МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РФ
ФГБОУ ВО «АЛТАЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Н.М. ОСКОРБИН., Г.И. АЛГАЗИН, С.И. СУХАНОВ

МЕТОДЫ И МОДЕЛИ ИНТЕРВАЛЬНОГО АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Учебное пособие

Барнаул

Издательство
Алтайского государственного
университета
2023

УДК 519.8(075.8)
ББК 22.181я73
О-744

Рецензенты:

О.И. Пятковский, доктор технических наук, профессор, профессор кафедры информационных систем в экономике Алтайского технического университета им. И.И. Ползунова

Д.Ю. Козлов, кандидат физико-математических наук, доцент, заведующий кафедрой информатики Алтайского государственного университета

О-744 **Оскорбин, Николай Михайлович.**

Методы и модели интервального анализа экспериментальных данных : учебное пособие / Н.М. Оскорбин, Г.И. Алгазин, С.И. Суханов. – Барнаул : Изд-во Алт. ун-та, 2023. – 95 с.

ISBN 978-5-7904-2755-8.

Учебное пособие предназначено для студентов вузов, обучающихся по специальностям «Прикладная информатика», «Математика и информатика».

УДК 519.8(075.8)
ББК 22.181я73

ISBN 978-5-7904-2755-8

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Материал данного учебного пособия посвящен традиционным и новым методам и математическим моделям, используемым для идентификации параметров причинно-следственных связей. В современной литературе это направление получило название анализа данных и включает как задачи математической статистики, так и задачи, решаемые методами машинного обучения.

Математические модели и методы прикладного интервального анализа граничат с математической статистикой по характеру формирования баз данных и баз знаний. Интервальный анализ данных имеет средства проверки согласованности исходных данных, выявления и устранения противоречий в них, что позволяет получать гарантированные оценки искомых параметров.

С другой стороны интервальный анализ данных подобно методам машинного обучения существенно использует эмпирический подход к знаниям на основе наблюдений, т.е. использует принцип «так было – так будет».

Эти возможности прикладного интервального анализа данных дают основание надеется на его востребованность в будущем, как при анализе хорошо структурированных экспериментальных данных, так и для решения задач в системах искусственного интеллекта.

Пособие состоит из пяти разделов. В первом рассмотрены методы и задачи регрессионного анализа данных. В разделе 2 эти задачи для линейных процессов рассмотрены методами интервального анализа данных.

В разделах 3 – 5 анализ данных проводится с использованием теории интервальных систем линейных алгебраических уравнений (ИСЛАУ). Изложение задач и методов интервального анализа данных в этих разделах ведётся в обозначениях, которые традиционно используются в линейной алгебре.

Авторы надеются, что изложенный материал будет востребован магистрантами и аспирантами, изучающими дисциплины, связанные с теорией управления, с методами анализа структурированных экспериментальных данных.

РАЗДЕЛ 1. Основы статистических методов анализа данных

ТЕМА 1.1. Теоретические и эмпирические модели процессов

1.1.1. Моделирование процессов с использованием базы наблюдений и базы знаний

Структурно модели процессов представляют в виде «черного ящика» (рисунок 1.1). Задача математического моделирования процессов при условии достаточно точных наблюдений за входными переменными и выходной переменной формулируется следующим образом:

Оценить функцию $y = F(x, a)$ *и интервал* $[\varepsilon_H,\ \varepsilon_V]$ *для значений* ε_y. (1.1)

Тогда на практике можно найти оценки значения выходной переменной при известных значениях вектора x: $y_0 + \varepsilon_H \le y \le y_0 + \varepsilon_V$.

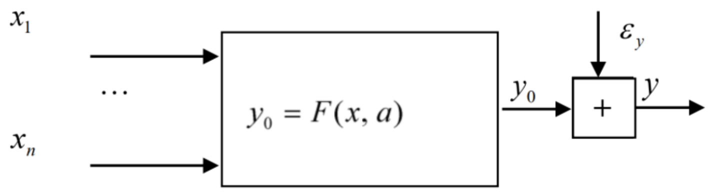


Рис. 1.1. Модель процесса в виде «черного ящика». Обозначения: $x_1, \dots, x_n$ – истинные значения вектора входных переменных; $y_0,\ y$ – значения детерминированной и случайной составляющих выходной переменной моделируемого процесса; ε_y – ненаблюдаемые внутренние аддитивные «шумы»; a – вектор параметров.

