названа Оскорбиным «метод центра неопределенности». В 1987 г. в Алтайском госуниверситете Н.М. Оскорбиным и А.В. Максимовым разработана программа MCN, реализующая метод центра неопределенности. Программа позволяет для зависимости вида (2.1) отыскивать интервальные оценки параметров (2.7), строить интервальный прогноз значения выходной переменной (2.9) при известных значениях входных переменных, получать оценки «ценности» каждого из наблюдений, основанные на решении двойственной задачи линейного программирования и, наконец, отыскивать центр неопределенности (строить точечную оценку параметров) следующим образом.

В описание множества неопределенности B вводится параметр w , позволяющий изменять «размер» множества неопределенности:

$$B_w = \left\{ \beta = (\beta_0, \dots, \beta_n) \left| y_j - w\varepsilon_j \leq \sum_{i=0}^n \beta_i x_{ij} \leq y_j + w\varepsilon_j, j = 1, \dots, N \right. \right\}. \quad (2.12)$$

При $w = 1$ множество B_w совпадает с исходным множеством неопределенности. Увеличение w соответствует растяжению множества неопределенности, а уменьшение параметра w влечет сжатие множества неопределенности. Среди возможных значений параметра w отыскивается наименьшее, при котором B_w остается непустым, то есть

$$w_0 = \operatorname{argmin} \{w \mid B_w \neq \emptyset\}. \quad (2.13)$$

Оптимальное значение w_0 отыскивается методом деления отрезка пополам, при этом непустота множества B_w проверяется методами линейного программирования.

Построение точечной оценки $\hat{\beta} = \left(\hat{\beta}_0, \dots, \hat{\beta}_n \right)$ параметров зависимости

предлагается производить, используя формулу (2.8) для множества B_{w_0} :

$$\hat{\beta}_i = \frac{1}{2} \left(\min_{\beta \in B_{w_0}} \beta_i + \max_{\beta \in B_{w_0}} \beta_i \right). \quad (2.14)$$

Среди прочих работ Оскорбина и его соавторов отметим работу [40], в которой рассматривается задача рекуррентного оценивания множества неопределенности в случае обработки больших массивов эмпирической информации. Указанные подходы и программные средства использовались при выполнении прикладных исследований [5].

Работы А.П. Воцинина, А.Ф. Бочкова и Г.Р. Сотирова [30, 31] содержат описание метода анализа данных при интервальной нестатистической ошибке, основанного на построении интервальных оценок параметров линейно параметризованной зависимости (2.7), интервальных оценок выходной переменной (2.9), построении минимаксной точечной оценки параметров (2.11) и точечной оценки, наиболее точной в среднем:

$$\hat{\beta} = \min_{\beta \in B} \frac{1}{V_B} \sum_{i=0}^{V_B} \|\beta - b_i\|, \quad (2.15)$$

где b_i , $i = 0, \dots, V_B$ – вершины множества B . Работа [30] знаменательна тем, что редакция журнала «Заводская лаборатория» провела на ее основе круглый стол, и в полном объеме опубликовала результаты обсуждения [25, 34, 56, 60, 61, 67], содержащие некоторые ценные критические замечания.

Среди прочих результатов, связанных с методом центра неопределенности, выделим работы В.А. Суханова с соавторами [17-19, 21-22, 88] и О.Е. Родионовой и А.Е. Померанцева [133, 134, 138-140].

Значительная часть работ Суханова с соавторами посвящена аппроксимации множества неопределенности эллипсоидом [18, 22] и, в том числе, разработке рекуррентных алгоритмов построения эллипсоида. Более поздние работы [88] этой группы авторов касались исследования теоретико-вероятностных свойств оценок метода центра неопределенности и, в частности, оценок (2.7)-(2.8), и являются своеобразным ответом на критические замечания, сформулированные в полемических заметках А.И. Хлебникова [94, 95].

В основу метода, предлагаемого в работах О.Е. Родионовой и А.Е. Померанцева [140] под названием «простые интервальные вычисления» (simple interval calculation, SIC), положен способ построения интервального прогноза линейной зависимости, описываемый формулой (2.9). Простые интервальные вычисления предлагается использовать для решения задачи калибровки многофакторной линейной модели и прогнозирования выходной переменной по новым измерениям совместно с проекционными методами многофакторного анализа такими, как регрессия на главные компоненты, регрессия на латентные структуры и т.п. [1, 116, 125]. В частности, в случае неограниченности множества неопределенности, то есть когда ранг матрицы наблюдений $X = (x_{ij})_{n \times N}$ меньше n , модельную зависимость $y = X\beta$ предлагается заменять проекционной моделью $y = Tq$, где T – матрица счетов, а q – вектор нагрузок (см. [125]), конструируемые методом главных компонент.

Кроме того, в работах О.Е. Родионовой и А.Е. Померанцева поднимается вопрос о выборе верхней границы ошибок наблюдений ε_{\max} при отсутствии информации о ней и предлагается несколько эмпирических правил выбора ε_{\max} , основанных на разного рода показателях качества предсказания (ширина прогнозного интервала, соотношение предсказанного интервала и интервала, соответствующего тестовым калибровочным образцам и т.п.).

Итак, обобщая результаты разных авторов, в качестве ключевых (характеристических) идей, лежащих в основе рассмотренных нестатистических методов построения и анализа зависимостей, выделим следующие:

1. Гипотеза о точном (с пренебрежимо малой ошибкой) измерении значений входных переменных в модели исследуемого объекта.
2. Гипотеза об ограниченности по модулю ошибок измерения выходной переменной (невязок модели).
3. Аппроксимация множества неопределенности (области допустимых значений) параметров модели.

4. Построение интервальных и точечных оценок параметров модели как некоторых характеристик сконструированной аппроксимации множества неопределенности.

В рамках «каркаса», определяемого этими идеями некоторые элементы нестатистических методов выступают в роли «параметров», то есть их «наполнение» может варьироваться. В частности, в качестве вида конструируемой зависимости можно задаваться не только линейно параметризованными функциями, но и прочими классами функций. Исходя из особенностей условий моделирования, могут выбираться различные способы аппроксимации множества неопределенности: гиперпрямоугольник, эллипсоид, и т.п. Выбранные вид зависимости и способ аппроксимации множества неопределенности определяют и математический аппарат, используемый для построения аппроксимации и конструирования точечных и интервальных оценок параметров модели: линейное программирование, нелинейное программирование, методы интервальной математики и т.д.

Отметим, что границы термина «метод центра неопределенности» до сих пор в литературе остаются довольно размытыми. Поэтому в целях конкретизации смысла, вкладываемого далее в указанный термин, договоримся обозначать им метод, опирающийся на сформулированные ключевые идеи нестатистических методов построения и анализа зависимостей, и в котором

а) вид конструируемой зависимости не ограничивается классом линейно параметризованных функций;

б) внешняя аппроксимация множества неопределенности производится гиперпараллелепипедом;

в) интервальные и точечные оценки параметров зависимости отыскиваются согласно (2.7) и (2.8) или (2.13) с использованием подходящих методов оптимизации;

г) интервальные и точечные оценки значения выходной переменной зависимости отыскиваются согласно (2.9) и (2.10) с использованием подходящих методов оптимизации.

2.2 Обработка информации методом центра неопределенности при правильных наблюдениях

В настоящем разделе рассмотрим, каким образом с помощью метода центра неопределенности могут решаться некоторые задачи, связанные с построением и анализом эмпирической зависимости, традиционно решаемые статистическими методами. При этом на протяжении всего раздела предполагается, что наблюдения, на основе которых выполняются построение зависимости, правильны, то есть не содержат выбросов. Методы обработки информации с выбросами рассматриваются в разделе 2.3.

Для решения задач построения и анализа эмпирических зависимостей, содержащих среди входных переменных только количественные факторы, только качественные факторы или одновременно факторы той и другой природы традиционно используют такие статистические инструменты, как регрессионный, дисперсионный и ковариационный анализы соответственно [37]. Метод центра неопределенности применим для решения подобного рода задач, если конструируемым моделям придать специальный вид.

2.2.1. Задачи регрессионного анализа

Задачи, решаемые с использованием регрессионного анализа, состоят в построении эмпирической зависимости априорно определенного вида (далее ограничимся лишь случаем линейно-параметризованной зависимости (2.1)) по совокупности эмпирической информации, главное место в которой занимает таблица наблюдений за значением выходной переменной при определенных значениях входной переменной (2.2). Система гипотез о стохастической структуре модели классического регрессионного анализа подробно обсуждалась в разделе 1.1. Система гипотез метода центра неопределенности в этом случае является частной формой (1.10)-(1.14) и состоит из предположений 1) о структуре модели, 2) об ограниченности ошибки измерения выходной переменной (2.3) и 3) об адекватности любой зависимости, описываемой

функцией заданного вида и проходящей через все интервальные измерения выходной переменной.

Основные идеи решения поставленной задачи с помощью метода центра неопределенности, а именно

- построение на основе таблицы данных (2.2) множества неопределенности (2.6):
- построение интервальных оценок параметров (2.7);
- построение точечных оценок (2.8), (2.11), (2.14) или (2.15);
- построение интервальных (2.9) и точечных (2.11) прогнозных значений зависимости в заданной точке;

описаны в разделе 2.1, поэтому ограничимся примером.

Данные для примера (см. таблицу 2.1) взяты из [37, с. 30]. В таблице приведены 25 наблюдений за работой выпарного аппарата на промышленном предприятии, а именно, значения переменных:

- количество используемого пара ежемесячно в фунтах (y);
- средняя температура в градусах Фаренгейта (x).

Необходимой информацией, используемой методом центра неопределенности, являются еще и сведения о границах ошибки каждого из измерений, содержащегося в таблице эмпирических данных. Такая информация отсутствует в [37]. Вопрос выбора границ ошибки измерения ε выходной переменной y в случае отсутствия такой информации в явном виде заслуживает отдельного обсуждения вне рамок данного примера. Однако, в нашем конкретном случае, мы не повредим достижению основной цели, преследуемой настоящим примером – проиллюстрировать схему решения задач регрессионного анализа при правильных наблюдениях, если положим верхнюю границу абсолютного значения ошибки измерения выходной переменной удвоенному стандартному отклонению ее табличных значений от линии МНК-регрессии, построенной в [37]: $\bar{\varepsilon} = 1,74$. Данные из таблицы 2.1 с границами ошибки измерения выходной переменной в графическом виде представлены

на рисунке 2.1.

Таблица 2.1

Данные о выпарном аппарате
(количество используемого пара ежемесячно в фунтах» (y),
средняя температура в градусах Фаренгейта (x))

Номер опыта	y	x	Номер опыта	y	x
1	10,98	35,3	14	9,57	39,1
2	11,13	29,7	15	10,94	46,8
3	12,51	30,8	16	9,58	48,5
4	8,40	58,8	17	10,09	59,3
5	9,27	61,4	18	8,11	70,0
6	8,73	71,3	19	6,83	70,0
7	6,36	74,4	20	8,88	74,5
8	8,50	76,7	21	7,68	72,1
9	7,82	70,7	22	8,47	58,1
10	9,14	57,5	23	8,86	44,6
11	8,24	46,4	24	10,36	33,4
12	12,19	28,9	25	11,08	28,6
13	11,88	28,1			

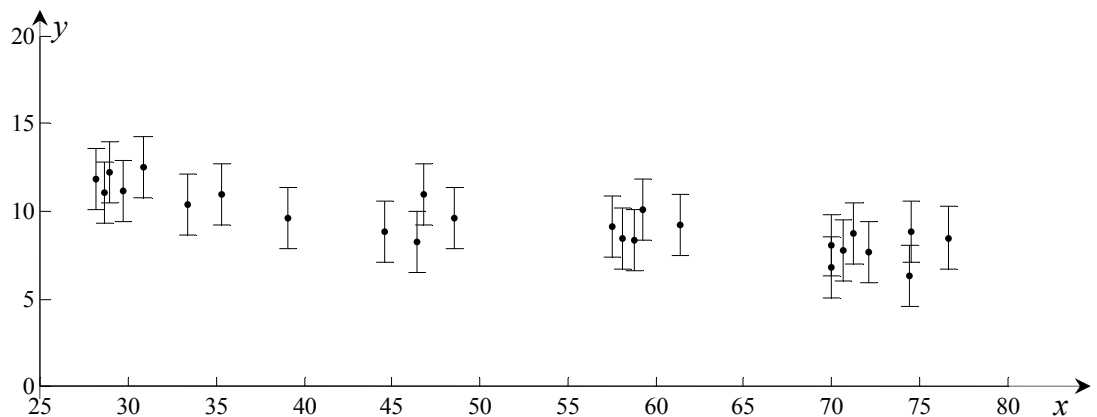


Рис. 2.1. Диаграмма рассеяния данных из таблицы 2.1.

В качестве гипотетической структуры конструируемой зависимости между входной переменной x и выходной переменной y может быть принята линейно параметризованная:

$$y = \beta_0 + \beta_1 x + \varepsilon. \quad (2.16)$$

При наличии каких-то априорных сведений их можно использовать для формирования стартовой информационной совокупности. Например, информация о монотонном возрастании или убывании зависимости позволяет поместить в стартовую информационную совокупность неравенства $\beta_1 > 0$ или

$\beta_1 < 0$ соответственно. В нашем случае такого рода сведениями мы не располагаем.

Основную информацию в себе несет таблица наблюдений 2.1 и факт ограниченности по модулю ошибки измерения выходной переменной y величиной $\bar{\varepsilon}$, которые позволяют добавить к множеству неопределенности ограничения вида

$$y_j - \bar{\varepsilon} \leq \beta_0 + \beta_1 x_j \leq y_j + \bar{\varepsilon}, \quad j = 1, \dots, 25. \quad (2.17)$$

Решая задачи линейного программирования (2.7) при ограничениях (2.17), получим интервальные оценки параметров зависимости β_0, β_1 :

$$\hat{\beta}_0 \in [12,65; 14,68], \quad \hat{\beta}_1 \in [-0,10; -0,06].$$

В качестве точечной оценки параметров может быть принята одна из оценок, приведенных в таблице 2.2.

Таблица 2.2

Точечные оценки параметров зависимости (2.16)

Вид оценки	$\hat{\beta}_0$	$\hat{\beta}_1$
Центр прямоугольника (2.8)	13,665	-0,081
Чебышевский центр (2.11)	13,665	-0,081
Центр Оскорбина (2.14)	13,581	-0,083
Наиболее точная в среднем (2.15)	13,155	-0,071
Центр тяжести множества неопределенности	13,439	-0,078

В нашем случае все пять оценок довольно близки между собой и близки к МНК-оценке ($\hat{\beta}_0 = 13,623$, $\hat{\beta}_1 = -0,080$). Однако, в общем случае значения точечных оценок могут заметно различаться, а МНК-оценка часто не принадлежит множеству неопределенности. Близость точечных оценок в нашем случае обусловлена довольно простой структурой множества неопределенности (см. рисунок 2.2), которое определяется всего четырьмя активными наблюдениями: 3, 7, 11 и 20-м. О степени активности (ценности) тех или иных наблюдений позволяют судить соответствующие им значения двойственных оценок, получаемые при решении задач линейного программирования. При рекуррентном оценивании знание списка активных (с ненулевой ценностью) наблюдений позволяет проводить сжатие информационной

совокупности, описывающей множество неопределенности, путем отбрасывания неактивных наблюдений. В нашем случае сжатая таблица наблюдений содержала бы только четыре строки с упомянутыми номерами.

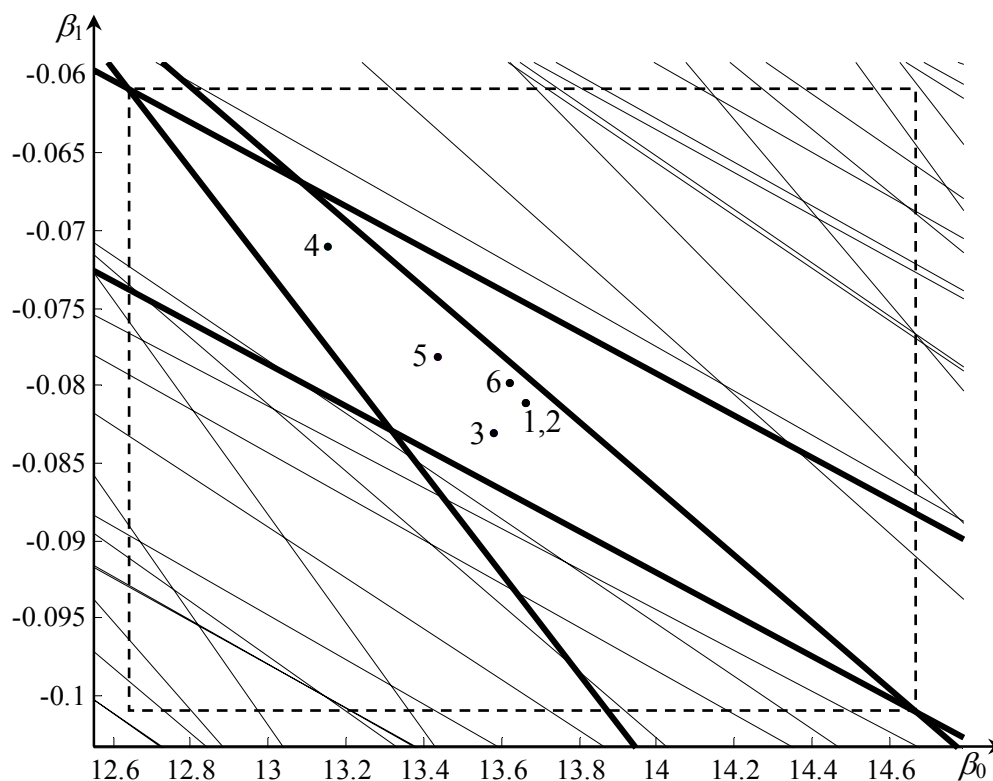


Рис. 2.2. Множество неопределенности и точечные оценки параметров зависимости (2.16) в пространстве параметров (β_0, β_1) : 1 – центр прямоугольника, 2 – чебышевский центр, 3 – центр Оскорбина, 4 – оценка, наиболее точная в среднем, 5 – центр тяжести, 6 – МНК-оценка.

Значительный интерес представляет проверка гипотезы о том, что коэффициент β_i равен нулю, или, другими словами, о значимости вклада переменной, соответствующей этому коэффициенту, в общее уравнение. Реже, но, тем не менее, в ряде случаев нас интересует проверка гипотезы о том, что угловой коэффициент равен некоторой константе, отличной от нуля. Такого рода гипотезы могут непосредственно проверяться путем добавления к системе ограничений, описывающих множество неопределенности, ограничений вида $\beta_i = b_i$ и выяснения совместности вновь полученной системы ограничений. Более того, таким же образом могут проверяться гипотезы о равенстве некоторой линейной комбинации параметров некоторой константе.

Одна из основных причин построения зависимости заключается в возможности использования ее для прогноза. Вычисление интервальной оценки прогноза выходной переменной y в заданной точке x сводится к решению двух задач линейного программирования (2.9). Интервалы прогнозных значений зависимости $[\underline{y}(x), \bar{y}(x)]$ для каждого из значений входной переменной x_i , содержащихся в таблице 2.1, приведены в таблице 2.3, а на рисунке 2.3 показаны их верхняя и нижняя огибающие, являющиеся, по сути, границами множества неопределенности в пространстве (x, y) .

Таблица 2.3

Интервалы прогнозных значений зависимости
в экспериментальных точках

Номер опыта	\underline{y}	\bar{y}	y	x	Номер опыта	\underline{y}	\bar{y}	y	x
1	10,39	11,11	10,98	35,3	14	10,08	10,73	9,57	39,1
2	10,83	11,67	11,13	29,7	15	9,44	9,96	10,94	46,8
3	10,77	11,56	12,51	30,8	16	9,30	9,84	9,58	48,5
4	8,44	9,15	8,40	58,8	17	8,40	9,12	10,09	59,3
5	8,23	8,98	9,27	61,4	18	7,51	8,40	8,11	70,0
6	7,40	8,31	8,73	71,3	19	7,51	8,40	6,83	70,0
7	7,15	8,10	6,36	74,4	20	7,14	8,10	8,88	74,5
8	6,91	7,96	8,50	76,7	21	7,34	8,26	7,68	72,1
9	7,45	8,35	7,82	70,7	22	8,50	9,20	8,47	58,1
10	8,55	9,24	9,14	57,5	23	9,62	10,17	8,86	44,6
11	9,47	9,98	8,24	46,4	24	10,55	11,30	10,36	33,4
12	10,88	11,75	12,19	28,9	25	10,90	11,79	11,08	28,6
13	10,93	11,84	11,88	28,1					

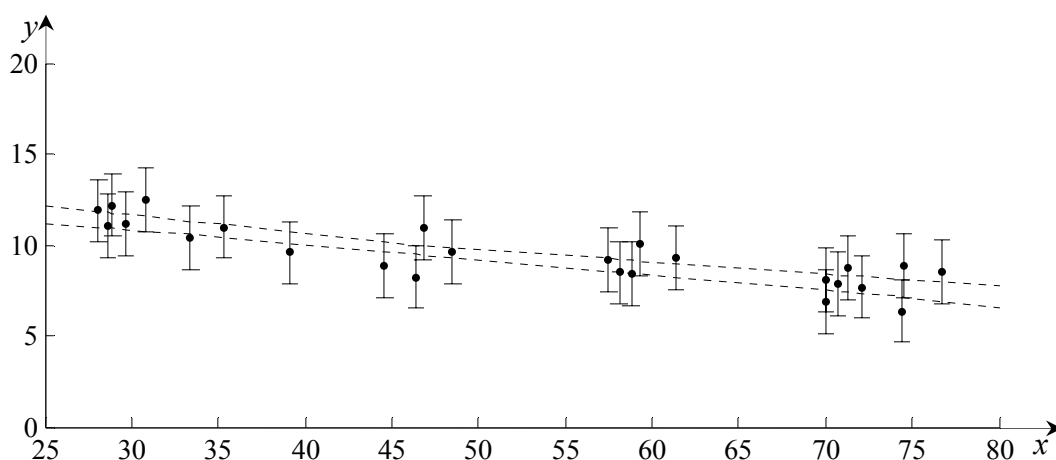


Рис. 2.3. Верхняя и нижняя огибающие интервалов прогнозных значений (показаны пунктирными линиями).

Продемонстрированный подход к решению задач регрессионного анализа с помощью метода центра неопределенности совместно с приемом преобразования входных переменных и использованием фиктивных переменных позволяет строить и более сложные модели. В частности, возможно построение и анализ зависимостей при наличии в списке входных переменных качественных факторов, то есть решение задач дисперсионного и ковариационного анализов.

2.2.2. Задачи дисперсионного и ковариационного анализов

Основной прием, позволяющий при построении зависимости ввести в рассмотрение качественные факторы, состоит в использовании фиктивных переменных. В классическом статистическом анализе хорошо известны [36, 37] способы применения этого аппарата для сведения задач дисперсионного анализа (все факторы качественные) и ковариационного анализа (часть факторов – количественные, а часть – качественные) к задаче регрессионного анализа. В настоящем разделе будет показано, что использование того же приема при нестатистическом подходе делает возможным решение задач дисперсионного и ковариационного анализов с помощью метода центра неопределенности.

Для учета влияния на значение выходной переменной каждого из качественных факторов x_i , принимающих значения на L_i уровнях $X_i = \{x_{i0}, \dots, x_{iL_i-1}\}$, в зависимость вводятся $L_i - 1$ фиктивных переменных, значения которых в совокупности кодируют уровень фактора x_i , соответствующий каждому из наблюдений. Способ выбора фиктивных переменных не единственен, но одним из наиболее простых для реализации и интерпретации является следующий вариант сопоставления уровней фактора и значений совокупности фиктивных переменных.

Один из уровней фактора выбирается в качестве эталонного, например, x_{i0} , а для остальных определяются фиктивные переменные $d_{i1}, \dots, d_{i(L_i-1)}$, при-

нимающие значения 0 или 1. Ситуация, когда все переменные $d_{i1}, \dots, d_{i(L_i-1)}$ равны нулю, соответствует эталонному уровню фактора x_{i0} . Равенство единице переменной d_{ik} при нулевых значениях остальных фиктивных переменных соответствует уровню фактора x_{ik} .

Коэффициент δ_{ik} при каждой из заданных таким способом фиктивных переменных d_{ik} представляет собой оценку так называемого «чистого» эффекта, то есть разницы в значении выходной переменной, обусловленной переходом фактора x_i с эталонного уровня x_{i0} на уровень x_{ik} при фиксированных значениях прочих переменных, входящих в зависимость.

После пополнения фиктивными переменными структура зависимости приобретает вид

$$y = \sum_{i=0}^{m-1} \beta_i x_i + \sum_{i=m}^n \sum_{k=1}^{L_i-1} \delta_{ik} d_{ik}, \quad (2.18)$$

где входные переменные x_0, \dots, x_{m-1} являются количественными факторами, а качественные факторы x_m, \dots, x_n представлены группами фиктивных переменных $d_{i1}, \dots, d_{iL_i-1}$, $i = m, \dots, n$. При $m = 0$ задача построения и анализа зависимости вида (2.18) соответствует задаче дисперсионного анализа, а при $m > 0$ – задаче ковариационного анализа. Для оценивания коэффициентов β_i и δ_{ik} используются методы, изложенные в предыдущем разделе.

Рассмотрим пример, данные для которого (таблица 2.4) взяты из [37, с. 301] и представляют собой вес (y) в фунтах и возраст в неделях для тринадцати индеек. Четыре из этих индеек выращены в штате Джорджия, четыре – в Виргинии и пять – в Висконсине. Попытаемся связать вес и возраст птицы простой линейной зависимостью и выяснить, какое влияние на зависимость оказывает место ее происхождения. Для учета влияния этого качественного фактора, принимающего значение на трех уровнях, введем две фиктивные переменные d_1 и d_2 , определив их значения, как показано в табли-

це 2.4. Конструируемая зависимость имеет вид:

$$y = \beta_0 + \beta_1 x + \delta_1 d_1 + \delta_2 d_2 + \varepsilon. \quad (2.19)$$

Отсутствующую информацию о верхней границе абсолютного значения ошибки ($\bar{\varepsilon}$) измерения выходной переменной восполним, положив ее равной 1 фунту. Множество неопределенности в нашей задаче определяется неравенствами вида

$$y_j - \bar{\varepsilon} \leq \beta_0 + \beta_1 x_j + \delta_1 d_{1j} + \delta_2 d_{2j} \leq y_j + \bar{\varepsilon}, \quad j = 1, \dots, 13, \quad (2.20)$$

где $(y_j, x_j, d_{1j}, d_{2j})$ – данные из j -й строки таблицы 2.4.

Таблица 2.4

Данные об индейках

(вес в фунтах (y), возраст в неделях (x), место происхождения)

Номер опыта	y	x	Происхождение	d_1	d_2
1	13,3	28	Джорджия	1	0
2	8,9	20	Джорджия	1	0
3	15,1	32	Джорджия	1	0
4	10,4	22	Джорджия	1	0
5	13,1	29	Виргиния	0	1
6	12,4	27	Виргиния	0	1
7	13,2	28	Виргиния	0	1
8	11,8	26	Виргиния	0	1
9	11,5	21	Висконсин	0	0
10	14,2	27	Висконсин	0	0
11	15,4	29	Висконсин	0	0
12	13,1	23	Висконсин	0	0
13	13,8	25	Висконсин	0	0

Используя те же процедуры, что и при решении задачи регрессионного анализа, получаем интервальные оценки параметров зависимости:

$$\hat{\beta}_0 \in [1,750; 5,570], \quad \hat{\beta}_1 \in [0,350; 0,450],$$

$$\hat{\delta}_1 \in [-3,350; -0,850], \quad \hat{\delta}_2 \in [-2,600; -0,450].$$

В качестве точечной оценки примем наиболее просто вычисляемый центр прямоугольника:

$$\hat{\beta}_0 = 3,750, \quad \hat{\beta}_1 = 0,400, \quad \hat{\delta}_1 = -2,100, \quad \hat{\delta}_2 = -1,525.$$

Однозначно отрицательные интервальные оценки коэффициентов δ_1 , δ_2 при фиктивных переменных указывают на различия в индексах, первая – из Джорджии и Висконсина, а вторая – из Виргинии и Висконсина соответственно. Подставляя три различных набора значений фиктивных переменных (d_1, d_2) и используя точечные оценки параметров, получим зависимости для трех разных штатов:

$$\text{для Джорджии при } d_1 = 1, d_2 = 0: \quad \hat{y} = 1,650 + 0,400x;$$

$$\text{для Виргинии при } d_1 = 0, d_2 = 1: \quad \hat{y} = 2,225 + 0,400x;$$

$$\text{для Висконсина при } d_1 = 0, d_2 = 0: \quad \hat{y} = 3,750 + 0,400x.$$

Полученные результаты не противоречат результатам обработки этих данных методами регрессионного анализа [37].

2.2.3. Временные тренды в данных

Прием использования фиктивных переменных оказывается полезным и при решении задач, связанных с обработкой временных рядов в рамках нестатистического подхода. Во многих практических случаях в выходной переменной проявляется временной тренд. Иногда тренд представляет собой единственный фактор, а иногда он налагается на эффекты других входных переменных. Мы можем описать временной тренд, добавив в структуру зависимости один или несколько подходящим образом определенных фиктивных факторов, а затем оценивая параметры зависимости аналогично тому, как показано в предыдущем примере.

Когда в данных представлен простой линейный тренд, для его учета достаточно ввести одну фиктивную переменную d . Если данные получены в результате наблюдений через равные промежутки времени, то значения переменной d , соответствующие каждой строке таблицы наблюдений, могут определяться как узлы регулярной сетки с целочисленными узлами $d_i = i$, $i = 1, \dots, N$, где N – количество строк таблицы. Иногда, в целях улучшения вычислительной структуры матрицы наблюдений, значения d можно цен-

трировать, то есть использовать фиктивную переменную $d' = d - \bar{d}$, где \bar{d} – среднее значение. Зависимость в этих случаях выражается уравнениями:

$$y = \delta_0 + \delta_1 d + \sum_{i=1}^n \beta_i x_i + \varepsilon \quad (2.21)$$

и
$$y = (\beta_0 + \delta_1 \bar{d}) + \delta_1 d' + \sum_{i=1}^n \beta_i x_i + \varepsilon = \beta_0' + \delta_1 d' + \sum_{i=1}^n \beta_i x_i + \varepsilon. \quad (2.22)$$

Если данные не равномерны по времени, то значения фиктивной переменной выбираются в узлах сетки (желательно целочисленной), расстояние между узлами которой пропорционально временным промежуткам, через которые проводились наблюдения.

В случае квадратичного временного тренда к зависимости следует добавлять члены $y = \delta_0 + \delta_1 d + \delta_2 d^2$ или $y = \delta_0' + \delta_1' d' + \delta_2 d'^2$. Используя члены более высоких порядков, таким же образом можно учитывать и тренды более высоких порядков.

Когда в данных представлены два временных тренда, то в зависимость добавляются фиктивные переменные для каждого из них. Уровень сложности этой задачи зависит от того, известно ли какому тренду принадлежат данные или нет. При отсутствии информации о принадлежности данных областям того или иного трендов, приходится просматривать все возможные варианты разбиения точек данных между трендами, подбирая наилучшую в некотором смысле модель.

В ситуации, когда известно, какие данные принадлежат каким трендам, можно выделить два подтипа задач: (1) когда имеется информация о том, что абсцисса точки «стыковки» тренда совпадает с одной из точек наблюдения, и (2) когда абсцисса точки пересечения линий трендов неизвестна.

Проиллюстрируем на примерах, как решаются задачи каждого из двух подтипов в случае двух линейных трендов.

Данные, приведенные в таблице 2.5 (взяты из [37, с.310]), относятся к первому подтипу. Известно, что с точностью $\bar{\varepsilon} = 0,5$ первые пять точек лежат

на первой прямой, а последние пять точек – на второй. Таким образом, пятая точка является общей для обеих прямых. Пара фиктивных переменных d_1 , d_2 для двух этих прямых вводятся так, как указано в таблице 2.5. Если в предположении, что нет других входных переменных, подобрать зависимость

$$y = \delta_0 + \delta_1 d_1 + \delta_2 d_2 + \varepsilon, \quad (2.23)$$

то полученные оценки будут иметь следующий смысл:

$\hat{\delta}_0$ – значение \hat{y} в точке пересечения (при $d_1 = d_2 = 0$),

$\hat{\delta}_1$ – угловой коэффициент прямой первого тренда,

$\hat{\delta}_2$ – угловой коэффициент прямой второго тренда.

Интервальные оценки этих коэффициентов:

$$\hat{\delta}_0 \in [9,700; 9,800], \quad \hat{\delta}_1 \in [1,800; 1,900], \quad \hat{\delta}_2 \in [0,967; 1,100].$$

Точечные оценки центра прямоугольника:

$$\hat{\delta}_0 = 9,750, \quad \hat{\delta}_1 = 1,850, \quad \hat{\delta}_2 = 1,033.$$

Графики прямых

$$\hat{y} = 9,750 + 1,850d_1 \quad \text{и} \quad \hat{y} = 9,750 + 1,033d_2,$$

описывающие линейные тренды приведены на рисунке 2.4.

Таблица 2.5

Данные для примера с двумя прямыми,
абсцисса пересечения которых известна

Номер опыта	Дата	y	d_1	d_2
1	1970 г.	2,3	-4	0
2		3,8	-3	0
3	1971 г.	6,5	-2	0
4		7,4	-1	0
5	1972 г.	10,2	0	0
6		10,5	0	1
7	1973 г.	12,1	0	2
8		13,2	0	3
9	1974 г.	13,6	0	4

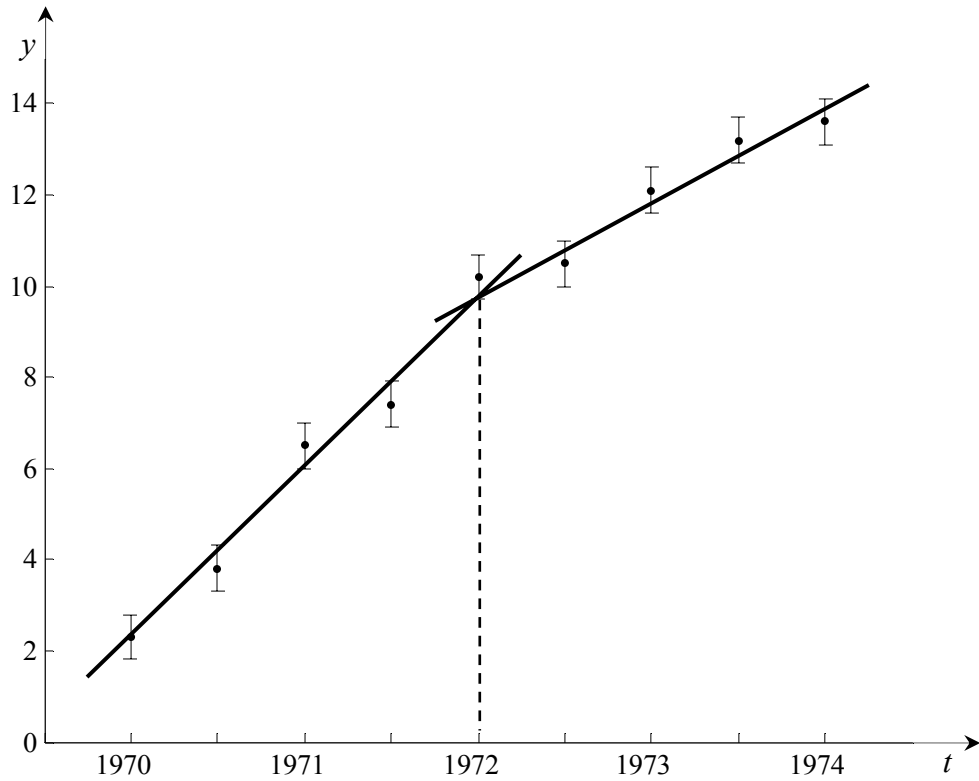


Рис. 2.4. Графики двух линейных трендов, стыкующиеся в точке с известной абсциссой.

Данные, приведенные в таблице 2.6 (взяты из [37, с.312]), относятся ко второму подтипу. Здесь известно, что первые четыре точки лежат на одной прямой, а остальные пять – на другой с точностью $\bar{\varepsilon} = 0,5$. Однако точка пересечения не известна. Чтобы обнаружить неизвестную точку, необходимо ввести третью фиктивную переменную d_3 . Ее полагают равной нулю для всех точек первой прямой и единице для точек второй прямой, чтобы отразить скачок от первой прямой ко второй. Значения фиктивных переменных d_1 , d_2 выбираются несколько иначе, чем в предыдущем примере, что иллюстрирует неединственность способов представления фиктивных переменных. Здесь значения d_1 сдвинуты на пять единиц вправо относительно значений этой переменной в предыдущем примере. Это представление даст такие же оценки угловых коэффициентов, что и предыдущее, но свободный член d_0 , соответствующий значениям $d_1 = d_2 = 0$, станет свободным коэффициентом первого уравнения с абсциссой в точке 1969,5.

Данные для примера с двумя прямыми,
абсцисса пересечения которых не известна

Номер опыта	Дата	y	d_1	d_2	d_3
1	1970 г.	1,8	1	0	0
2		4,3	2	0	0
3	1971 г.	5,6	3	0	0
4		8,2	4	0	0
5	1972 г.	9,1	5	0	1
6		10,7	5	1	1
7	1973 г.	11,5	5	2	1
8		12,5	5	3	1
9	1974 г.	14,0	5	4	1

Если в зависимость не включать прочих входных переменных, то конструируемая модель будет иметь вид:

$$y = \delta_0 + \delta_1 d_1 + \delta_2 d_2 + \delta_3 d_3 + \varepsilon \quad (2.24)$$

Параметр δ_3 представляет собой шаг изменения, приводящий к эффекту в пятой точке, то есть, другими словами, он равен расстоянию, на котором в этой точке вторая прямая проходит выше (или ниже при $\delta_3 < 0$) первой.

При $\bar{\varepsilon} = 0,5$ для данных из таблицы 2.6 получаются следующие интервальные оценки коэффициентов:

$$\hat{\delta}_0 \in [0,100; 0,500], \hat{\delta}_1 \in [1,800; 1,900],$$

$$\hat{\delta}_2 \in [0,767; 1,400], \hat{\delta}_3 \in [-0,800; 0,933].$$

Разные знаки интервальной оценки $\hat{\delta}_3$ означают, что в пятой точке может оказаться как первая прямая выше второй, так и наоборот, или, что то же самое, абсцисса точки пересечения трендовых прямых может располагаться как по левую, так и по правую сторону от пятой точки.

Центр прямоугольника дает следующие точечные оценки:

$$\hat{\delta}_0 = 0,300, \hat{\delta}_1 = 1,825, \hat{\delta}_2 = 1,083, \hat{\delta}_3 = 0,067.$$

Соответствующие полученным точечным оценкам графики трендовых линий, выражающиеся в шкале переменной d_1 уравнениями

$$\hat{y} = 0,300 + 1,850d_1 \text{ и } \hat{y} = 4,200 + 1,033d_1,$$

приведены на рисунке 2.5.

Абсцисса точки пересечения прямых в шкале переменной d_1 при этих значениях коэффициентов равна 5,087 или, после перевода во временную шкалу, 15.01.1972 19:41.

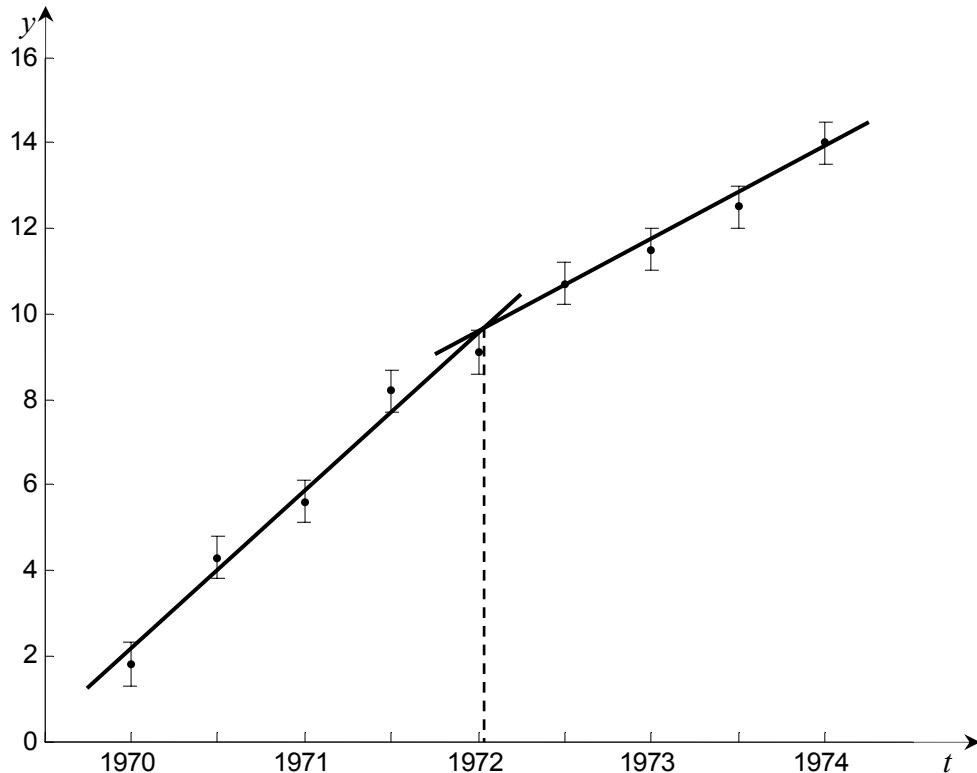


Рис. 2.5 Графики двух линейных трендов, стыкующиеся в точке с заранее неизвестной абсциссой.