Задание 1.1. Изучить задачи моделирования процессов и их отличие от моделей обоснования оптимальных решений по работе [1, стр. 8 – 12].

1.1.2. Теоретические и эмпирические модели процессов

Математическая модель (1.1) называется эмпирической (ЭМП), если основная информация для ее построения – результаты наблюдений моделируемого процесса или наблюдений за процессами – аналогами моделируемого, и теоретической (ТМП), если существенно используются знания соответствующей теории.

Ниже мы рассматриваем методы построения эмпирических зависимостей (эмпирических моделей процессов) в различных исходных предположениях.

Пусть $\{x_j,\ y_j\}_{j=1}^{N}$ – N наблюдений проявления исследуемой зависимости на одном объекте в разные моменты времени или на классе N объектов, выступающих для исследуемого объекта O в качестве объектов-аналогов.

В существующем подходе, например, в регрессионном анализе, выдвигаются некоторые предположения относительно «истинной» зависимости $y = F(x)$, условий наблюдения за совокупностью N объектов, характера неконтролируемых факторов. Далее, используя выбранный критерий качества приближения, отыскивается вариант модели, наиболее близкий к наблюдаемым данным. Эту схему анализа данных иллюстрирует рисунок 2.

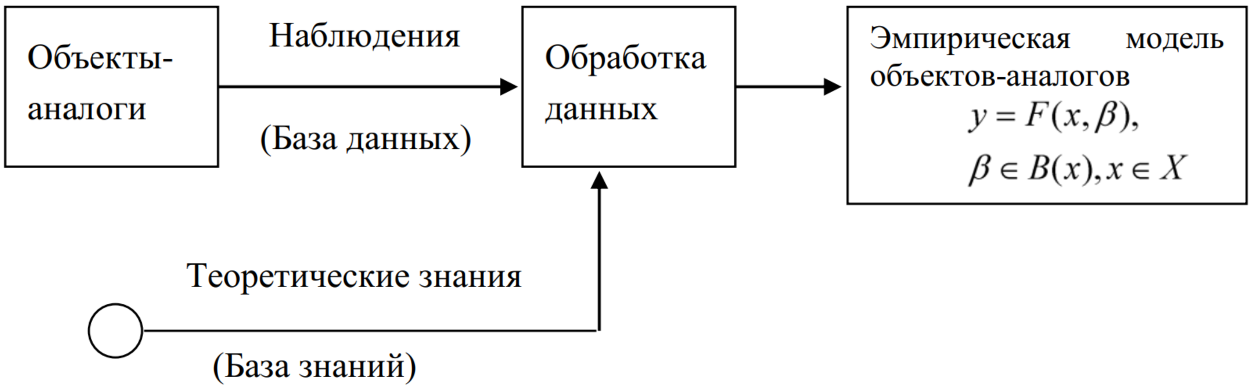


Рис. 1.2. Традиционная схема анализа данных при построении моделей реальных процессов

Задание 1.2. Изучить методы моделирования процессов [2, стр. 13 – 17].

Вопросы по теме 1.1:

1. Пояснить утверждение: «Под эмпирическим моделированием слабоструктурированных реальных процессов мы понимаем получение математических моделей и оценок их достоверности при низком уровне теоретических знаний» [2, стр. 13].

2. Пояснить отличия теории и эмпирики и основной принцип эмпирического метода познания – «Так было – так будет», который определяет уровень доверия.

3. Пояснить термин: «Объекты-аналоги». Аналоги чего? Почему наблюдения за моделируемым объектом О в прошлом при строгом подходе мы должны называть наблюдениями объекта-аналога? При построении формулы рыночной стоимости городских квартир выделяют класс квартир с известной рыночной стоимостью. Укажите критерии отбора этого класса, которые выступят в качестве аргументов искомой формулы. Используйте работу [2, стр. 15].

4. Сравните рисунки 1 и 2 в книге [2, стр. 14, 15]. Поясните ссылку 2 на стр. 15 [2].

5. Поясните утверждение [2, стр. 15]: «…мы «забыли» при моделировании об основном объекте O, для которого конструировали модель, и ограничились лишь объектами-аналогами».

6. Поясните утверждение [2, стр. 15]: «…используя на практике полученную по схеме (рисунка 1.2) эмпирическую модель, мы сталкиваемся с противоречивыми требованиями точности описания объекта и простоты использования модели. Для теории моделирования необходимо точное и непротиворечивое с исходными базами данных и знаний отражение объекта в модели. Для практики чаще всего достаточно простой, легко интерпретируемой «копии» объекта, требования к которой диктуются потребителем».

7. Сравните рисунок 1.1 и рисунок 1.2. Почему для объекта O необходимо выполнение условия: $a \in B(x),\ x \epsilon X$?

ТЕМА 1.2. МНК. Историческая справка. Постановка задачи. Методы численного анализа

1.2.1. МНК. Историческая справка

Метод наименьших квадратов (МНК) – математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных. Он может использоваться для «решения» переопределенных систем уравнений (когда количество уравнений превышает количество неизвестных), для поиска решения в случае обычных (не переопределенных) нелинейных систем уравнений, для аппроксимации точечных значений некоторой функции. МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

Задание 1.3. Изучить метод наименьших квадратов:

1. История МНК от Гаусса до Маркова (https://ru.wikipedia.org/wiki/Метод_ наименьших_квадратов).

2. Переопределенная система линейных уравнений и оценка ее решения методом наименьших квадратов (https://ru.wikipedia.org/wiki/Переопределённая_система).

3. Математическая постановка задачи решения переопределенной системы уравнений методом наименьших квадратов.

4. Привести пример переопределенной системы линейных уравнений – системы линейных уравнений, в которой число неизвестных меньше числа уравнений.

1.2.2. МНК. Постановка задачи

Метод наименьших квадратов с полным правом можно отнести к классическим методам эмпирического моделирования процессов. Начиная с Гаусса этот простой и во многом изящный способ поиска линии наилучшего сглаживания зависимостей привлекал внимание ученых и специалистов-практиков как средство наглядного отражения наблюдаемых связей между группами переменных, выявления и представления их закономерностей.

Существенное развитие МНК получил в XX в. с разработкой процедур оценивания параметров в рамках математической статистики. Было показано, что для нормально распределенных совокупностей случайных величин $(y,\ x)$ при независимой выборке оценки параметров функции регрессии $M[y/x]$ методом максимального правдоподобия сводятся к МНК и являются несмещенными, состоятельными и эффективными.

Задача моделирования в содержательной постановке заключается в построении по экспериментальным данным функциональной зависимости

$$y = F(x, z), \tag{1.2}$$

где y – вектор выходных переменных; $x = (x_1, \dots, x_n)$ – вектор входных контролируемых переменных; z – вектор неконтролируемых переменных. Отметим, что обозначения переменных в (1.2) являются общепринятыми в задачах статистического описания процессов.

Вектор z в (1.2) рассматривается обычно как случайный вектор, поэтому выход y также является случайным. Если между компонентами вектора y отсутствует стохастическая зависимость, то построение модели (1.2) можно проводить по каждой компоненте раздельно. В дальнейшем рассматривается задача моделирования для класса объектов, допускающих в заданной области X пространства входных переменных описание моделью следующего вида:

$$y = y_0 + \varepsilon(z);\ y_0 = F(x, \beta),\ x \epsilon X. \tag{1.3}$$

где ε – случайная величина, являющаяся случайной составляющей выходной скалярной переменной y; y_0 – детерминированная составляющая переменной y; $\beta = (\beta_1, \dots, \beta_m)$ – вектор неизвестных параметров.