Таким образом, в случае правильных наблюдений модели, полученные с помощью приема использования фиктивных переменных, примененного в настоящем разделе для решения задач дисперсионного и ковариационного анализов, позволяют при построении и анализе зависимостей принимать в расчет не только количественные, но и качественные входные переменные. Кроме того, использование фиктивных переменных позволяет вести обработку данных, содержащих временные тренды.

2.3 Обработка информации в случае неправильных наблюдений

Наиболее значимым с практической точки зрения свойством метода центра неопределенности является его потенциальная способность выявлять ситуации, в которых собранные для построения зависимости совокупности знаний и данных противоречивы. Основными источниками противоречий являются либо нарушение гипотезы о структуре конструируемой зависимости, либо наличие выбросов в данных. Выбор способа разрешения противоречий в конечном итоге определяется исследователем по результатам всестороннего анализа. Однако результаты анализа во многом зависят и от того, какой информацией располагает для этого исследователь. Настоящий раздел посвящен описанию возможностей, предоставляемых нестатистическими методами, для получения информации, позволяющей разрешить противоречия, возникающие в случае неправильных наблюдений. Предлагаемый метод выявления выбросов является развитием идеи, изложенной в главе 1, и состоящей в решении задачи (1.28)-(1.29).

Присутствие выбросов в совокупности собранных данных является одной из главных причин пустоты множества неопределенности на первичных этапах построения зависимости. Выброс представляет собой определенную особенность, нетипичное наблюдение по отношению к остальным данным. Это означает, что выбросы должны подвергаться особенно тщательному рассмотрению с целью выяснения причин их возникновения. Иногда выброс дает такую информацию, которую не могут дать другие наблюдения, и является результатом измерений при необычной комбинации условий. В этом случае требуется дальнейшее углубленное исследование. Однако чаще выбросы вызваны грубыми промахами при регистрации значений наблюдаемых величин. В этом случае производится исключение наблюдений из общей информационной совокупности.

Выброс, обусловленный грубым промахом при регистрации результатов измерений, можно трактовать как наблюдение, предельная погрешность

которого ε_j занижена по отношению к реальной ошибке, имевшей место при измерении. Чтобы такое наблюдение стало фиктивно правильным, необходимо найти такое новое значение ε_j' , являющееся нижней границей реальной ошибки, при котором наблюдение не будет вступать в противоречие с остальными.

Нижние границы предельных ошибок наблюдений, при которых множество неопределенности становится непустым, можно отыскивать, решая задачу

$$\min_{\beta, w} \sum_{j=1}^N w_j, \quad (2.25)$$

$$y_j - w_j \varepsilon_j \leq \sum_{i=0}^n \beta_i x_{ji} \leq y_j + w_j \varepsilon_j, \quad w_j \geq 1, \quad j = 1, \dots, N, \quad (2.26)$$

где w_j – масштабирующие коэффициенты, при умножении на которые исходные значения ошибок наблюдений ε_j дают искомые величины ε_j' . При этом выполнение неравенства $w_j > 1$ означает, что j -е наблюдение является выбросом и для совместимости с общей информационной совокупностью необходимо увеличить соответствующую ему предельную ошибку наблюдения в w_j раз. Если у исследователя есть основания считать, что надежность некоторых наблюдений одинакова, то система ограничений (2.26) может быть пополнена ограничениями вида $w_{j_1} = w_{j_2} = \dots = w_{j_K}$. В случае, когда в надежности каких-либо наблюдений исследователь уверен полностью, при решении задачи (2.25)-(2.26) соответствующие им величины w_j можно положить равными единице.

Количество наблюдений, для которых масштабирующие коэффициенты w_j , полученные в результате решения задачи (2.15)-(2.26), превосходят единицу, позволяет судить о доле выбросов в совокупности данных. Большая доля выбросов может говорить либо о неверно выбранной структуре зависимости, либо о том, что предельные ошибки измерения занижены во многих

наблюдениях (например, в результате неверной оценки точности измерительного прибора).

Проиллюстрируем высказанные предложения примером. Данные из таблицы 2.7 получены четырьмя различными способами A , B , C и D и имеют диаграмму рассеяния, приведенную на рисунке 2.6. Требуется построить простую линейную зависимость $y = \beta_0 + \beta_1 x$. Построив, как и ранее, на основе таблицы данных систему ограничений

$$y_j - \varepsilon_j \leq \beta_0 + \beta_1 x_j \leq y_j + \varepsilon_j, \quad j = 1, \dots, 10,$$

выясним, что она противоречива.

Таблица 2.7

Данные с выбросами

Номер опыта	Способ наблюдения	y	x	ε
1	A	2,13	1	0,20
2	A	2,95	2	0,20
3	A	5,01	3	0,20
4	A	4,99	4	0,20
5	A	5,97	5	0,20
6	B	7,04	6	0,40
7	B	8,02	7	0,40
8	C	8,15	8	0,40
9	C	10,01	9	0,40
10	D	10,98	10	0,50

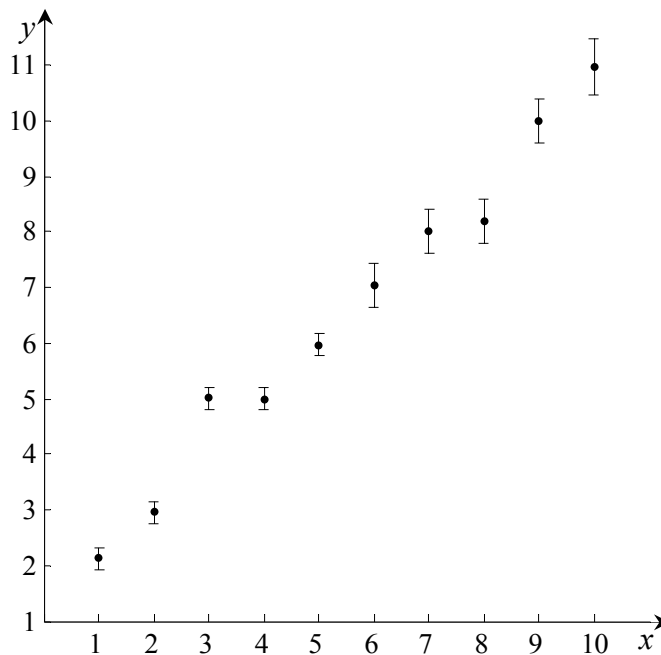


Рис 2.6. Диаграмма рассеяния данных из таблицы 2.7.

Чтобы понять, имеются ли выбросы, для начала пытаемся решить в отношении наших данных задачу (2.25)-(2.26). Полученные значения масштабирующих коэффициентов предельной ошибки приведены в первом столбце таблицы 2.8. Анализируя эти числа, можно прийти к выводу, о том, что третье и восьмое наблюдения вступают в противоречие с остальными. Причем третье наблюдение особенно не вписывается в общую картину, поскольку посчитать его правильным возможно только при увеличении предельной ошибки измерения более чем в четыре с половиной раза, и это притом, что способ *A*, которым получено третье наблюдение, является наиболее точным из всех четырех. Эти соображения позволяют нам склониться к заключению, что третье наблюдение есть результат грубых промахов при его проведении и должно быть исключено из всех дальнейших построений.

Что касается восьмого наблюдения, то его принадлежность к неправильным менее выражена. Здесь следует отработать как гипотезу о грубых промахах во время проведения этого измерения, так и о возможной переоценке точности способа *C*. С этой целью задачу (2.1)-(2.2) нужно решать, либо исключая из рассмотрения восьмое наблюдение, либо предполагая, что способ *C* менее точен, но все выполненные им наблюдения правильны, т.е.

$$w_8 = w_9. \quad (2.27)$$

Таблица 2.8

Масштабирующие коэффициенты w_j

предельной ошибки измерений для данных таблицы 2.7

Номер опыта	Решение задачи (2.25)-(2.6)	Решение задачи (2.25)-(2.27)
1	1,000	1,000
2	1,000	1,000
3	4,686	—
4	1,000	1,000
5	1,000	1,000
6	1,000	1,000
7	1,000	1,000
8	1,343	1,143
9	1,000	1,143
10	1,000	1,000

Результат решения задачи (2.25)-(2.27), приведенный во втором столбце таблицы 2.8, говорит о том, что точность измерений способом C завышена и обеспечиваемая им предельная ошибка не может иметь значение менее чем $w_8\varepsilon_8 = w_9\varepsilon_9 = 1,143 \cdot 0,40 \approx 0,46$. Такой вывод, конечно, не может служить основанием автоматического увеличения ε_8 и ε_9 до указанного уровня, а может означать лишь необходимость дополнительных исследований точности способа измерений C .

Таким образом, в каждом из случаев

- 1) если и третье, и восьмое наблюдения исключить из совокупности наблюдений;
- 2) если в результате дополнительных исследований выяснится, что способ измерений C , действительно, менее точен и предельная ошибка восьмого и девятого измерений будет увеличена;

множество неопределенности в задаче (2.25)-(2.26) станет непустым, что соответствует ситуации правильных наблюдений. Дальнейший анализ может проводиться с использованием приемов, описанных в разделе 2.2.

2.4. Метод центра неопределенности и статистические методы оценивания: сравнительный анализ

Любой метод построения эмпирических зависимостей опирается на некоторую систему гипотез, которые должны выполняться для получения качественных оценок параметров зависимостей. Одним из методов выяснения качества оценок при нарушении тех или иных гипотез является имитационный эксперимент. На актуальность проведения имитационного эксперимента, позволяющего сравнить статистические и нестатистические методы построения эмпирических зависимостей, указывалось, в частности, при обсуждении метода анализа данных, предложенного А.П. Воцининым [25]. Имитационный эксперимент является главным способом сравнения методов, опирающихся на различные системы гипотез.

Настоящий раздел посвящен экспериментальному изучению поведения оценок прогнозных значений эмпирической зависимости, полученных методом центра неопределенности (МЦН) и статистическими методами – методом максимума правдоподобия (ММП) и методом наименьших квадратов (МНК) – при различных распределениях ошибки в условиях правильных наблюдений. Как указано в [45] ММП в достаточно широком круге практически важных случаев является в известном смысле наилучшим, что способствует его достаточно широкому употреблению. Свойства МНК, который является частной формой ММП, и причины его широкого использования в задачах эмпирического моделирования подробно освещены в главе 1.

Описательную способность каждого из рассматриваемых методов предлагается выяснить по результатам вычислительного эксперимента, состоящего в многократном решении следующей задачи прогноза:

По таблице экспериментальных данных $T = \{(x, y)\}$ построить оценку $\hat{\beta}$ параметров функциональной зависимости $y = f(x, \beta)$ и вычислить оценку прогнозного значения функции $\hat{y}^* = f(x^*, \hat{\beta})$ в заданной точке x^* .

(2.28)

Схема вычислительного эксперимента состоит в следующем. Для заданной линейно параметризованной функции $y = f(x, \beta^0)$ с известными параметрами β^0 предварительно формируется совокупность пар значений

$$T^0 = \{(x_i, y_i^0) \mid y_i^0 = f(x_i, \beta^0), i = 1, \dots, P\}. \quad (2.29)$$

Далее, в каждом n -м испытании ($n = 1, \dots, N$) в каждой точке x_i имитируется Q «наблюдений» за значением выходной переменной y с ошибкой ε , имеющей заданное распределение, то есть генерируется таблица экспериментальных данных

$$T^{(n)} = \{(x_i, y_{ij}^{(n)}) \mid y_{ij}^{(n)} = y_i^0 + \varepsilon_{ij}, i = 1, \dots, P, j = 1, \dots, Q\}, \quad (2.30)$$

где ε_{ij} – значение ошибки j -го наблюдения за значением y в точке x_i . Таким образом, полная таблица наблюдений $T^{(n)}$ содержит $P \cdot Q$ строк.

По сгенерированной таблице экспериментальных данных $T^{(n)}$ в каждом n -м испытании производится оценивание параметров функциональной зависимости f двумя сравниваемыми методами. Полученные оценки параметров для первого и второго методов обозначим, соответственно, $\hat{\beta}_{ММП}$ и $\hat{\beta}_{МЦН}$. На основе этих оценок параметров вычисляются прогнозные значения

функциональной зависимости $y_m^{(n)} = f(x^*, \hat{\beta}_m^{(n)})$, $m = ММП, МЦН$.

По результатам N испытаний вычисляются среднеквадратичные отклонения прогнозных значений $y_m^{(n)}$ от истинного значения $y^* = f(x^*, \beta)$:

$$d_m = \sqrt{\frac{1}{N} \sum_{n=1}^N (y^* - y_m^{(n)})^2}, \quad m = ММП, МЦН. \quad (2.31)$$

Сравнение значений $d_{ММП}$ и $d_{МЦН}$ позволяет выяснить «качество» прогнозных значений, обеспечиваемое различными методами при заданном распределении ошибки наблюдений.

Что касается выбора распределения ошибки, то интерес представляют ситуации «наилучшие» для каждого из сравниваемых методов, а также некоторые близкие к ним варианты. Наилучшими условиями для статистических методов являются ситуации, когда ошибка распределения подчиняется некоторому унимодальному распределению, в частности, для МНК таковым является нормальное распределение ошибки. Базовому для МЦН предположению о равнозначности всех элементов множества неопределенности в вероятностных терминах наиболее адекватно соответствует равномерное распределение. Таким образом, сравнительный эксперимент предлагается провести для унимодального и равномерного распределений ошибки, а также некоторых промежуточных распределений.

2.4.1. Метод центра неопределенности и метод максимального правдоподобия

Сравнение МЦН и ММП предлагается произвести для распределения ошибки, изменяющегося от треугольного до равномерного. При этом пути «эволюции» треугольного распределения к равномерному могут быть различны. Предлагается три варианта описания этого процесса с помощью семейств распределений, имеющих плотности вероятностей с носителем $[-\varepsilon, \varepsilon]$ и зависящими от параметра α :

$$p_{\alpha}^1(x) = \begin{cases} \frac{\varepsilon + x}{\varepsilon^2 - \alpha^2}, & -\varepsilon \leq x < -\alpha \\ \frac{1}{\varepsilon + \alpha}, & -\alpha \leq x < \alpha \\ \frac{\varepsilon - x}{\varepsilon^2 - \alpha^2}, & \alpha \leq x \leq \varepsilon \end{cases}, \quad \alpha \in [0, \varepsilon]; \quad (2.32)$$

$$p_{\alpha}^2(x) = \begin{cases} \frac{1 - 2\varepsilon\alpha}{\varepsilon^2}x + \frac{1 - \varepsilon\alpha}{\varepsilon}, & -\varepsilon \leq x < 0 \\ \frac{2\varepsilon\alpha - 1}{\varepsilon^2}x + \frac{1 - \varepsilon\alpha}{\varepsilon}, & 0 \leq x \leq \varepsilon \end{cases}, \quad \alpha \in \left[0, \frac{1}{2\varepsilon}\right]; \quad (2.33)$$

$$p_{\alpha}^3(x) = \begin{cases} \frac{1}{\varepsilon(\alpha + \varepsilon)}x + \frac{\alpha^2 + \varepsilon\alpha + 2\varepsilon^2}{2\varepsilon^2(\alpha + \varepsilon)}, & -\varepsilon \leq x < -\alpha \\ \frac{\alpha^2 - \varepsilon\alpha + 2\varepsilon^2}{2\varepsilon^2(\alpha + \varepsilon)}, & -\alpha \leq x < \alpha \\ -\frac{1}{\varepsilon(\alpha + \varepsilon)}x + \frac{\alpha^2 + \varepsilon\alpha + 2\varepsilon^2}{2\varepsilon^2(\alpha + \varepsilon)}, & \alpha \leq x \leq \varepsilon \end{cases}, \quad \alpha \in [0, \varepsilon]. \quad (2.34)$$

Изменение параметра α от нулевого значения до максимального α_{\max} вызывает «перетекание» треугольного распределения в равномерное. При $\varepsilon = 1$ графики функций p_{α}^1 , p_{α}^2 и p_{α}^3 для граничных и двух внутренних значений параметра α приведены на рисунках 2.7-2.9. При проведении эксперимента в качестве значений параметра α выбираются узлы регулярной сетки

$$\alpha_k = \frac{k}{K}\alpha_{\max}, \quad k = 0, 1, \dots, K. \quad (2.35)$$

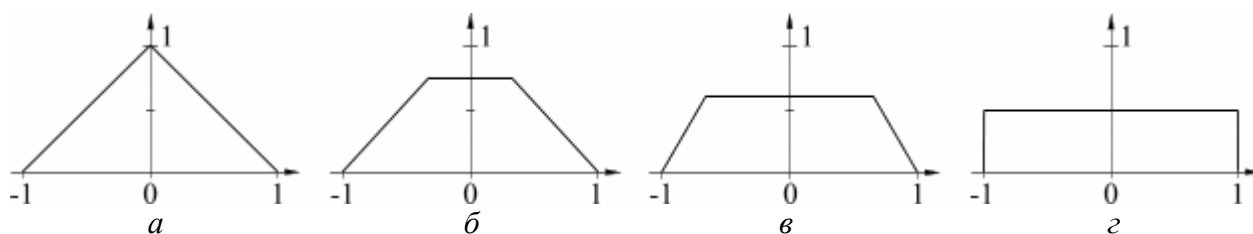


Рис. 2.7. Графики функции плотности p_α^1 при $\varepsilon = 1$ и а) $\alpha = 0$, б) $\alpha = 1/3$, в) $\alpha = 2/3$, г) $\alpha = 1$.

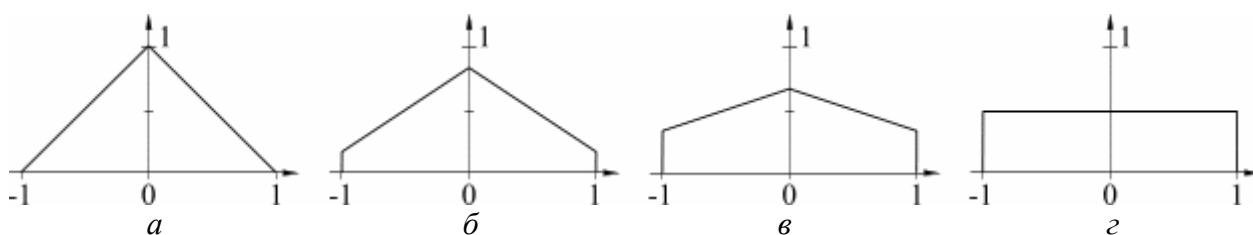


Рис. 2.8. Графики функции плотности p_α^2 при $\varepsilon = 1$ и а) $\alpha = 0$, б) $\alpha = 1/6$, в) $\alpha = 1/3$, г) $\alpha = 1/2$.

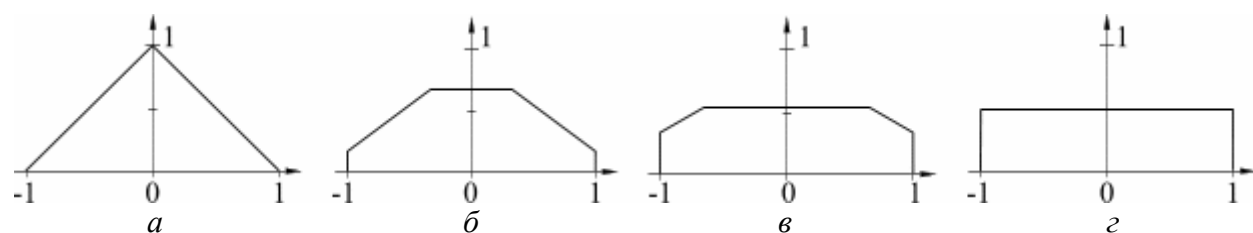


Рис. 2.9. Графики функции плотности p_α^3 при $\varepsilon = 1$ и а) $\alpha = 0$, б) $\alpha = 1/3$, в) $\alpha = 2/3$, г) $\alpha = 1$.