Задача построения модели (1.3) ставится в следующем виде. По результатам наблюдения за входными и выходной переменными найти оценки $\hat{\beta}$, оценки вероятностных характеристик переменной ε и провести статистический анализ полученных оценок. При этом структура модели (1.3), область X, случайность величины ε являются априорной информацией. Величины y_j, x_j измеренные в j-м опыте, связаны выражением

$$y_j = F\left(x_j, a\right) + \varepsilon_j; \;\; j = 1, \dots, N. \tag{1.4}$$

Предполагается, что выборка случайной переменной ε удовлетворяет следующим условиям:

$$M\left[\varepsilon_j, \varepsilon_v\right] = \begin{cases} \sigma_\varepsilon^2, \; j = v; \\ 0, j \neq v; \end{cases} \;\; M\left[\varepsilon_j\right] = 0, j, v = 1, \dots, N. \tag{1.5}$$

Оценка $\hat{\beta}$ вектора параметров a в (1.4) находится методом наименьших квадратов решением экстремальной задачи

$$S_\varepsilon^2 = \min_{\beta \epsilon R^m} \frac{1}{N-m} \sum_{j=1}^{N} \left(y_j - F\left(x_j, \beta\right)\right)^2, \;\; N > m, \tag{1.6}$$

где минимальное значение целевой функции S_ε^2 – несмещенная оценка дисперсии σ_ε^2.

В большинстве работ, посвященных данной задаче моделирования, функция $F\left(x_j, \beta\right)$ принимается линейной относительно коэффициентов β:

$$y_0 = \sum_{i=1}^{m} \beta_i \cdot f_i\,(x). \tag{1.7}$$

Тогда необходимые (и достаточные) условия оптимальности задачи (1.6) записываются системой линейных уравнений, а задача моделирования становится наиболее простой.

1.2.3. Методы численного анализа и компьютерного моделирования

В нашем случае решение задач моделирования процессов и исследование свойств оценок МНК будем проводить в среде Excel с использованием инструмента «Поиск решения».

Для исследования оценок МНК необходимо использовать методы генерирования совокупности таблиц наблюдений с заданными свойствами. Для этого в исследовательских компьютерных программах достаточно иметь наборы равномерно и нормально распределенных псевдослучайных чисел. Ниже кратко опишем методы их получения.

Случайное число равномерно распределено на интервале $[a,\ b]$, если его плотность вероятности на этом интервале является постоянной величиной, равной $\frac{1}{b-a}$.

Равномерно распределенные на $[a,\ b]$ числа u имеют следующие характеристики: математическое ожидание – $M_u = \frac{b+a}{2}$ и дисперсию – $D_u = \frac{(b-a)^2}{12}$.

Генерация равномерно распределенных чисел на интервале [0, 1] в MS Excel осуществляется с помощью встроенной функции СЛЧИС().

Если необходимо сгенерировать число u на некотором интервале $[a,\ b]$, используют следующее преобразование: $u = \ b + (b-a) \cdot \text{СЛЧИС}()$.

Генерирование нормально распределенных случайных чисел с заданными математическим ожиданием и дисперсией можно проводить в среде Excel с использованием функции СЛЧИС(). Теоретическое основание метода – центральная предельная теорема.

Пусть необходимо получить N таких чисел $h_j,\ j = 1, \dots, N$ с заданными дисперсией D_h и математическим ожиданием M_h. Для генерации каждого числа используем k равномерно распределенных на интервале [0, 1] чисел $p_{lj}, l = 1, \dots, k$. Преобразование проводим следующими формулами: $H_j = \sum_{l=1}^{k} p_{lj}$; $\overline{H} = k \cdot 0.5$; $D_H = \frac{k}{12}$. Тогда $h_j = \frac{\sqrt{D_h}}{\sqrt{D_H}} \cdot \left(H_j - \overline{H} + M_h\right)$.

Более точные способы генерирования нормально распределенных псевдослучайных чисел можно выполнить с использованием, например, рекомендаций, приведенных в интернет по ссылке https://nesteruk.wordpress.com/2013/04/16/random-number-generation/.