При построении оценок ММП при распределении ошибки, близком к равномерному, возникают сложности в выборе оценки, обусловленные неединственностью максимума функции правдоподобия. Выход из этой ситуации видится в регуляризации задачи поиска максимума функции правдоподобия

$$L(\beta) = \prod_{i=1}^{P,Q} p(y_i - f(x_i, \beta)), \quad (2.36)$$

где p – функция плотности распределения ошибки. С этой целью предлагается к максимизируемой функции добавить регуляризирующее слагаемое вида

$$\delta |L(\beta)| (\beta - \beta^0)^2, \quad (2.37)$$

где $\delta < 0$ – постоянный весовой коэффициент, β^0 – известные значения оцениваемых параметров.

Описанная схема статистических испытаний была реализована в виде программы на языке C++ (компилятор IBM C Set++ for AIX 4.2) для рабочей станции IBM RS/6000 Model 43P-140. В качестве исследуемой зависимости рассматривалась функция

$$y = x + 1, \quad (2.38)$$

то есть вектор истинных значений параметров $\beta^0 = (1, 1)$. Совокупность пар значений T^0 представляла собой значения функции (2.38) в узлах регулярной сетки с шагом 1 на интервале $[1, 10]$, то есть в (2.29) $x_i = i$, $i = 1, \dots, 10$, а $P = 10$. Совокупность экспериментальных данных $T^{(n)}$ в n -м испытании строилась на основе T^0 добавлением к каждому y_i^0 случайной ошибки ε_{ij} – случайной величины из интервала $[-\frac{1}{2}, \frac{1}{2}]$ с одной из функций плотности (2.32)-(2.34). Кратность наблюдений Q в каждой из точек x_i равнялась пяти. Таким образом, суммарное количество строк-наблюдений в таблице $T^{(n)}$ составляло $P \cdot Q = 50$.

Датчики случайных чисел с заданными плотностями были реализованы на основе следующего известного факта. Случайные числа с произвольной функцией распределения $F(x)$ могут быть построены по последовательности равномерно распределенных чисел ξ_i как $\eta_i = F^{-1}(\xi_i)$, то есть найдены из уравнения $\xi_i = F(\eta_i)$, $i = 1, 2, \dots$. Источником равномерно распределенных псевдослучайных чисел служила стандартная функция `rand()`.

Прочие параметры схемы статистических испытаний выбирались следующими. Весовой коэффициент регуляризирующего слагаемого (2.37) $\delta = -0,01$. Количество узлов в сетке (2.35), определяющей значения параметра α , $K + 1 = 21$. Количество повторений эксперимента при фиксированных параметрах метода статистических испытаний принималась равной $N = 10000$. В этом случае дисперсии оценок параметров была настолько ма-

ла, что их можно считать точными. Оценки прогнозного значения функции $\hat{y}^* = f(\hat{x}^*, \hat{\beta})$ строились в точке $x^* = 5,5$, где истинное значение функции $y^* = 6,5$.

Результаты эксперимента для каждого из семейств плотностей распределения ошибки (2.5)-(2.7) приведены в таблице 2.9 и на рисунках 2.10-2.12.

Таблица 2.9

Среднеквадратичные отклонения оценок прогнозных значений, полученных ММП ($d_{ММП}$) и МЦН ($d_{МЦН}$) при различных распределениях ошибки

k	$p_{\alpha_k}^1(x)$		$p_{\alpha_k}^2(x)$		$p_{\alpha_k}^3(x)$	
	$d_{ММП}$	$d_{МЦН}$	$d_{ММП}$	$d_{МЦН}$	$d_{ММП}$	$d_{МЦН}$
0	0,000242	0,000572	0,000241	0,000572	0,000227	0,000596
1	0,000235	0,000576	0,000253	0,000598	0,000238	0,000593
2	0,000242	0,000573	0,000254	0,000594	0,000262	0,000586
3	0,000239	0,000552	0,000269	0,000588	0,000261	0,000576
4	0,000239	0,000568	0,000269	0,000560	0,000273	0,000560
5	0,000241	0,000536	0,000289	0,000548	0,000275	0,000521
6	0,000239	0,000523	0,000296	0,000517	0,000275	0,000475
7	0,000232	0,000514	0,000308	0,000507	0,000283	0,000451
8	0,000232	0,000482	0,000313	0,000473	0,000286	0,000438
9	0,000227	0,000458	0,000313	0,000448	0,000284	0,000397
10	0,000227	0,000437	0,000319	0,000423	0,000268	0,000380
11	0,000213	0,000407	0,000326	0,000389	0,000278	0,000351
12	0,000207	0,000378	0,000320	0,000374	0,000269	0,000304
13	0,000204	0,000349	0,000323	0,000347	0,000251	0,000293
14	0,000185	0,000313	0,000325	0,000312	0,000226	0,000261
15	0,000178	0,000278	0,000333	0,000295	0,000220	0,000257
16	0,000156	0,000247	0,000328	0,000269	0,000197	0,000226
17	0,000144	0,000228	0,000327	0,000254	0,000173	0,000210
18	0,000118	0,000208	0,000318	0,000228	0,000137	0,000209
19	0,000081	0,000191	0,000309	0,000219	0,000095	0,000212
20	0,000000	0,000206	0,000000	0,000212	0,000000	0,000215

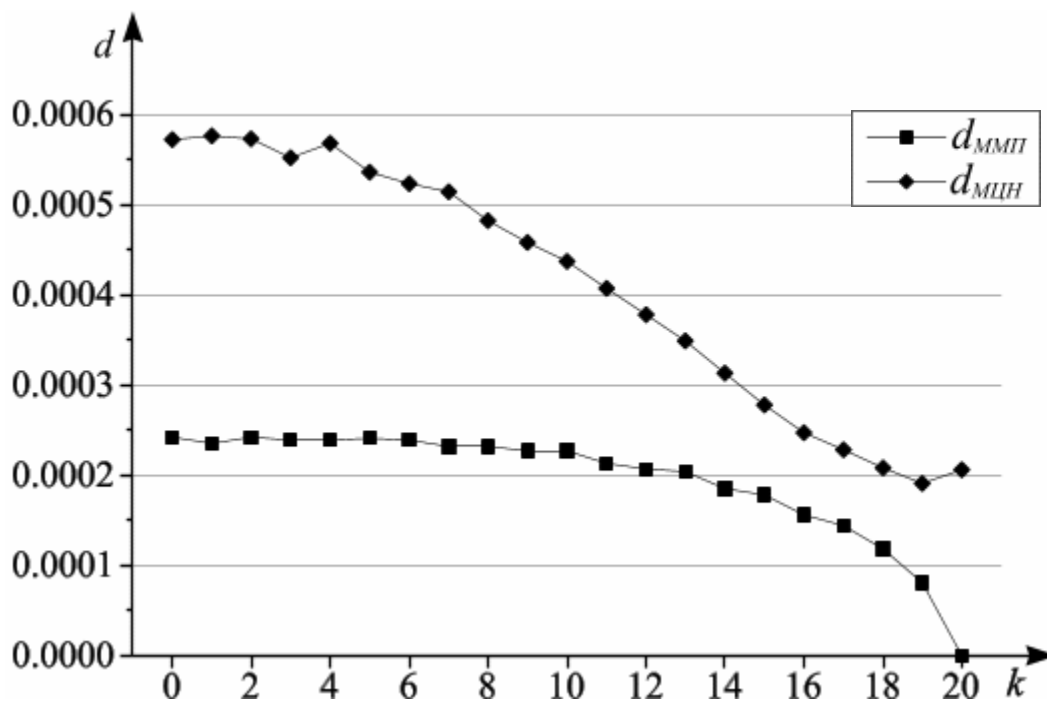


Рис. 2.10. Среднеквадратичные отклонения оценок прогнозных значений от истинных значений для ММП ($d_{ММП}$) и МЦН ($d_{МЦН}$) при ошибке с плотностью распределения $p_{\alpha_k}^1(x)$.

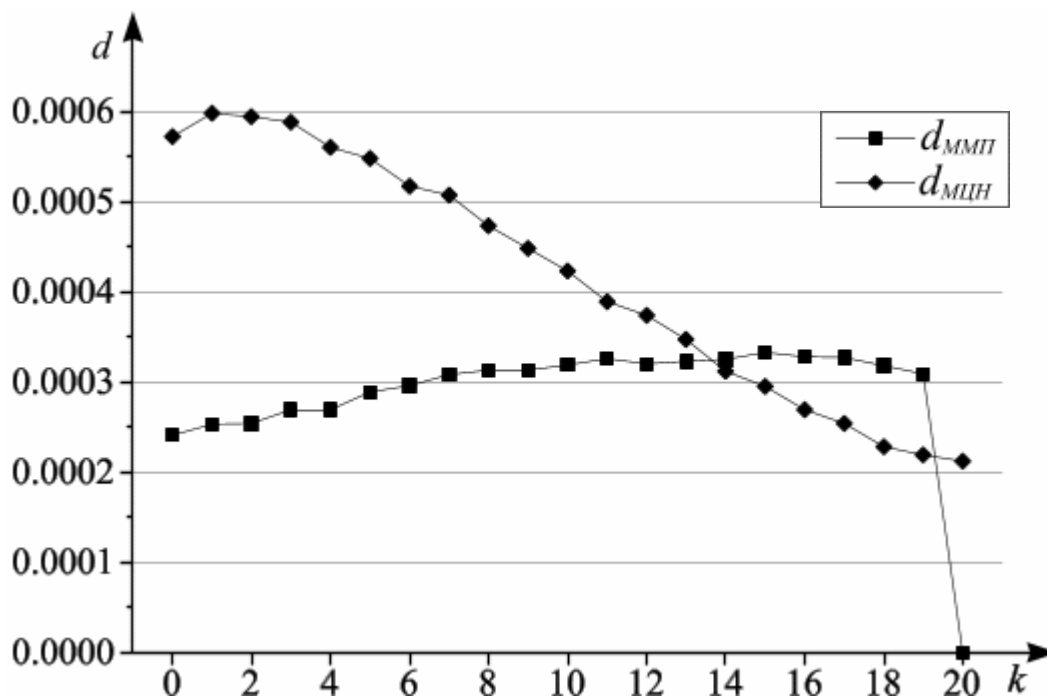


Рис. 2.11 Среднеквадратичные отклонения оценок прогнозных значений от истинных значений для ММП ($d_{ММП}$) и МЦН ($d_{МЦН}$) при ошибке с плотностью распределения $p_{\alpha_k}^2(x)$.

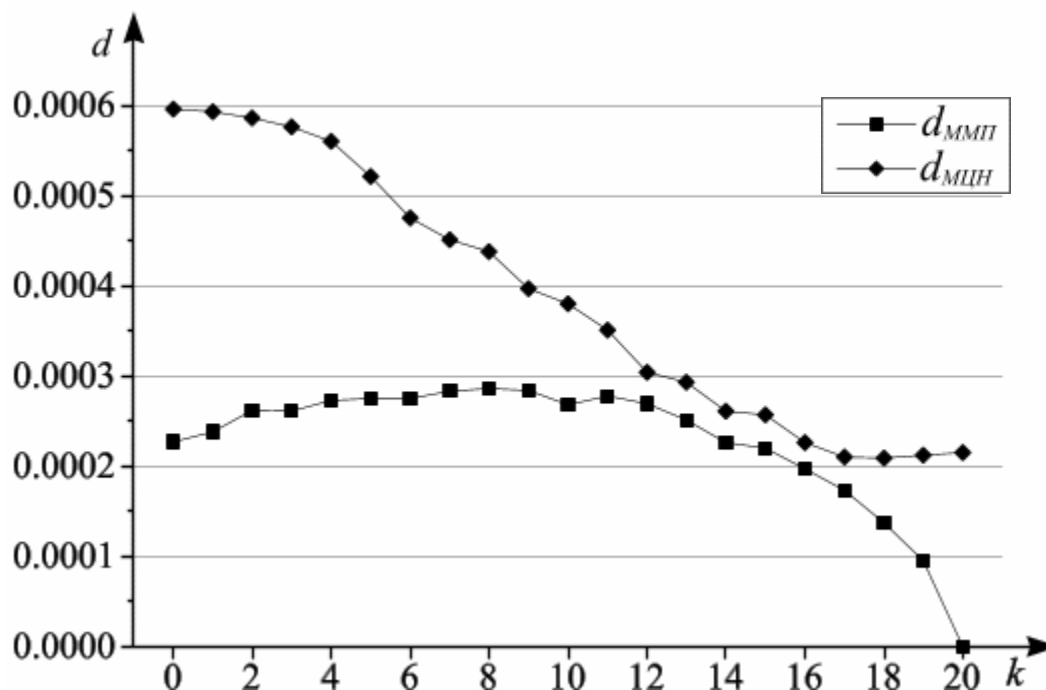


Рис. 2.12. Среднеквадратичные отклонения оценок прогнозных значений от истинных значений для ММП ($d_{ММП}$) и МЦН ($d_{МЦН}$) при ошибке с плотностью распределения $p_{\alpha_k}^3(x)$.

Анализ результатов проведенного сравнения ММП и МЦН показывает, что при распределениях погрешности, близких к «треугольным», характер поведения ошибки прогноза соответствует известным соотношениям и закономерностям, свойственным использованным методам оценивания. Действительно, МЦН как нестатистическая процедура не учитывает дополнительную информацию, связанную с характером распределения, и, соответственно, имеет большую ошибку прогноза. Кроме того, ошибка прогноза, обеспечиваемая МЦН, снижается по мере приближения распределения погрешности к равномерному. Это объясняется тем, что в такой ситуации становится выполненным базовое для МЦН предположение о равноценности всех элементов множества неопределенности и устойчивость оценок повышается. Одинаковое поведение оценок МЦН для всех семейств распределений служит подтверждением достоверности результатов.

Рассматривая оценки метода максимального правдоподобия, следует отметить, что среднеквадратичное отклонение прогнозного и истинного значений при совпадении распределений с треугольным, практически, совпадает

во всех трех случаях ($d_{ММП} \in [0,000227;0,000242]$). Существенным представляется также заметить, что при распределениях погрешности, удаляющихся от треугольного, эта величина не убывает. Рост погрешности объясняется тем, что по мере приближения к равномерному распределению количество используемой для оценивания информации уменьшается. Дальнейшее нарушение этой тенденции объясняется «притяжением» оценок к истинным значениям искомых коэффициентов зависимости, вызванным возрастающим весом регуляризирующего слагаемого (2.37).

Проведем сравнение МЦН и статистических процедур по точности оценок в предельном случае равномерного распределения. Исходя из выше сказанного и анализируя результаты эксперимента, мы можем утверждать, что ошибка прогноза МЦН будет меньше соответствующей ошибки для ММП. Действительно, среднеквадратичное отклонение ошибки МЦН для равномерного распределения оценена в пределах $d_{МЦН} \in [0,000206;0,000215]$, что меньше указанной выше величины $d_{ММП}$ для статистических процедур в случае треугольного распределения. При этом $d_{ММП}$ возрастает по мере приближения распределения к равномерному.

Данный результат, который мы готовы защищать, вообще говоря, противоречит известным закономерностям соотношения точности статистических и нестатистических процедур оценивания. Причины этого явления могут быть рассмотрены как дискуссионные.

2.4.2. Метод центра неопределенности и метод наименьших квадратов

Сравнение МЦН и МНК предлагается провести при ошибке, имеющей нормальное распределение, усеченное на разных уровнях, что позволяет имитировать условия оценивания, соответствующие гипотезам каждого из методов и некоторые промежуточные ситуации.

На практике, при использовании МНК для построения эмпирических зависимостей распределение ошибки, как правило, полагают усеченным

нормальным $N_k(a, \sigma^2)$, то есть полагают, что ошибка принимает значения из интервала $[a - k\sigma, a + k\sigma]$, где a – математическое ожидание, σ – средне-квадратическое отклонение, k – некоторая константа, определяющая уровень отсечения.. Использование усеченного нормального распределения обосновывается тем, что значения нормально распределенной случайной величины с достаточно большой вероятностью сосредоточены на конечном интервале.

Значение уровня отсечения k нормального распределения, описывающего ошибку измерения, на практике выбирают из интервала $[1,5; 2,5]$ (см., например, [56]). Это эмпирическое правило недавно получило и теоретическое обоснование [131]. В проведенной серии экспериментов уровень отсечения выбирался из более широкого интервала $[0,2, 3]$ с шагом 0,2. По мере роста уровня отсечения получаемые распределения принимали вид от почти равномерного, при $k = 0,2$, до почти нормального, при $k = 3$.

Совокупность пар значений T^0 представляла собой значения функции (2.38) в узлах регулярной сетки с шагом 1 на интервале $[1; 10]$, то есть в (2.29) $x_i = i$, $i = 1, \dots, 10$, а $P = 10$. Совокупность экспериментальных данных $T^{(n)}$ в n -м испытании строилась на основе T^0 добавлением к каждому y_i^0 ошибки ε_{ij} – случайной величины с распределением $N_k(0,1)$. Для каждого из значений уровня отсечения нормального распределения ошибки k , а также одного из значений кратности «измерений» $Q = 1, 2, 3, 5, 7, 9$ осуществлялось $N = 1000$ испытаний. Оценки прогнозного значения функции $y^* = f(x^*, \beta)$ строились в точке $x^* = 5,5$, где истинное значение функции $y^* = 6,5$. Схема статистических испытаний была реализована в виде сценария (m-файла) системы MATLAB. Датчик случайных чисел с усеченным нормальным распределением был реализован на основе стандартной функции `randn()`, возвращающей нормально распределенные числа.

Результаты вычислительных экспериментов приведены в таблицах 2.10 и 2.11, а также графически представлены на рисунках 2.13-2.15.

Таблица 2.10

Среднеквадратическое отклонение МНК-прогноза ($d_{\text{МНК}}$) при различных уровнях усечения нормального распределения ошибки (k) и кратности измерений (Q)

$k \backslash Q$	1	2	3	5	7	9
0,2	0,090	0,066	0,054	0,038	0,033	0,031
0,4	0,090	0,066	0,052	0,041	0,033	0,030
0,6	0,089	0,064	0,050	0,041	0,033	0,031
0,8	0,088	0,064	0,052	0,040	0,033	0,029
1,0	0,083	0,059	0,047	0,038	0,033	0,029
1,2	0,079	0,060	0,048	0,037	0,032	0,027
1,4	0,081	0,058	0,045	0,034	0,030	0,027
1,6	0,079	0,055	0,044	0,034	0,028	0,025
1,8	0,075	0,054	0,043	0,034	0,026	0,024
2,0	0,069	0,051	0,041	0,030	0,026	0,023
2,2	0,066	0,047	0,036	0,029	0,025	0,021
2,4	0,063	0,043	0,037	0,028	0,024	0,021
2,6	0,058	0,041	0,034	0,026	0,023	0,020
2,8	0,055	0,038	0,031	0,025	0,021	0,019
3,0	0,052	0,038	0,031	0,023	0,019	0,018

Таблица 2.11

Среднеквадратическое отклонение МЦН-прогноза ($d_{\text{МЦН}}$) при различных уровнях усечения нормального распределения ошибки (k) и кратности измерений (Q)

$k \backslash Q$	1	2	3	5	7	9
0,2	0,095	0,056	0,041	0,025	0,019	0,015
0,4	0,093	0,062	0,038	0,026	0,019	0,015
0,6	0,100	0,060	0,042	0,026	0,020	0,016
0,8	0,099	0,061	0,043	0,030	0,021	0,017
1,0	0,101	0,063	0,048	0,031	0,023	0,018
1,2	0,099	0,068	0,052	0,034	0,025	0,020
1,4	0,102	0,069	0,052	0,037	0,028	0,024
1,6	0,105	0,071	0,054	0,039	0,033	0,027
1,8	0,107	0,070	0,059	0,044	0,036	0,030
2,0	0,109	0,077	0,061	0,046	0,039	0,035
2,2	0,112	0,076	0,062	0,051	0,044	0,039
2,4	0,111	0,078	0,063	0,054	0,046	0,042
2,6	0,117	0,089	0,065	0,057	0,050	0,045
2,8	0,115	0,097	0,070	0,057	0,051	0,047
3,0	0,109	0,099	0,077	0,058	0,052	0,048

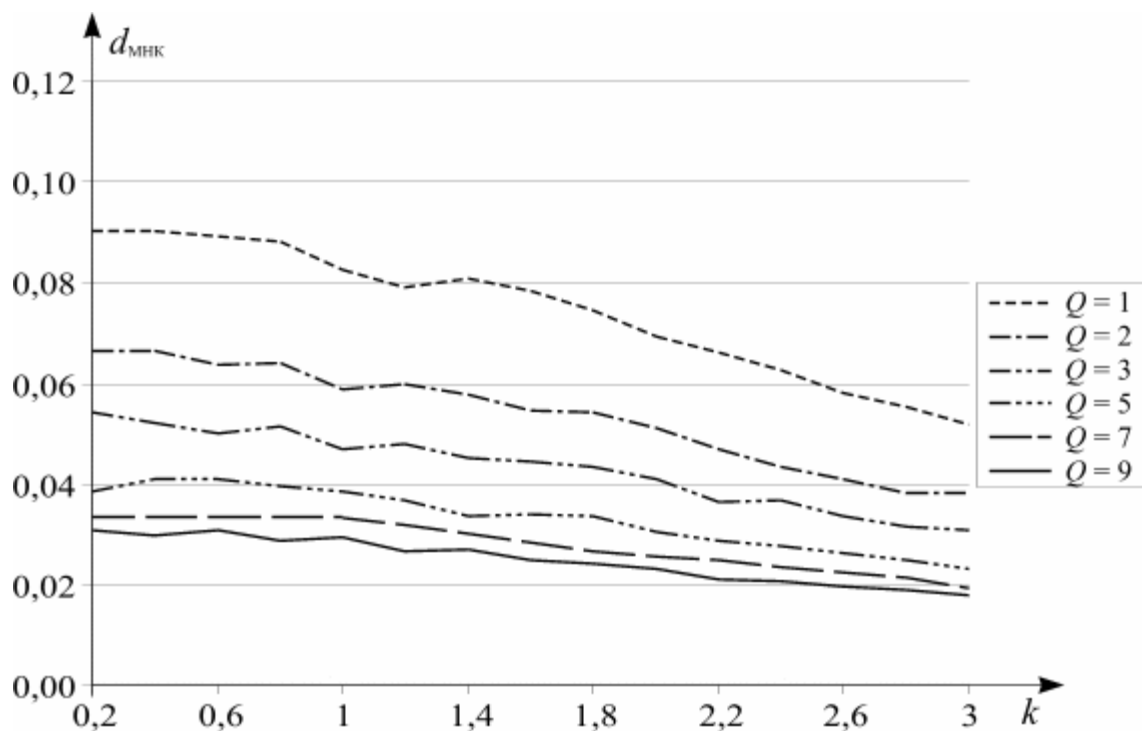


Рис 2.13. Зависимость среднеквадратического отклонения прогнозных значений, полученных МНК (d_{MNK}), от уровня отсечения нормального распределения (k) и кратности наблюдений (Q).

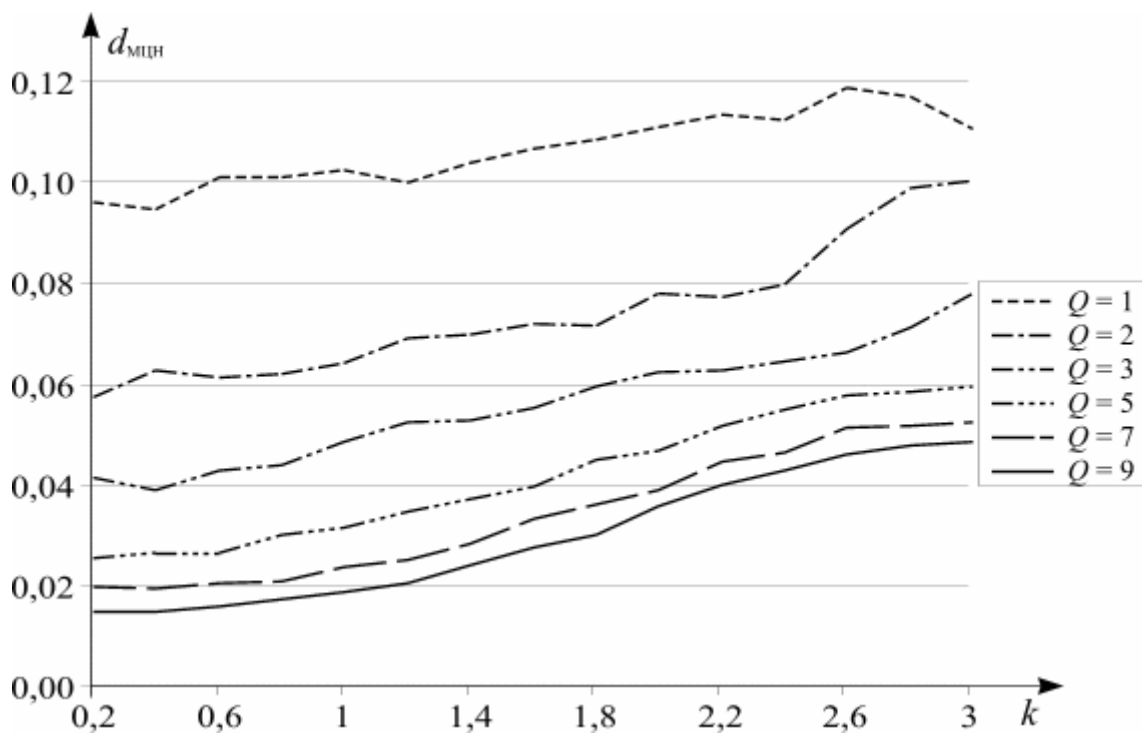


Рис 2.14. Зависимость среднеквадратического отклонения прогнозных значений, полученных МЦН (d_{MCN}), от уровня отсечения нормального распределения (k) и кратности наблюдений (Q).

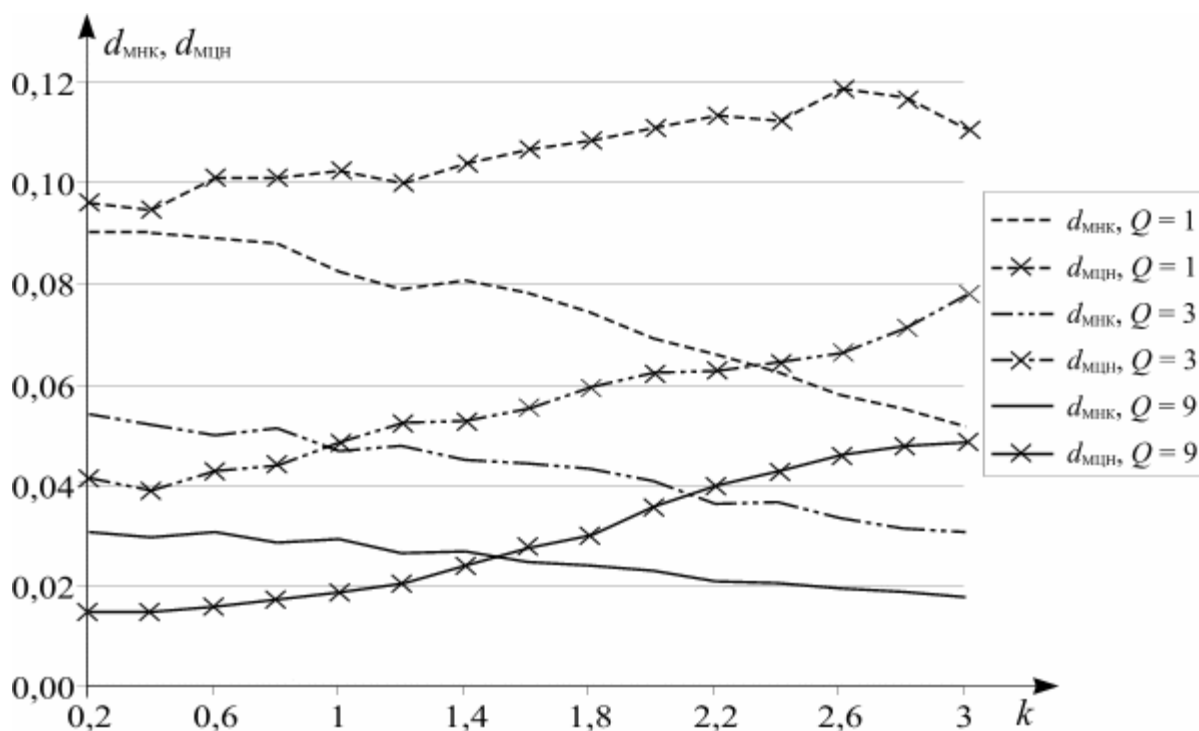


Рис 2.15. Зависимость среднеквадратического отклонения прогнозных значений, полученных МНК ($d_{\text{МНК}}$) и МЦН ($d_{\text{МЦН}}$), от уровня отсечения нормального распределения (k) и кратности наблюдений (Q).

Качественный анализ взаимосвязей среднеквадратичных отклонений МНК- и МЦН-прогнозов с уровнем усечения нормального распределения ошибки и кратностью измерений позволяет сделать следующие наблюдения:

1. По мере уменьшения уровня усечения нормального распределения ошибки измерений k среднеквадратичные отклонения и МНК-, и МЦН-прогнозов также убывают. При этом для МНК-прогноза скорость убывания можно качественно охарактеризовать как логарифмическую или линейно-логарифмическую, в то время как для МЦН-прогноза – как полиномиальную.

2. При больших значениях k оценки МНК-прогноза более устойчивы, чем МЦН-оценки. Однако с уменьшением k их преимущество утрачивается. Кроме того, с увеличением кратности измерений Q более устойчивыми становятся МЦН-оценки. Объяснение этому факту состоит в уменьшении степени соответствия распределения ошибки измерения гипотезе о нормальности, в рамках которой МНК дает наилучшие результаты. В то же время, приближение распределения ошибки к равномерному все более соответствует

одному из базовых предположений МЦН о равноценности всех элементов множества неопределенности.

3. С увеличением кратности измерений Q устойчивость МЦН-оценок растет несколько быстрее. Тенденция усиливается по мере уменьшения уровня усечения нормального распределения ошибки измерений, то есть по мере приближения распределения к равномерному. Этот факт свидетельствует о способности МЦН неявно накапливать информацию о распределении ошибки, незадействуемую явным образом в отличие от статистических процедур оценивания, и подтверждает теоретический вывод В.А. Суханова о высокой скорости сходимости МЦН-оценок с ростом числа наблюдений.

Экспериментально выявленные качественные характеристики МНК- и МЦН-оценок свидетельствуют о более высокой эффективности МНК при оценивании параметров эмпирической зависимости при соблюдении гипотез о нормальности и независимости ошибок измерений, хотя при нарушении гипотез о распределении ошибок и увеличении кратности наблюдений МЦН не уступает в качестве оценок. При распределении ошибки, близком к равномерному более эффективными являются МЦН-оценки. Полученные результаты позволяют осуществить выбор процедуры оценивания в зависимости от условий наблюдений в конкретных случаях и, в частности, при отсутствии достоверной информации о распределении ошибки.

Выводы по главе 2

1. Анализ литературы, посвященной нестатистическим методам построения и анализа линейно параметризованных зависимостей по эмпирическим данным, позволяет говорить о принципиальном родстве подходов и приемов, независимо предложенных рядом авторов в различное время, и отсутствии единой устоявшейся терминологии нестатистического подхода.

2. На основе техники применения фиктивных переменных разработаны способы использования метода центра неопределенности для решения в условиях правильных наблюдений 1) задачи построения и анализа зависимостей, включающих наряду с количественными качественными входными пере-

менные; 2) задачи обработки данных, содержащих временные тренды, в том числе кусочно-линейные при известных и неизвестных абсциссах точек излома.

3. Разработан метод обработки противоречивых наблюдений, позволяющий выявлять выбросы путем поиска минимальных коэффициентов увеличения предельной ошибки наблюдений-выбросов, при которых множество неопределенности является непустым.

4. Разработана схема статистических испытаний, позволяющая сравнить статистические и нестатистические методы построения зависимостей. Схема состоит в построении и сравнении разброса оценок прогнозных значений для модельных зависимостей каждым из сравниваемых методов при различных распределениях ошибки. Функция плотности вероятностей ошибки выбирается из параметрического семейства плотностей, в различной степени соответствующих базовым гипотезам сравниваемых методов в зависимости от параметра.

5. На основе разработанной схемы статистических испытаний проведен сравнительный анализ метода центра неопределенности и метода максимального правдоподобия, а также метода центра неопределенности и метода наименьших квадратов. Результаты сравнительного анализа свидетельствуют 1) о более высокой эффективности оценок метода центра неопределенности по отношению к оценкам метода наименьших квадратов при нарушении исходной гипотезы метода наименьших квадратов о виде распределения ошибки и 2) о более высокой эффективности оценок метода центра неопределенности по отношению к оценкам метода максимального правдоподобия при распределении ошибки, близком к равномерному.

ГЛАВА 3. ИСПОЛЬЗОВАНИЕ МЕТОДА ЦЕНТРА НЕОПРЕДЕЛЕННОСТИ ПРИ РЕШЕНИИ ЗАДАЧ ОБРАБОТКИ ПРОСТРАНСТВЕННЫХ ДАННЫХ

3.1 Геометрические преобразования и привязка изображений

Во многих задачах обработки пространственных данных возникает потребность во взаимном сопоставлении изображений различной природы между собой. Наиболее ярким примером может служить использование данных дистанционного зондирования в геоинформационных технологиях. В частности, при решении задач мониторинга экосистем необходимо сопоставлять спутниковые снимки одной и той же территории, сделанные в разное время, и осуществлять их координатную привязку с одновременной геометрической коррекцией для последующего совместного анализа.

Для построения геометрического преобразования изображения требуется установить соответствие между элементами преобразуемого и эталонного изображения, что сводится к выделению так называемых сопряженных (или, по-другому, реперных, контрольных, опорных) точек на изображениях. Точки на двух изображениях называются сопряженными, если они являются образами одной и той же точки сцены.

Если известен некоторый набор сопряженных точек, то задача геометрического преобразования изображения решается в два этапа. На первом этапе производится построение зависимости, связывающей координаты соответствующих сопряженных точек на разных изображениях:

$$\begin{aligned}x &= f(u, v), \\y &= g(u, v),\end{aligned}\tag{3.1}$$

где (u, v) – координаты точек эталонного изображения, (x, y) – соответствующие им координаты на преобразуемом изображении. На втором этапе производится собственно геометрическая коррекция изображения, состоящая в вычислении для каждого элемента изображения в эталонных координатах соответствующего ему элемента преобразуемого изображения с помощью

построенной зависимости (3.1) и восстановлении уровня яркости преобразованного элемента.

Отметим, что координаты опорных точек (u, v) на эталонном изображении известны, как правило, с довольно высокой точностью, поскольку определяются либо по крупномасштабной карте, либо с помощью систем спутниковой навигации, так что ошибкой их измерения можно пренебречь. Координаты опорных точек (x, y) на преобразуемом изображении, как правило, указываются оператором с возможной последующей автоматизированной подгонкой. Это означает, что точность измерения величин x и y не может превышать одного пиксела, то есть предельная ошибка их измерений $\bar{\varepsilon} \geq 1$ и устанавливается оператором.

Структура преобразования (3.1) зависит от того, какой информацией мы располагаем относительно характера геометрических искажений в исходном изображении. Формирование спутниковых изображений сопровождается различного рода нелинейными искажениями (см., например [55]), связанными с особенностями регистрирующей камеры (например, оптической дисторсией линзы), неравномерностью движения платформы камеры по орбите и т.п. Поэтому геометрическая коррекция изображений в таком случае возможна только с помощью нелинейной функции.

На практике для аппроксимации преобразующей зависимости используют (см. [79, 97, 118]) полиномы двух переменных

$$\begin{aligned} x &= \sum_{i=0}^m \sum_{j=0}^{m-i} a_{ij} u^i v^j, \\ y &= \sum_{i=0}^m \sum_{j=0}^{m-i} b_{ij} u^i v^j. \end{aligned} \tag{3.2}$$

Количество параметров в каждом из выражений зависимости (3.2), а, следовательно, и наименьшее число пар опорных точек, необходимых для оценивания этих параметров, равно $(m+1)(m+2)/2$.

Как правило, для оценивания параметров a_{ij} и b_{ij} преобразующей зависимости (3.2) по таблице опорных точек используется метод наименьших квадратов. Использование для решения этой задачи метода центра неопределенности позволяет получить два следующих преимущества.

Во-первых, при расстановке оператором опорных точек становится возможным автоматическое отслеживание неправильно заданных точек. Обнаружение некорректно установленных опорных точек осуществляется методом, описанным в разделе 2.3. После устранения противоречий построение геометрического преобразования изображения проводится с помощью метода центра неопределенности.

Во-вторых, изображение, полученное после геометрической коррекции, может сопровождаться картой неопределенности

$$U(u, v) = U_x(u, v) + U_y(u, v),$$

$$U_x(u, v) = \bar{x}(u, v) - \underline{x}(u, v),$$

$$U_y(u, v) = \bar{y}(u, v) - \underline{y}(u, v),$$

где $[\underline{x}(u, v), \bar{x}(u, v)]$ и $[\underline{y}(u, v), \bar{y}(u, v)]$ – соответствующие точке (u, v) интервальные оценки координат (x, y) , полученные методом центра неопределенности по таблице опорных точек (см. (2.9)). В некоторых ситуациях удобно использовать не только интегральную карту неопределенности $U(u, v)$, но и ее покоординатные слагаемые $U_x(u, v)$ и $U_y(u, v)$. На практике непрерывную функцию $U(u, v)$ удобнее представлять в дискретизированном виде с некоторым заданным шагом по u и v .

Карта неопределенности может использоваться оператором в качестве руководства к расстановке дополнительных опорных точек с целью понижения пространственной неопределенности в некоторых областях изображения. Однако получаемая карта неопределенности не только служит вспомогательным средством при выполнении операции геометрической коррекции изображения, но и является в некоторой степени самоценным продуктом.

Разработка методов анализа пространственной неопределенности при выполнении любых операций над пространственными данными является одним из приоритетных направлений развития геоинформационных технологий [107, 111, 115, 119, 128, 141, 142]. Такого рода средства обязаны появиться в перспективе во всех геоинформационных системах, так как лица, принимающие решения на основе геоинформационных продуктов, должны располагать информацией об их точности и степени надежности, поскольку в некоторых ситуациях некорректные решения могут повлечь за собой тяжелые последствия.

Пример результатов геометрической коррекции спутникового изображения приведен в приложении 1. Исходное многоканальное изображение получено 4 июля 1999 с помощью орбитального сканера МСУ-СК с борта спутника Ресурс-01. Приложение содержит таблицу опорных точек, исходное и откорректированное изображения с картой неопределенности.

Описанный способ геометрических преобразований и привязки изображений использован в ряде прикладных исследований [11-16, 41, 106]

3.2 Совместная обработка неравноточных инженерно-геодезических измерений

При проведении измерений каждый результат измерения должен сопровождаться сообщением о его погрешности и тем самым указываться интервал его неопределенности [46, 63]. При вероятностном подходе к описанию погрешностей средств и результатов измерений в качестве интервала неопределенности используется доверительный интервал с заданным уровнем доверительной вероятности. Как уже упоминалось в разделе 2.1, долгое время считалось, что случайные погрешности приборов или результатов измерений распределены нормально. Однако по мере накопления данных о фактических распределениях погрешностей стало очевидным, что они весьма разнообразны и очень часто далеки от нормального. Это большое разнообразие законов распределения погрешностей обуславливает практическую

сложность определения доверительных значений погрешностей. В этой ситуации нестатистический подход к интервальному оцениванию погрешностей измерений может служить одним из способов, позволяющих преодолеть эти сложности.

Задачи инженерной геодезии состоят в оценке истинных значений и указании погрешностей оценок тех или иных геометрических характеристик инженерных объектов по совокупности разнородных и неравноточных измерений. Основным математическим инструментом построения такого рода оценок является обобщенный метод наименьших квадратов [47, 104].

В настоящем разделе описывается опирающаяся на метод центра неопределенности технология расчета интервалов неопределенности плановых координат стоек опор линий электропередач (ЛЭП) по измерениям различной точности, выполненная в рамках работ по межеванию земельных участков под опорами ЛЭП. В ходе работ предусматривалось определение координат с указанием оценок неопределенности, для угловых точек основания $A_i B_i C_i D_i$ каждой опоры ЛЭП (см. рисунок 3.1) в местной прямоугольной системе координат. Основание каждой i -й опоры $A_i B_i C_i D_i$ представляет собой прямоугольник с известными размерами, определяемыми типом опоры.