Задание 1.4. (индивидуальное задание). Для объекта на рисунке 1.1. сгенерировать выборку линейного объекта с двумя входами $x_1 \in [-10, 10]$; $x_2 \in [-10, 10]$; $a_0 = 1$; $a_1 = 1$; $a_2 = 2$. Принять для нормально распределенного случайного внутреннего шума следующие характеристики: математическое ожидание: $M_\varepsilon = 0$; $\sigma_\varepsilon = \sqrt{D_\varepsilon} = {}^{1,5}/_{1,96}$. Размер выборки принять равной 100 $(N = 100)$. Входные переменные в заданных границах сгенерировать с использованием оператора СЛЧИС(). Получить МНК оценки параметров процесса. При защите результатов сравнить полученные оценки с расчетными данными других магистрантов. Пример слайдов компьютерной модели генерации наблюдений и анализа данных приведен в приложении 1.

Вопросы по теме 1.2: С использованием изложенного учебного материала в теме 1.2 настоящего учебного пособия и раздела 2.1. работы [2, стр. 73 – 78] подготовить ответы на следующие вопросы:

1. Сформулируйте отношения оценок МНК и оценок метода максимального правдоподобия [2, стр. 73].

2. Поясните высказывание: «Роль математической статистики состояла еще и в том, что для оценок МНК были найдены вероятностно-статистические распределения. Это позволило давать не только точечные, но и интервальные оценки искомым параметрам зависимостей и способствовало развитию специальных разделов математической статистики» [2, стр. 73].

3. Поясните формулы (1.3) – (1.6).

4. Почему в формуле (1.4) приведен параметр a, а далее он заменен на параметр β. Может здесь ошибка?

5. Объем выборки в задаче моделирования процесса, описанной в задании 1.4, равен 100. Соответствует ли это число рекомендации работы [2, стр. 78]?

ТЕМА 1.3. Регрессионный анализ. Свойства оценок МНК

Перечислим предположения регрессионного анализа относительно свойств объекта описания и объектов-аналогов, представленных на рисунке 1.2.

1. Структура принятой математической модели (1.3) (или (1.6)) верна, т.е. «истинная» функция $y_0 = F(x, \beta)$, $x \epsilon X$ является детерминированной для всех $x \epsilon X$; β – неслучайные параметры.

2. Функция $F(x, \beta)$, $x \epsilon X$ известна вычислителю модели.

3. Переменная ε является случайной величиной с нулевым математическим ожиданием и дисперсией, независимой от $x \epsilon X$ и β.

Наблюдатель не вносит ошибок при измерении x (не делает их случайными) и обеспечивает независимость выборки (условие 1.5)).

Предположения относительно объекта описания являются наиболее сильными, поскольку, во-первых, значительно ограничивают класс допустимых объектов, и, во-вторых, предполагают большой объем априорных сведений об объекте описания.

В прикладных задачах, например, при описании производственных, социальных систем производится сознательное загрубение модели, а значит, модели выступают только как аппроксимирующие выражения. Еще большие различия между «истинной» и принятой структурой функций $F(x,\beta),\ x\epsilon X$ проявляются в моделях прогноза параметров. При аппроксимации детерминированной функции величина ε не является случайной. Все эти случаи не удовлетворяют исходным предпосылкам регрессионного анализа.

Вопросы по теме 1.3:

1. Приведите список предположений регрессионного анализа относительно свойств объекта описания [2, стр. 77].

2. Поясните свойства статистических оценок при выполнении условий справедливости неравенства Крамера – Рао: несмещенность, состоятельность и эффективность (смотри ссылки: https://ru.wikipedia.org/wiki/Неравенство_Крамера_—_Рао).

3. О каких статистических оценках в регрессионном анализе идет речь в п.2?

4. Выберите примеры эффективных оценок, которые имеют место в рассматриваемом случае (http://www.nsu.ru/mmf/tvims/chernova/ms/lec/node24.html).

5. Проблемы регрессионного анализа выделены в работе [2, стр. 78 – 81]. Составьте их краткий перечень.

6. Поясните определение адекватных моделей: «Под адекватностью модели здесь понимается статистическое тождество модели и объекта. Модель считается адекватной, если, по имеющимся статистическим данным, невозможно найти существенные отклонения модели от объекта» [2, стр. 79].