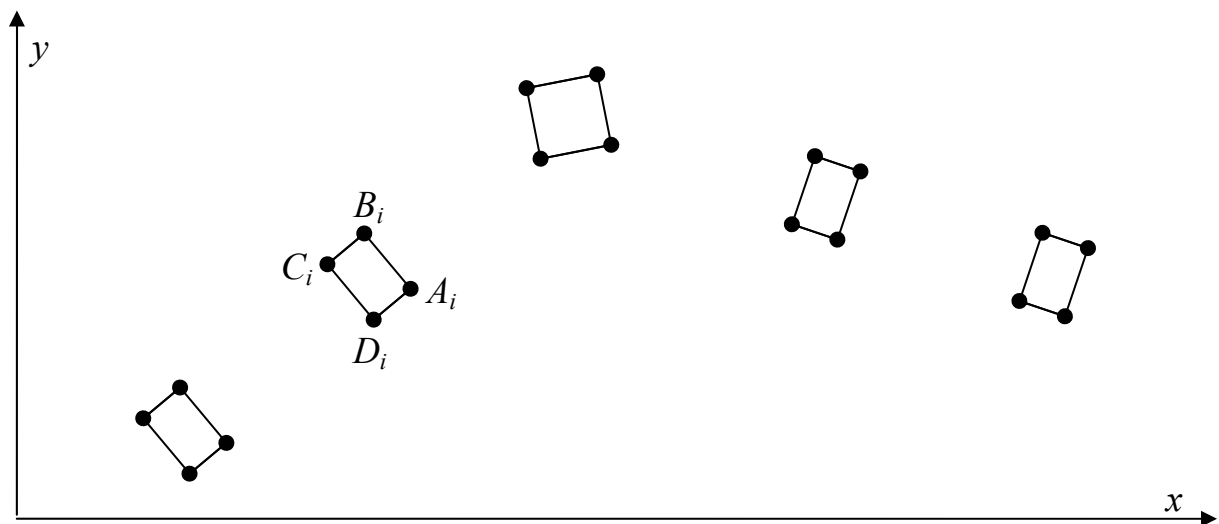


Рис. 3.1 Схема участка линии электропередач в прямоугольной системе координат.

Основным инструментом определения плановых координат опор ЛЭП являлся измерительный комплекс Trimble GPS 4600 SL, обеспечивающий точность измерения плановых координат 2 см.

Условия съемки не всегда допускали проведение прямых измерений. Поэтому в некоторых ситуациях приходилось прибегать к косвенным измерениям, используя дополнительные средства измерений (рулетка, тахеометр и т.д.). В зависимости от конкретных условий на местности использовалась одна из следующих схем косвенных измерений.

Схема косвенных измерений 1. При невозможности прямого измерения с помощью GPS-комплекса координат одной из точек A или B (см. рисунок 3.2) прямоугольники неопределенности их координат $R(A) = [\underline{x}_A, \bar{x}_A] \times [\underline{y}_A, \bar{y}_A]$ и $R(B) = [\underline{x}_B, \bar{x}_B] \times [\underline{y}_B, \bar{y}_B]$ отыскиваются по совокупности следующей информации. GPS-измерения координат производятся в точках R и S , которые выносятся от точек A и B в направлениях $-\overrightarrow{AB}$ и \overrightarrow{AB} на расстояния AR и BS соответственно. Расстояния AR и BS измеряются с помощью геодезической рулетки. Направления \overrightarrow{AR} и \overrightarrow{BS} откладываются либо с помощью тахеометра, либо «на глаз». Для каждого вида измерений известны предельные ошибки:

- 1) ошибка измерения размеров основания опоры ε_p ,
- 2) ошибка измерения координат точек R и S ε_g ,
- 3) ошибка измерения расстояний с помощью рулетки ε_r ,
- 4) предельное отклонение точек A и B от прямой R и S ε_l ,
- 5) ошибка угловых измерений ε_a .

Указанные знания и данные позволяют сформировать следующую систему ограничений:

$$AB^{u3M} - \varepsilon_p \leq \text{dist}(A, B) \leq AB^{u3M} + \varepsilon_p, \quad (3.3)$$

$$AR^{u3M} - \varepsilon_r \leq \text{dist}(A, R) \leq AR^{u3M} + \varepsilon_r, \quad (3.4)$$

$$BS^{u3M} - \varepsilon_r \leq \text{dist}(B, S) \leq BS^{u3M} + \varepsilon_r, \quad (3.5)$$

$$x_R^{uzM} - \varepsilon_g \leq x_R \leq x_R^{uzM} + \varepsilon_g, \quad (3.6)$$

$$y_R^{uzM} - \varepsilon_g \leq y_R \leq y_R^{uzM} + \varepsilon_g, \quad (3.7)$$

$$x_S^{uzM} - \varepsilon_g \leq x_S \leq x_S^{uzM} + \varepsilon_g, \quad (3.8)$$

$$y_S^{uzM} - \varepsilon_g \leq y_S \leq y_S^{uzM} + \varepsilon_g, \quad (3.9)$$

$$-\varepsilon_l \leq \text{line}_{RS}(A) \leq \varepsilon_l, \quad (3.10)$$

$$-\varepsilon_l \leq \text{line}_{RS}(B) \leq \varepsilon_l, \quad (3.11)$$

$$\cos \varepsilon_a \leq \cos \angle(\overrightarrow{BA}, \overrightarrow{AR}) \leq \cos 0, \quad (3.12)$$

$$\cos \varepsilon_a \leq \cos \angle(\overrightarrow{AB}, \overrightarrow{BS}) \leq \cos 0, \quad (3.13)$$

где индекс uzM обозначает значения соответствующих величин, полученные при измерениях; $dist(U, V) = \sqrt{(x_U - x_V)^2 + (y_U - y_V)^2}$ – расстояние между точками U и V ; $line_{UV}(P) = ax_P + by_P + c$, $a = (y_V - y_U)/d$, $b = (x_U - x_V)/d$, $c = x_V y_U - x_U y_V / d$, $d = \sqrt{(y_V - y_U)^2 + (x_U - x_V)^2}$ – значение канонического уравнения прямой, проходящей через точки U и V , в точке P ; и, наконец, $\cos \angle(\overrightarrow{UV}, \overrightarrow{PQ}) = \frac{\overrightarrow{UV} \cdot \overrightarrow{PQ}}{dist(U, V) \cdot dist(P, Q)}$.

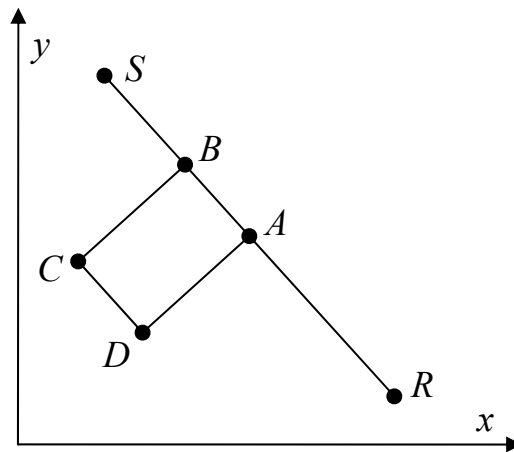


Рис. 3.2 Чертеж к схеме косвенных измерений 1.

Ограничения (3.3)-(3.5) соответствуют измерениям рулеткой, (3.6)-(3.9) – GPS-измерениям, (3.12)-(3.13) – измерениям углов, а (3.10)-(3.11) опреде-

ляют «коридор», в который укладываются точки A, B, R, S . Сформировать и вовлечь в вычисления две последние группы ограничений одновременно удавалось не всегда, поскольку, они дублируют друг друга в некотором смысле, и соответствующие им измерения редко проводились одновременно для одной и той же опоры.

В случаях, когда вынос точек R и S в противоположных направлениях невозможен, использовалась следующая, несколько видоизмененная, схема измерений.

Схема косвенных измерений 2. Вторая схема измерений отличается от первой только тем, что точки R и S , в которых производятся непосредственные GPS-измерения координат, выносятся от точки A в одном и том же направлении $-\overrightarrow{AB}$ на расстояния AR и AS соответственно (см. рисунок 3.3). Естественно, вынос точек R и S , при необходимости может производиться и в направлении \overrightarrow{AB} . Прочие приемы измерений (расстояний AR и AS , направлений \overrightarrow{AR} и \overrightarrow{AS}) и предельные ошибки измерений те же, что и в схеме 1.

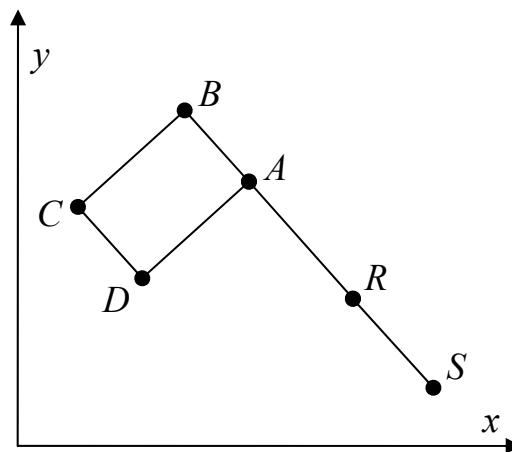


Рис. 3.3 Чертеж к схеме косвенных измерений 2.

Система ограничений в этой схеме приобретает следующий вид.

$$AB^{u_{3M}} - \varepsilon_p \leq \text{dist}(A, B) \leq AB^{u_{3M}} + \varepsilon_p, \quad (3.14)$$

$$AR^{u_{3M}} - \varepsilon_r \leq \text{dist}(A, R) \leq AR^{u_{3M}} + \varepsilon_r, \quad (3.15)$$

$$AS^{u_{3M}} - \varepsilon_r \leq \text{dist}(A, S) \leq AS^{u_{3M}} + \varepsilon_r, \quad (3.16)$$

$$BR^{u_{3M}} - \varepsilon_r \leq \text{dist}(B, R) \leq BR^{u_{3M}} + \varepsilon_r, \quad (3.17)$$

$$BS^{u_{3M}} - \varepsilon_r \leq \text{dist}(B, S) \leq BS^{u_{3M}} + \varepsilon_r, \quad (3.18)$$

$$x_R^{u_{3M}} - \varepsilon_g \leq x_R \leq x_R^{u_{3M}} + \varepsilon_g, \quad (3.19)$$

$$y_R^{u_{3M}} - \varepsilon_g \leq y_R \leq y_R^{u_{3M}} + \varepsilon_g, \quad (3.20)$$

$$x_S^{u_{3M}} - \varepsilon_g \leq x_S \leq x_S^{u_{3M}} + \varepsilon_g, \quad (3.21)$$

$$y_S^{u_{3M}} - \varepsilon_g \leq y_S \leq y_S^{u_{3M}} + \varepsilon_g, \quad (3.22)$$

$$- \varepsilon_l \leq \text{line}_{RS}(A) \leq \varepsilon_l, \quad (3.23)$$

$$- \varepsilon_l \leq \text{line}_{RS}(B) \leq \varepsilon_l, \quad (3.24)$$

$$\cos \varepsilon_a \leq \cos \angle(\overrightarrow{BA}, \overrightarrow{AR}) \leq \cos 0, \quad (3.25)$$

$$\cos \varepsilon_a \leq \cos \angle(\overrightarrow{AB}, \overrightarrow{AS}) \leq \cos 0. \quad (3.26)$$

Замечание относительно одновременного присутствия в системе ограничений последних двух групп ((3.23)-(3.24) и (3.25)-(3.26)), сделанное при изложении первой схемы измерений, в этом случае также остается справедливым.

Нижние (верхние) границы интервалов неопределенности координат точек A и B возможно отыскать путем решения задач

$$x_A \rightarrow \min(\max), \quad (3.27)$$

$$y_A \rightarrow \min(\max), \quad (3.28)$$

$$x_B \rightarrow \min(\max), \quad (3.29)$$

$$y_B \rightarrow \min(\max) \quad (3.30)$$

при ограничениях (3.3)-(3.13) или (3.14)-(3.26), поставляемых первой или второй схемой косвенных измерений соответственно.

Вычисленные границы интервалов неопределенности координат точек A и B наряду с данными о размерах основания опоры позволяют сформировать систему ограничений для задач определения интервалов неопределенности координат точек C и D :

$$x_C \rightarrow \min(\max), \quad (3.31)$$

$$y_C \rightarrow \min(\max), \quad (3.32)$$

$$x_D \rightarrow \min(\max), \quad (3.33)$$

$$y_D \rightarrow \min(\max). \quad (3.34)$$

$$\underline{x}_A \leq x_A \leq \bar{x}_A, \quad (3.35)$$

$$\underline{y}_A \leq y_A \leq \bar{y}_A, \quad (3.36)$$

$$\underline{x}_B \leq x_B \leq \bar{x}_B, \quad (3.37)$$

$$\underline{y}_B \leq y_B \leq \bar{y}_B, \quad (3.38)$$

$$AB^{u_{3M}} - \varepsilon_p \leq \text{dist}(A, B) \leq AB^{u_{3M}} + \varepsilon_p, \quad (3.39)$$

$$BC^{u_{3M}} - \varepsilon_p \leq \text{dist}(B, C) \leq BC^{u_{3M}} + \varepsilon_p, \quad (3.40)$$

$$CD^{u_{3M}} - \varepsilon_p \leq \text{dist}(C, D) \leq CD^{u_{3M}} + \varepsilon_p, \quad (3.41)$$

$$DA^{u_{3M}} - \varepsilon_p \leq \text{dist}(D, A) \leq DA^{u_{3M}} + \varepsilon_p, \quad (3.42)$$

$$AC^{u_{3M}} - \varepsilon_p \leq \text{dist}(AC) \leq AC^{u_{3M}} + \varepsilon_p, \quad (3.43)$$

$$BD^{u_{3M}} - \varepsilon_p \leq \text{dist}(B, D) \leq BD^{u_{3M}} + \varepsilon_p. \quad (3.44)$$

Следует заметить, что сформулированные задачи относятся к задачам нелинейного программирования в отличие от большинства ранее рассмотренных случаев, в которых идеи метода центра неопределенности приводили нас к задачам линейного программирования.

Пример результатов расчета интервалов неопределенности для координат стоек одной из опор согласно схеме косвенных измерений 1 приведен в приложении 4. В приложении 3 приведена схема ЛЭП, в состав которой входит описываемая опора.

Выводы по главе 3

1. Использование метода выявления выбросов и метода центра неопределенности для построения трансформирующей зависимости по опорным точкам при решении задачи геометрического преобразования и привязки изображений позволяет 1) устранять грубые промахи оператора при расстановке опорных точек и 2) производить обоснованный выбор вида зависимости, не вступающего в противоречие с системой опорных точек, установленных с заданной точностью.

2. Возможность вычисления интервальных оценок метода центра неопределенности в заданной точке изображения, подвергнутого геометрическому преобразованию, позволяет сопровождать результирующее изображение картой позиционной неопределенности. Карта неопределенности может использоваться как для внесения дополнений и изменений в систему опорных точек геометрического преобразования, так и для последующего точностного анализа геоинформационных продуктов, построенных на основе преобразованного изображения.

3. При обработке неравноточных инженерно-геодезических измерений с известной предельной погрешностью каждого вида измерений отношения между измеренными величинами могут формулироваться в виде системы неравенств, что позволяет с помощью метода центра неопределенности находить интервалы неопределенности для искомых величин.

ЗАКЛЮЧЕНИЕ

Основные результаты исследований, составившие содержание диссертационной работы, могут быть сформулированы следующим образом.

1. Предложена концептуальная теоретико-множественная модель процесса эмпирического моделирования при интервально заданной нестатистической ошибке.

2. Построена логическая схема процесса эмпирического моделирования. В составе логической схемы выделен ряд информационных задач и указаны необходимые для их решения математические модели и методы.

3. Предложены способы использования метода центра неопределенности для решения в условиях наблюдений без выбросов задач построения и анализа зависимостей, включающих наряду с количественными и качественные входные переменные, а также задач обработки данных, содержащих временные тренды, в том числе линейные и кусочно-линейные.

4. Разработан метод обработки противоречивых наблюдений, позволяющий выявлять выбросы путем поиска минимальных коэффициентов увеличения предельной ошибки наблюдений-выбросов, при которых множество неопределенности является непустым.

5. Разработана схема статистических испытаний, позволяющая сравнить статистические и нестатистические методы построения зависимостей в условиях наблюдений без выбросов. На основе разработанной схемы испытаний проведен сравнительный анализ МЦН и ММП, а также МЦН и МНК. Результаты сравнительного анализа позволяют производить обоснованный выбор процедур построения и анализа зависимостей в соответствии с условиями наблюдений.

6. Разработанные нестатистические методы выявления выбросов и оценивания параметров зависимостей апробированы на реальных данных при построении трансформирующей зависимости по опорным точкам в задаче геометрического преобразования и привязки изображений, а также при обработке неравноточных косвенных инженерно-геодезических измерений.

Результаты работы переданы для использования в ФГУП «Алтайский институт мониторинга земель и экосистем» и Лабораторию обработки изображений физико-технического факультета Алтайского государственного университета. Материалы работы используются в учебном процессе математического факультета Алтайского государственного университета при подготовке студентов по специальности «Прикладная математика».

ЛИТЕРАТУРА

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
2. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Исследование зависимостей. – М.: Финансы и статистика, 1985. – 488 с.
3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. – М.: Финансы и статистика, 1983. – 471 с.
4. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М.: ЮНИТИ-ДАНА, 2001. – 656 с.
5. Алгазин А.И., Шойхет Я.Н., Кисилев В.И., Оскорбин Н.М. Оценка риска радиационного заражения населения Алтайского края // Семипалатинский полигон – Алтай. – 1997. – №8.
6. Алгоритмы и программы восстановления зависимостей / Под ред. В.Н. Вапника. – М.: Наука, Гл. ред. физ.-мат. лит., 1984. – 816 с.
7. Алефельд Г.Ш., Херцбергер Ю. Введение в интервальные вычисления. – М.: Мир, 1987. – 370 с.
8. Алимов Ю.И. Альтернатива методу математической статистики. – М.: Знание, 1980. – 64 с.
9. Алимов Ю.И., Кравцов Ю.А. Является ли вероятность «нормальной» физической величиной? // УФН. – 1992. – Т.162. – №7. – С. 149-182.
10. Алтунин А.Е., Семухин М.В. Модели и алгоритмы принятия решений в нечетких условиях. – Тюмень: Издательство Тюменского гос. ун-та, 2000. – 352 с.
11. Байкалова Т.В., Давыдов Е.С., Дубина И.Н., Евтюшкин А.В., Жилин С.И., Оскорбин Н.М., Поляков Ю.А., Юшаков В.Н. Комплекс цифрового картографирования для решения задач экологии и природопользования. // Региональные проблемы информатизации: Труды Респ. науч.-тех. конф. Барнаул: Изд-во АлтГТУ, 1995. С. 54-55.

12. Байкалова Т.В., Евтюшкин А.В., Жилин С.И., Казанцев К.В., Редькин А.Г. Дешифрирование ледниковых форм рельефа на радиолокационных изображениях // ИНТЕРКАРТО 4: ГИС для оптимизации природопользования в целях устойчивого развития территорий: Материалы международной конференции. – Барнаул: Изд-во Алт. гос. ун-та, 1998. С. 226-231.
13. Байкалова Т.В., Евтюшкин А.В., Жилин С.И., Оскорбин Н.М., Юшаков В.Н. Автоматизация построения экологических карт. // Методы дистанционного зондирования и ГИС-технологии для контроля и диагностики состояния окружающей среды: Тезисы докладов 3-й Международной конференции. Москва: МИИГАиК, 1996.
14. Байкалова Т.В., Евтюшкин А.В., Жилин С.И., Юшаков В.Н. Информационные технологии автоматизации построения экологических карт. // Проблемы предотвращения деградации земель Западной Сибири и осуществление государственного контроля за их использованием и охраной: Сб. науч. тр. – Барнаул: Минсельхозпрод РФ, 1997. С. 108-113
15. Байкалова Т.В., Евтюшкин А.В., Жилин С.И., Юшаков В.Н. Мониторинг состояния земных покровов Алтая по данным ИСЗ "NOAA". // Известия Алтайского государственного университета. – 1998. №1. – С. 49-52.
16. Байкалова Т.В., Евтюшкин А.В., Жилин С.И., Юшаков В.Н. Пространственно-временной мониторинг состояния земных покровов Алтая по данным ИСЗ "NOAA". // Историческая и современная картография в развитии Алтайского региона: Тезисы докладов международной научно-практической конференции. Барнаул: Изд-во Алт. гос. ун-та, 1997. С. 70-71.
17. Белов В.М., Суханов В.А., Гузеев В.В., Унгер Ф.Г. Оценивание параметров линейных физико-химических зависимостей прямоугольником метода центра неопределенности // Изв. вузов. Физика, 1991. – №8. – С. 35-45.

18. Белов В.М., Суханов В.А., Унгер Ф.Г. Аппроксимация эллипсом множества неопределенности параметров зависимостей, сводящихся к линейным. Томск, 1990. – 28 с. – (Препр. / АН СССР. Сиб. отд-ние. Ин-т химии нефти; №45).
19. Белов В.М., Суханов В.А., Унгер Ф.Г. Метод центра неопределенности при расчете линейных градуировочных графиков и метрологических характеристик результатов химического анализа. Томск, 1989. – 34 с. – (Препр. / АН СССР. Сиб. отд-ние. Ин-т химии нефти; №59).
20. Белов В.М., Суханов В.А., Унгер Ф.Г. Обзор основных статистических методов определения параметров аппроксимирующих функций. Препринт №46, ТНЦ СО АН СССР, Томск, 1990. – 34 с.
21. Белов В.М., Суханов В.А., Унгер Ф.Г. Оценка физико-химических величин методом центра неопределенности. Томск, 1990. – 45 с. – (Препр. / АН СССР. Сиб. отд-ние. Ин-т химии нефти; №16).
22. Белов В.М., Суханов В.А., Унгер Ф.Г. Теоретические и прикладные аспекты метода центра неопределенности. – Новосибирск: Наука. Сибирская издательская фирма РАН, 1995. – 144 с.
23. Болдырев В.И., Спивак С.И. Чебышевские приближения для кинетической модели с дробно-линейной зависимостью от параметров // Математические проблемы химии. – Новосибирск: ВЦ СО АН СССР, 1973. – Ч. 2. – С. 58-65.
24. Болч Б., Хуань К.Дж. Многомерные статистические методы для экономики. – М.: Статистика, 1979. – 317 с.
25. Бородюк В.П. Комментарий I к статье А.П. Воцинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке". – Заводская лаборатория. – 1990. – Т.56. – №.7. – С. 81-83.
26. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 447 с.

27. Вошинин А.П. Интервальный анализ данных: развитие и перспективы. // Заводская лаборатория. Диагностика материалов. – 2002. – Т.68. – №.1. – С. 118-126.
28. Вошинин А.П. Метод анализа данных с интервальными ошибками в задачах проверки гипотез и оценивания параметров неявных линейно параметризованных функций. // Заводская лаборатория. Диагностика материалов. – 2000. – Т.66. – №3. – С. 51-65.
29. Вошинин А.П. Решение задач оптимизации по интервальным оценкам критерия. // Заводская лаборатория. – 1987. – Т.54. – №7. – С. 45-48.
30. Вошинин А.П., Бочков А.Ф., Сотиров Г.Р. Метод анализа данных при интервальной нестатистической ошибке // Заводская лаборатория. – 1990. – Т.56. – № 7. – С. 76-81.
31. Вошинин А.П., Сотиров Г.Р. Оптимизация в условиях неопределенности. – М., София: Изд-во МЭИ, Техника, 1989. – 224 с.
32. Гермейер Ю.Б. Введение в теорию исследования операций. – М.: Наука, 1971. – 383 с.
33. Гурин Л.С. О состоятельности оценок метода наименьших квадратов. // Математическое обеспечение космических экспериментов. – М.: Наука, 1978.
34. Демиденко Е.З. Комментарий II к статье А.П. Вошинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке". – Заводская лаборатория. – 1990. – Т.56. – №.7. – С. 83-84.
35. Деревницкий Д.П., Фрадков А.Л. Прикладная теория дискретных адаптивных систем управления. – М.: Наука, 1981.
36. Доугерти К. Введение в эконометрику. – М.: ИФРА-М, 1999. – XIV, 402 с.
37. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. – М.: Финансы и статистика, 1987. – 717 с.

38. Еремин И.И. Противоречивые модели оптимального планирования. – М.: Наука. Гл. ред. физ.-мат. лит., 1988. – 160 с.
39. Еремин И.И., Астафьев Н.Н. Введение в теорию линейного и выпуклого программирования. – М.: Наука. Гл. ред. физ.-мат. лит., 1976. – 192 с.
40. Ерохин Г.И., Камышников А.И., Оскорбин Н.М. Обработка больших баз данных и анализ зависимостей методами линейного программирования. // Управление, математическое моделирование и оптимизация на базе ПЭВМ: Межвуз. сб. науч. раб. – Барнаул: Изд-во Алт. гос. ун-та, 1993. – С. 143-146.
41. Жилин С.И. Решение задачи трансформации векторных и растровых изображений с использованием методов линейного программирования. // Проблемы предотвращения деградации земель Западной Сибири и осуществление государственного контроля за их использованием и охраной: Сб. науч. тр. – Барнаул: Минсельхозпрод РФ, 1997. С. 152-156.
42. Жилин С.И. Эксперименты по нестатистическому оцениванию параметров эмпирических зависимостей // Материалы шестой краевой конференции по математике МАК-2003. – Барнаул: Изд-во Алт. гос. ун-та, 2003. С. 56–57.
43. Жилин С.И. Эксперименты по оцениванию параметров эмпирической зависимости методом наименьших квадратов и методом центра неопределенности. // Известия Алтайского государственного университета. – 2003. – №1. – С. 24-27.
44. Жилин С.И., Паршуков Л.Н. Математическая модель коррекции площадей элементов векторного топологического полигонального покрытия по априорным данным. // Материалы первой краевой конференции по математике. Барнаул: Изд-во Алт. гос. ун-та, 1998. С. 34-35.
45. Ибрагимов И.А. Максимального правдоподобия метод // Математическая энциклопедия. М., 1982. – Т. 3. – С. 483-484.
46. Иванников Д.А., Фомичев Е.Н. Основы метрологии и организации метрологического контроля [Электронный ресурс] : Учеб. пособие. / Ниже-

городский гос. техн. ун-т. – Н. Новгород, 2001. – Режим доступа: <http://www.nntu.sci-nnov.ru/RUS/VECH/metod/metrology/title.htm>. – Загл. с экрана.

47. Инженерная геодезия. Учеб. для вузов / Е.М. Ключин, М.И. Киселев, Д.Ш. Михелев, В.Д. Фельдман; Под ред. Д.Ш. Михелева. — М.: Высш. шк., 2001. – 464 с.
48. Информационные множества в задаче наблюдения за движением самолета в горизонтальной плоскости / Кумков С.И., Пацко В.С., Пятко С.Г., Решетов В.М., Федотов А.А. // Известия РАН. Теория и системы управления. – 2003. – №4. – С. 51-61.
49. Ицкович И.А., Спивак С.И. Анализ применения методов линейного программирования при построении кинетической модели сложной химической реакции // Управляемые системы. – Новосибирск: Наука. Сиб. отделение, 1970. – Вып. 4-5. – С. 142-147.
50. Калмыков С.А., Шокин Ю.И., Юлдашев З.Х.. Методы интервального анализа. – Новосибирск: Наука, Сибирское отделение, 1986. – 222 с.
51. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. – 1962. – Т.3. – № 5. – С. 701-709.
52. Кендалл М., Стюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 542 с.
53. Коковин Г.А., Титов В.А., Титов А.А., Спивак С.И. Некоторые методологические вопросы математической обработки экспериментальных данных по исследованию равновесий // Математика в химической термодинамике. – Новосибирск: Наука. Сиб. отд-ние, 1980. – С. 50-58.
54. Колеников С. Прикладной эконометрический анализ в статистическом пакете Stata. – М.: Российская экономическая школа, 2000. – 112 с.
55. Кравцова В.И. Космические методы картографирования. – М.: Изд-во МГУ, 1995. – 240 с.

56. Кузнецов В.П. Комментарий VI к статье А.П. Воцинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке". // Заводская лаборатория. – 1990. – Т.56. – №7. – С.90-93.
57. Куржанский А.Б. Задача идентификации – теория гарантированных оценок. // Автоматика и телемеханика. – 1991. – №4. – С. 3-26.
58. Куржанский А.Б. Управление и наблюдение в условиях неопределенности. – М.: Наука, 1977. – 390 с.
59. Куркин О.М., Коробочкин Ю.Б., Шаталов С.А. Минимаксная обработка информации. – М.: Энергоатомиздат, 1990. – 216 с.
60. Легостаева И.Л. Комментарий V к статье А.П. Воцинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. – 1990. – Т.56. – №7. – С.90-93.
61. Лецкий Э.К. Комментарий III к статье А.П. Воцинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. – 1990. – Т.56. – №7. – С.84-86.
62. Люблинский Р.Н., Оскорбин Н.М. Методы декомпозиции при оптимальном управлении непрерывными производствами. – Томск: Изд-во Томск. ун-та, 1979. – 220 с.
63. Метрология и электрорадиоизмерения в телекоммуникационных системах: Уч. для вузов. / Ред. Нефедов В. – М.: Высшая школа, 2001. – 383 с.
64. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия – М.: Финансы и статистика, 1982. – 556 с.
65. Налимов В.В. Применение математической статистики при анализе вещества. – М.: Физматгиз, 1960. – 430 с.
66. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. – Л.: Энергоатомиздат, 1985. – 248 с.
67. Орлов А.И. Комментарий IV к статье А.П. Воцинина, А.Ф. Бочкова, Г.Р. Сотирова "Метод анализа данных при интервальной нестатистической ошибке" // Заводская лаборатория. – 1990. – Т.56. – №7. – С. 86-89.

68. Орлов А.И. Некоторые нерешенные вопросы в области математических методов исследования // Заводская лаборатория. Диагностика материалов. – 2002. – Т.68. – №3. – С. 52-56.
69. Орлов А.И. Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений // Заводская лаборатория. – 1992. – Т.58. – №7. – С. 40-42.
70. Орлов А.И. Об оценивании регрессионного полинома // Заводская лаборатория. – 1994. – Т.60. – №5. – С. 43-47.
71. Орлов А.И. Современная прикладная статистика // Заводская лаборатория. – 1998. – Т.64. – №3. – С. 52-60.
72. Орлов А.И. Устойчивость в социально-экономических моделях. – М.: Наука, 1979. – 296 с.
73. Орлов А.И. Часто ли распределение результатов наблюдений является нормальным? // Заводская лаборатория. – 1991. – Т.57. – №7. – С. 64-66.
74. Орлов А.И. Термины и определения в области вероятностно-статистических методов // Заводская лаборатория. – 1999. – Т.65. – №7. – С. 46-54.
75. Оскорбин Н.М. Некоторые задачи обработки информации в управляемых системах // Синтез и проектирование многоуровневых иерархических систем. Материалы конференции. – Барнаул: Алтайский государственный университет, 1983.
76. Оскорбин Н.М., Жилин С.И., Дронов С.В. Сравнение статистической и нестатистической оценок параметров эмпирической зависимости. // Известия Алтайского государственного университета. – 1998. – №4. – С. 38-41.
77. Оскорбин Н.М., Максимов А.В., Жилин С.И. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия Алтайского государственного университета. – 1998. – № 1. – С. 35-38.
78. Прохоров А.В. Регрессия // Математическая энциклопедия. М., 1984. – Т.4. – С. 929-931.

79. Прэтт У. Цифровая обработка изображений. – М.: Мир, 1982. – 790 с.
80. Смоляк С.А., Титаренко Б.П. Устойчивые методы оценивания: Статистическая обработка неоднородных совокупностей. – М.: Статистика, 1980. – 208 с.
81. Спивак С.И. Детальный анализ применения методов линейного программирования при определении параметров кинетической модели // Математические проблемы химии. – Новосибирск: ВЦ СО АН СССР, 1975. – Ч. 2. – С. 35-42.
82. Спивак С.И. Информативность эксперимента и проблема неединственности решения обратных задач химической кинетики: Автореф. дис. ... д-ра физ.-мат. наук. – Черноголовка, 1984. – 30 с.
83. Спивак С.И. О неединственности решения задач восстановления констант химической кинетики и констант химических равновесий // Математические проблемы химической термодинамики. – Новосибирск: Наука. Сиб. отд-ние, 1980. – С. 63-72.
84. Спивак С.И., Слинько М.Г., Тимошенко В.И. Оценка значимости влияния измерений на кинетическую модель химической реакции // Математические проблемы химии. – Новосибирск: ВЦ СО АН СССР, 1973. – Ч. 2. – С. 3-9.
85. Спивак С.И., Тимошенко В.И., Слинько М.Г. Планирование методами математического программирования экспериментов для уточнения параметров кинетических моделей // Кинетика-2. II Всесоюз. конф. по кинетике катализ. реакций: Тез. докл. – Новосибирск: ИК СО АН СССР, 1975. – Т.3. – С. 36-42.
86. Спивак С.И., Шмелев А.С. Методологические аспекты определения физико-химических параметров по экспериментальным данным // Математика в химической термодинамике. – Новосибирск: Наука. Сиб. отд-ние, 1980. – С. 84-91.
87. Срагович В.Г. Адаптивное управление. – М.: Наука. Гл. ред. физ.-мат. лит., 1981. – 384 с.

88. Суханов В.А., Белов В.М., Унгер Ф.Г. Теоретико-вероятностные исследования свойств оценок прямоугольника в методе центра неопределенности. Препринт. – Томск: Институт химии нефти СО РАН, 1999. – 20 с.
89. Трусов В.С. Теория эксперимента. Томск: Изд-во Томск. ун-та, 1983. – 184 с.
90. Тутубалин В.Н. Вероятность, компьютеры и обработка результатов экспериментов. // УФН. – 1993. – Т.163. – №7. – С. 93-109.
91. Тутубалин В.Н. Границы применимости (Вероятностно-статистические методы и их возможности). – М.: Знание, 1977. – 64 с.
92. Устойчивые статистические методы оценки данных / Под ред. Р.Л. Лонера, Г.Н. Уилкинсона: Пер. с англ. – М.: Машиностроение, 1984. – 232 с.
93. Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния. – М.: Мир, 1989. – 512 с.
94. Хлебников А.И. О методе центра неопределенности // Журнал аналитической химии. – 1996. – Т. 51. – №3. – С. 347-348.
95. Хлебников А.И. О проблемах использования метода центра неопределенности для обработки экспериментальных данных // Вычислительные технологии. – 1999. – Т.4. – №4. – С. 80-81.
96. Хьюбер П. Робастность в статистике: Пер. с англ. Под ред. И.Г. Журбенко. – М.: Мир, 1984. – 304 с.
97. Цифровая обработка изображений в информационных системах: Учеб. пособие / И.С. Грузман, В.С. Киричук и др. – Новосибирск: Изд-во НГТУ, 2002. – 352 с.
98. Черноузько Ф. Л. Оценивание фазового состояния динамических систем. Метод эллипсоидов. – М.: Наука, 1988. – 320 с.
99. Шарый С.П. Еще раз о внутреннем оценивании множеств решений интервальных линейных систем // Материалы шестой краевой конференции по математике МАК-2003. – Барнаул: Изд-во Алт. ун-та, 2003. С. 73-74.

100. Шарый С.П. Интервальные алгебраические задачи и их численное решение. Дисс. ... докт. физ.-мат. наук. – Новосибирск, 2000. – 327 с.
101. Шокин Ю.И. Интервальный анализ. – Новосибирск: Наука, 1981. – 281 с.
102. Эльясберг П.Е. Об устойчивости оценок точности определения орбит по результатам измерений. – Космические исследования. – 1978. – Т.16. – №5.
103. Эльясберг П.Е., Измерительная информация. Сколько ее нужно, как ее обрабатывать? – М.: Наука, 1983. – 208 с.
104. Яковлев Н.В. Высшая геодезия: Учебник для вузов. – М.: Недра, 1989. – 445 с.
105. Akyildiz Y., Markov S. Curve Fitting and Interpolation of Biological Data Under Uncertainties // Journal of Universal Computer Science. – 1996. – Vol. 2. – №2. – P. 58-69.
106. Baikalova T.V., Dubina I.N., Yevtyushkin A.V., Oskorbin N.M., Polyakov Yu.A., Yushakov V.N., Zhilin S.I. Automation of digital mapping in systems of regional ecological monitoring. // Proceedings of the International Conference "Intercarto 2: GIS for Enviromental Studies and Mapping". Irkutsk, 1996, p. 67-71.
107. Beard K.M., Battenfield B.P., Clapham, S.B. Visualization of Spatial Data Quality. Technical Paper 91-26, National Center for Geographic Information and Analysis, Castine, Maine, 1991. – 54 p.
108. Beaumont O. Solving interval linear systems with linear programming techniques // Linear Algebra and its Applications. – 1998. – Vol. 281. – P. 293-309.
109. Belforte G., Tay T.T. Two New Estimation Algorithms for Linear Models with Unknown but Bounded Measurement Noise // IEEE Transactions on Automatic Control. – 1993. – Vol. 38. – № 8. – P. 1273-1279.
110. Bounding Approaches to System Identification / Milanese M., Norton J., Walter E., editors. – London: Plenum Press, 1996. – 586 p.

111. Chil J.-P., Delfiner P. Geostatistics: Modeling Spatial Uncertainty. – Wiley-Interscience, 1999. – 720 pp.
112. Combettes P.L. Foundations of set-theoretic estimation // Proc. of IEEE. – 1993. – Vol. 81. – №2. – P. 182-208.
113. de Vries D. K., Identification of Model Uncertainty for Control Design. PhD Thesis. – Delft: Delft University of Technology, 1994. – 272 p.
114. Deller J., Nayeri M., Odeh S. Least-square identification with error bounds for real time signal processing and control // Proc. of IEEE. – 1993. – Vol. 81. – №6. – P. 815-849.
115. Ehlschlaeger C.R., Goodchild M.F. Uncertainty in Spatial Data: Defining, Visualizing, and Managing Data Errors // Proc. of Conference GIS/LIS. – Phoenix, 1994. – P. 246-253.
116. Esbensen K.H. Multivariate Data Analysis – in practice. An introduction to multivariate data analysis and experimental design. – CAMO AS, 2000. – 600 p.
117. Fogel D.N., Tinney L.R. Image Registration using Multiquadric Functions, the Finite Element Method, Bivariate Mapping Polynomials and the Thin Plate Spline. Technical Report Number 96-01, National Center for Geographic Information and Analysis, University of California, Santa Barbara, 1996. – 63 p.
118. Fogel E., Huang Y.F. On the value of information in system identification bounded noise case // Automatica. – 1982. – Vol. 18. – P. 229-238.
119. Goodchild M.F. Accuracy of Spatial Databases. Research Initiative 1 Closing Report, National Center for Geographic Information and Analysis, University of California, Santa Barbara, 1992. – 18 p.
120. Jaulin L., Kieffer M., Didrit O., Walter E. Applied Interval Analysis. – Springer-Verlag, Londres, 2001. – 379 p.
121. Kearfott R. B. Rigorous Global Search: Continuous Problems. Dordrecht: Kluwer Academic Publishers, 1996. – 280 p.

122. Kreinovich V. Data Processing Beyond Traditional Statistics: Applications of Interval Computations. A Brief Introduction. // Extended Abstracts of International Workshop on Applications of Interval Computations, El Paso, 1995. – P. 13-21.
123. Kreinovich V., Lakeev A., Rohn J., Kahl P. Computational Complexity and Feasibility of Data Processing and Interval Computations. – Dordrecht: Kluwer, 1997. – 472 pp.
124. Kumkov S.I. Method of Informational Sets in Problems of Filtration, Identification and Control. – Tampere: Tampere International Center for Signal Processing, 2001, p. 33.
125. Martens H., Naes T. Multivariate Calibration. – Wiley, 1989. – XVIII, 420 p.
126. Milanese M., Belforte G. Estimation Theory and Uncertainty Intervals Evaluation in Presence of Unknown But Bounded Errors: Linear Families of Models and Estimators // IEEE Transactions on Automatic Control. – 1982. – Vol. 27. – № 2. – P. 408-414.
127. Moore R.E. Parameter sets for bounded-error data // Mathematics And Computers in Simulation. – 1992. – Vol. 34. – №2. – P. 113-119.
128. Mowrer H.T., Congalton R.G. Quantifying Spatial Uncertainty in Natural Resources: Theory and Applications for GIS and Remote Sensing. – Taylor & Francis, 2000. – 350 p.
129. Nakaya Y., Oishi S. Finding All Solutions of Nonlinear Systems of Equations Using Linear Programming with Guaranteed Accuracy // Journal of Universal Computer Science. – 1998. – Vol 4. – № 2. – P. 171-177.
130. Neumaier A. Interval Methods for Systems of Equations. Cambridge: Cambridge Univ. Press, 1990. – 255 p.
131. Nguyen H.T., Kreinovich V., Tao C.-W. Why 95% and Two Sigma? A Theoretical Justification for an Empirical Measurement Practice // Proc. International Workshop on Intelligent Systems Resolutions, Taipei, 2000. – P. 358-362.
132. Orlov A.I. Interval statistics // Interval Computations. – 1992. – №1. P. 44-52.

133. Pomerantsev A.E., Rodionova O.E., Esbensen K.H. Object status classification with SIC (Simple Interval Calculation) // J. Chemometrics (submitted), 2001.
134. Pomerantsev A.L., Rodionova O.Ye. Application of Simple Interval Calculation Method // Book of Abstracts, 7th Scandinavian Symposium on Chemometrics, Copenhagen, 2001.
135. Pronzato L., Walter E. Minimum-volume ellipsoids containing compact sets: application to parameter bounding // Proc. of IEEE. – 1994. – Vol. 30, – №11. – P. 1731-1739.
136. Pronzato L., Walter E. Volume-optimal inner and outer ellipsoids. In M. Milaneze, J.-P. Norton, H. Piet-Lahanier, and E. Walter, editors, Bounding Approaches to System Identification. – London: Plenum, 1996. – P. 119-138.
137. Pronzato L., Walter E. Minimal-volume ellipsoids // Int. Journal of Adaptive Control and Signal Processing. – 1994. – Vol 8. – P. 15-30.
138. Rodionova O.E., Pomerantsev A.E. SIC (Simple Interval Calculation) – a new approach for linear modeling. // J. Chemometrics (submitted), 2001.
139. Rodionova O.E., Pomerantsev A.E. Simple Interval Calculation – A Method for Linear Modeling // Book of Abstracts, 7th Scandinavian Symposium on Chemometrics, Copenhagen, 2001.
140. Rodionova O.Ye., Pomerantsev A.L. Antioxidants Activity Prediction Using DSC Measurements and SIC Data Processing. // II Conference on Experimental Methods in Physics of Heterogeneous Condensed Media, Barnaul, 2001. – P. 239-246.
141. Wenzhong S., Ehlers M., Tempfli K. Modeling and Visualizing Uncertainties in Multi-Data-Based Spatial Analysis. // Proc. of European GIS Conference, 1994. P. 454-464.
142. Wilke F.Z., Franciosi B.R.T., Oliveira P.W., Claudio D.M. Modeling the Measurement Uncertainty by Intervals // Journal of Universal Computer Science. – 1998. – Vol 4. – № 1. – P. 82-88.

143. Yamamura K. Interval solution of nonlinear equations using linear programming // Proc. of IEEE International Symposium on Circuits and Systems, 1997. – P. 837-840.

**Пример геометрического преобразования
и привязки изображения**

**Обзорный вариант исходного изображения
с нанесенными опорными точками**
(МСУ-СК, 04.07.1999,
оригинальный размер: 5854×5737 пикселей,
разрешение: 40 м)

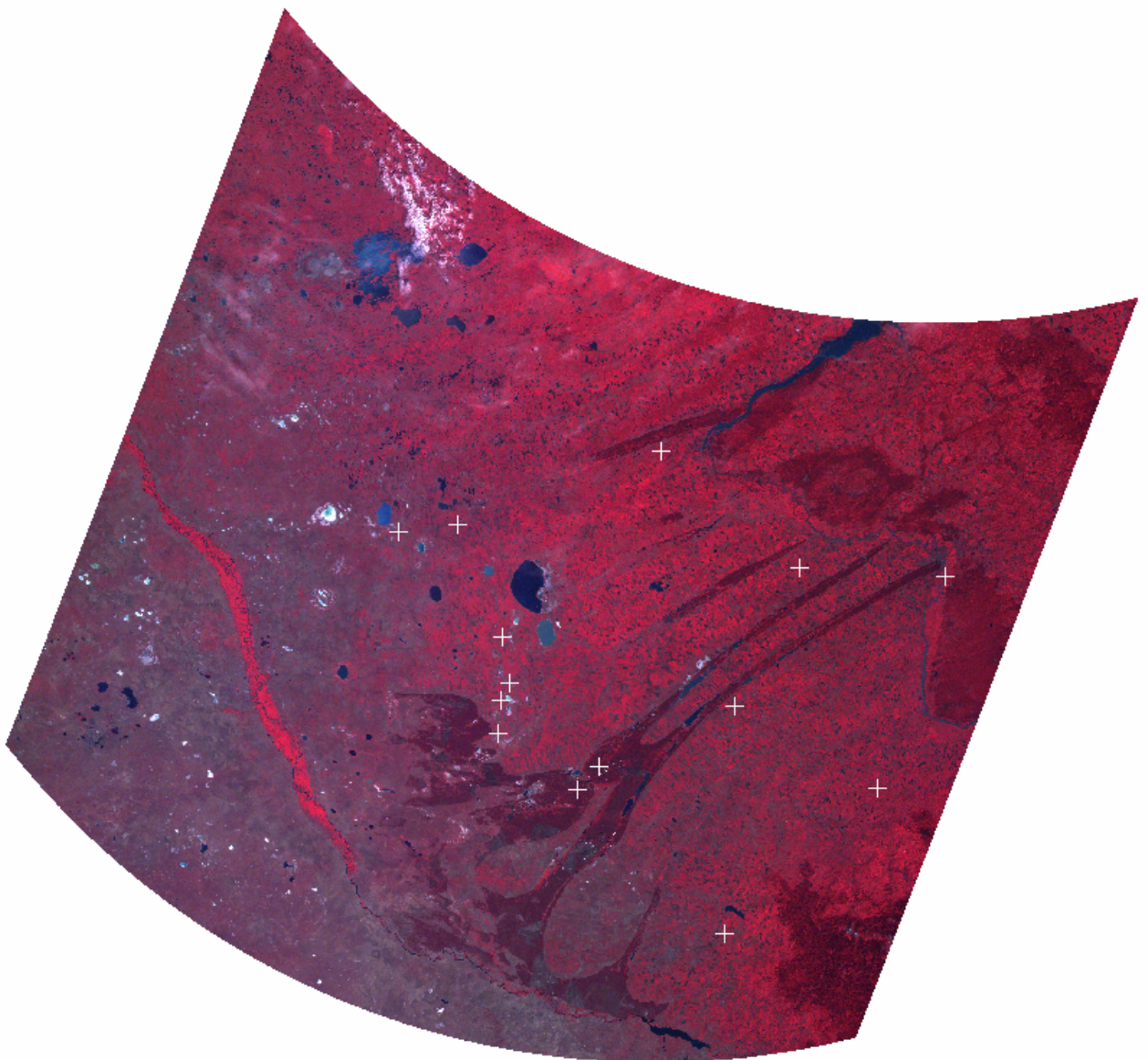


Таблица опорных точек*

Номер точки	Координаты на исходном изображении		Координаты на эталонном изображении	
	u	v	x	y
1	2935,16	3072,66	14486,30	5991,49
2	2045,81	2745,70	14349,30	5927,55
3	1795,21	2714,94	14309,60	5919,30
4	3518,59	2560,89	14586,30	5917,89
5	4147,04	2518,88	14687,20	5919,87
6	2248,77	2268,73	14388,00	5854,12
7	2268,31	2062,65	14393,80	5821,74
8	2235,14	1987,57	14389,70	5809,22
9	3238,83	1966,60	14550,00	5819,64
10	2220,98	1843,67	14389,40	5786,27
11	2690,13	1673,81	14466,60	5765,55
12	2563,95	1605,97	14447,30	5753,20
13	3753,66	1621,51	14637,00	5772,36
14	3199,48	985,54	14557,60	5664,14

* Предельная ошибка измерения координат на исходном изображении $\varepsilon = 1,5$ пиксела.

Вид трансформирующей функции

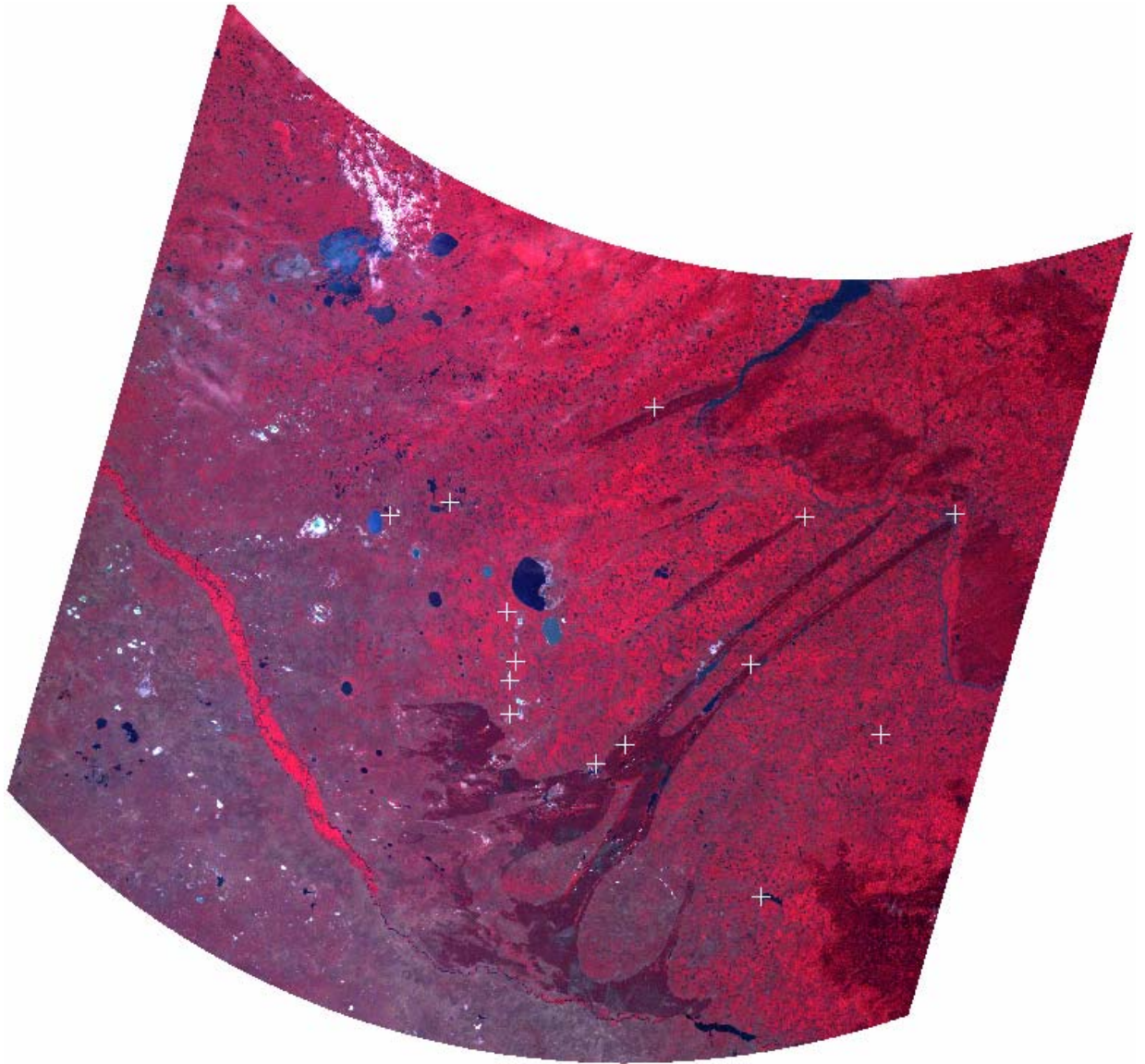
$$x = a_{00} + a_{10}u + a_{01}v + a_{11}uv + a_{20}u^2 + a_{02}v^2,$$

$$y = b_{00} + b_{10}u + b_{01}v + b_{11}uv + b_{20}u^2 + b_{02}v^2.$$

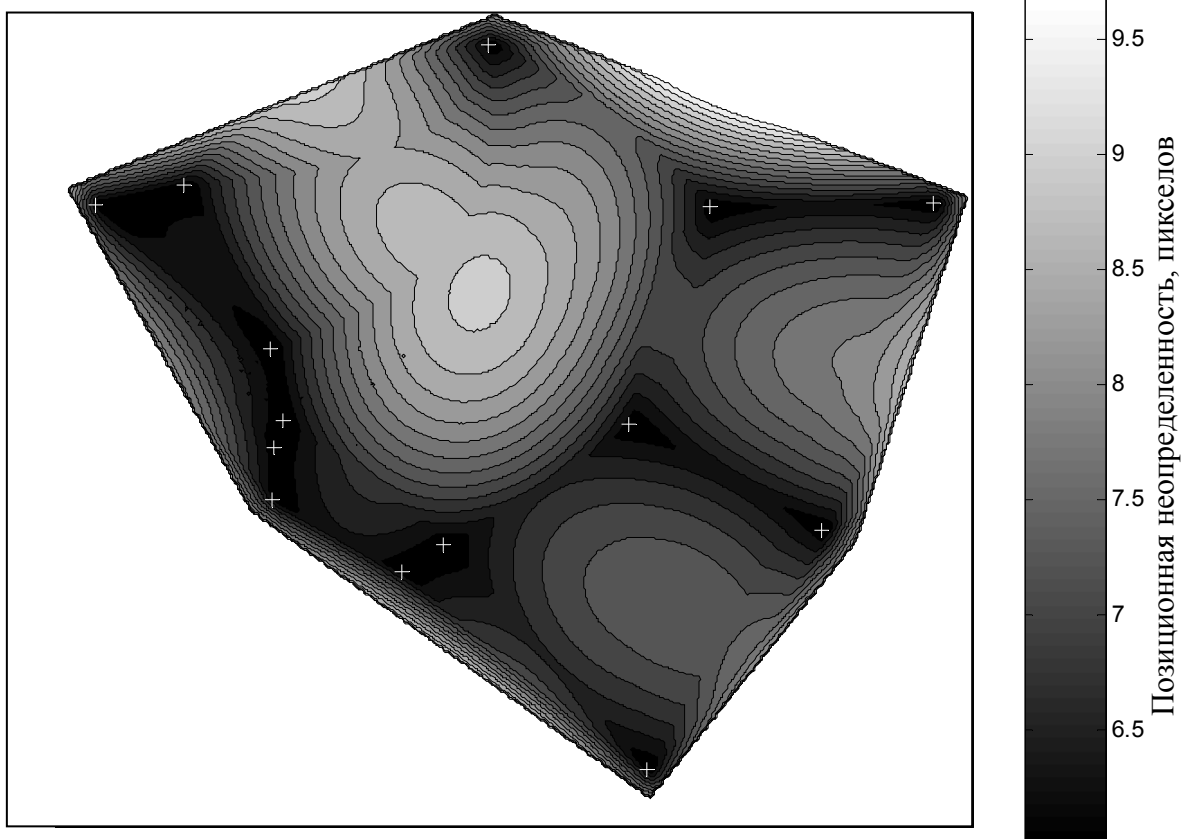
Оценки коэффициентов трансформирующей функции

Индекс коэффициента	\hat{a}	\hat{b}
00	13549,18263943	5511,70942290
10	6,31108517	-0,53720265
01	0,64043523	6,58753970
11	-0,00001521	0,00012764
20	-0,00007649	-0,00004857
02	-0,00011458	-0,00094673

**Обзорный вариант преобразованного изображения
с нанесенными опорными точками**

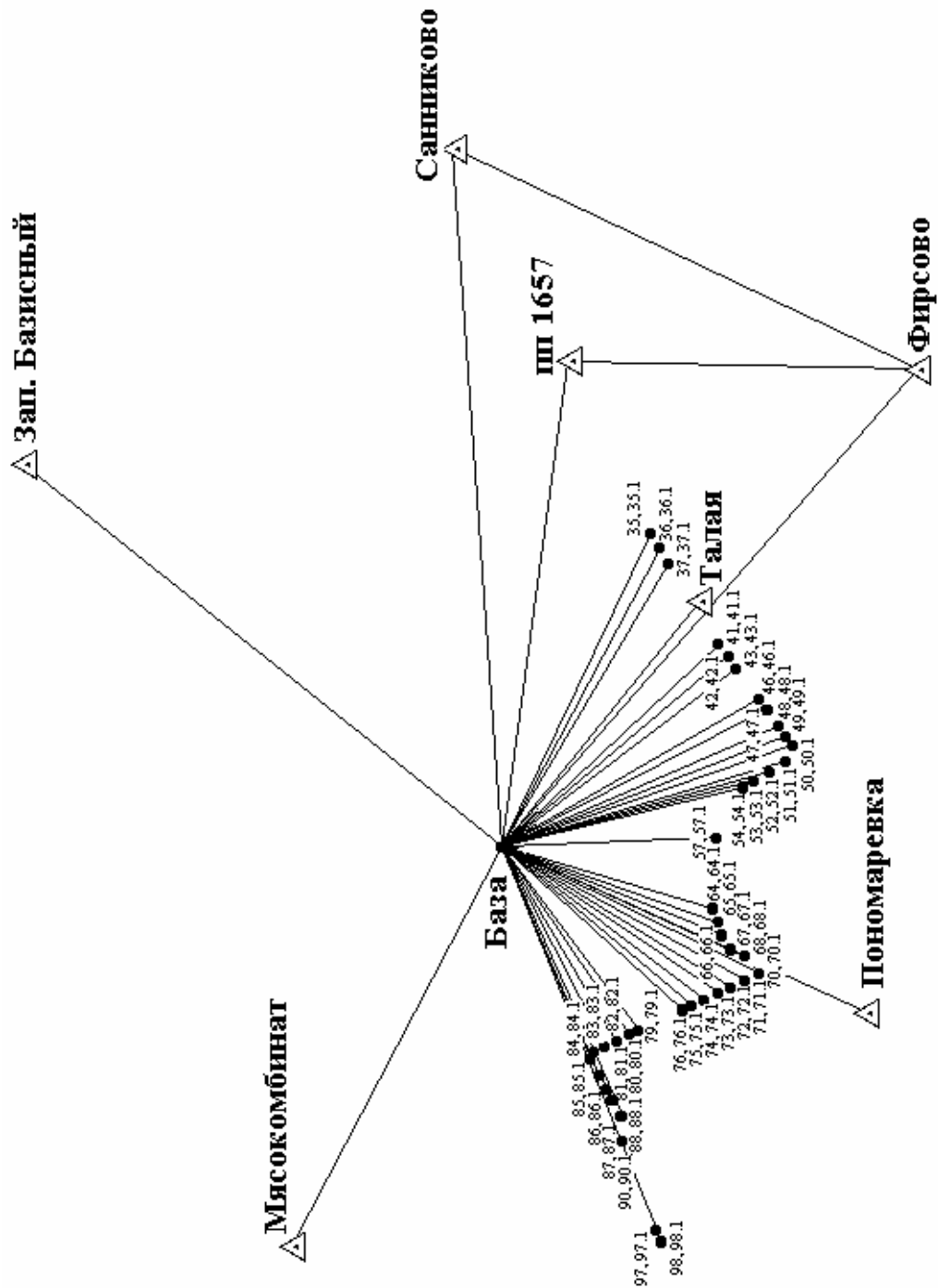


**Фрагмент карты суммарной неопределенности
с нанесенными опорными точками**



Пример результатов расчета интервалов неопределенности
для координат стоек опоры ЛЭП

Схема ЛЭП на территории г. Барнаула
с прилегающими триангуляционными пунктами
и точкой размещения базового GPS-приемника



**Исходные данные и результаты расчета
интервалов неопределенности
для координат стоек опоры №67**

Исходные данные по опоре

1. GPS-измерения координат точек в местной системе координат (в метрах):

$$R = (7788,905; 7276,036),$$

$$S = (7769,764; 7259,235).$$

Предельная ошибка измерений:

$$\varepsilon_g = 0,02 \text{ м.}$$

2. Расстояния, измеренные рулеткой:

$$AR = 9,05 \text{ м}$$

$$BS = 8,22 \text{ м.}$$

Предельная ошибка измерений:

$$\varepsilon_r = 0,05 \text{ м.}$$

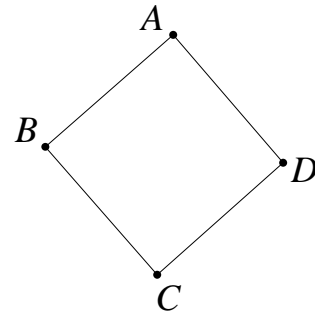


Схема косвенных измерений
для опоры №67

3. Размеры основания опоры (тип У39):

$$AB = BC = CD = DA = 8,10 \text{ м.}$$

Предельная ошибка измерений:

$$\varepsilon_p = 0,005 \text{ м}$$

4. Предельная угловая ошибка выноса точек R и S в створе AB «на глаз»:

$$\varepsilon_a = 1'.$$

Расчетные интервалы неопределенности координат стоек опоры

Точка	Интервал неопределенности		Ширина интервала, м	
	x	y	x	y
A	[7782,045; 7782,104]	[7270,014; 7270,059]	0,059	0,045
B	[7775,939; 7776,007]	[7264,665; 7264,719]	0,068	0,053
C	[7781,208; 7781,432]	[7258,567; 7258,648]	0,224	0,081
D	[7787,358; 7787,450]	[7263,848; 7264,056]	0,097	0,208