7. Поясните, что влечет замена предположения о знании истинной модели предположением о ее адекватности [2, стр. 79].

ТЕМА 1.4. Ошибки измерения входных переменных: конфлюэнтный анализ

Задачами конфлюэнтного анализа является изучение структуры случайных величин, и оценка значений коэффициентов модели процесса в условиях, когда все переменные измеряются с ошибками.

Конфлюэнтный анализ определяется как совокупность методов математического моделирования процессов в условиях, когда входные переменные $x_1, \dots, x_n$ являются количественными (случайными или не случайными) и наблюдаются в каждом опыте с ошибками. Этим метод конфлюэнтного анализа отличается от методов классического регрессионного анализа. Один из подходов анализа данных в рассмотренном случае предложен в работе [2], который кратко изложим в данном разделе.

При исследовании метода конфлюэнтного анализа относительно ошибок обычно предполагается их случайность и независимость от входных переменных, от прошлых значений. Полагаем, что ошибки наблюдений, как входных факторов, так и выходной переменной процесса являются нормальными с нулевыми математическими ожиданиями и постоянной в каждом наблюдении дисперсией.

С учетом предположений относительно ошибок, таблица наблюдений $\{Z_1^j, \dots, Z_n^j; V^j\}, j = 1, \dots, N$ за всеми переменными процесса определяется следующими формулами:

$$Z_i^j = x_i^j + \varepsilon_{xi}^j;\ V^j = y^j + \varepsilon_y^j;$$

$$M[\varepsilon_{xi}^j] = M[\varepsilon_y^j] = 0;\quad D[\varepsilon_{xi}^j] = d_i;\quad D[\varepsilon_y^j] = d_{y,}\ i = 1, \dots, n; j = 1, \dots, N.$$

Здесь N– общее число наблюдений; n– число входных переменных. Связь входных и выходных переменных предполагаем заданной и линейной формулой следующего вида:

$$f(x) = a_1 x_1 +_{\dots} + a_n x_n\ ,$$

а задачу конфлюэнтного анализа можно ограничить оценкой истинных значений коэффициентов приведенного уравнения.

Рассмотрим метод оценки истинных значений параметров моделируемого процесса. Пусть без потери общности средние значения переменных в таблице наблюдений являются нулевыми. Можно показать, что оценки истинных коэффициентов моделируемого процесса находятся решением следующей системы уравнений:

$$a_i(R_{ii} - d_i) + \sum_{k=1,k\neq i}^{n} a_k R_{ik} = R_{iy},\ i = 1, \dots, n. \quad (1.8)$$

В приведенной системе n уравнений, коэффициентами выступают ковариации соответствующих переменных, вычисленные по таблице наблюдений. $\{Z_1^j, \dots, Z_n^j, V^j\}, j = 1, \dots, N$; d_i – дисперсии ошибок измерения входных переменных соответственно, значения которых считают известными.

Отличие системы уравнений (1.8) от аналогичной системы регрессионного анализа состоит в коррекции диагональных элементов левой части. Исследование оценок (1.8) методами вычислительных экспериментов показывает, что они являются несмещенными и состоятельными. Таким образом, представленный метод конфлюэнтного анализа может быть использован на практике.

ТЕМА 1.5. Критерии работоспособности и адекватности моделей

Работоспособность эмпирических моделей процессов выражается как количественным показателем, так и качественно. Количественный показатель работоспособности характеризует свойство модели с точки зрения потребителя и может быть выражен одной или несколькими величинами [2, стр. 82 – 83].

Описательную способность моделей МНК предлагается характеризовать величиной множественного коэффициента корреляции. В [3, стр. 7] приведена удобная шкала его значений (см. таблицу).

Таблица 1.1.

Относительная шкала изменения коэффициента корреляции

Значение R^2	< 0,3	0,3–0,5	0,5–0,7	0,7–0,9	> 0,9
Качественная характеристика взаимосвязи	слабая	умеренная	заметная	высокая	очень высокая

Поскольку получить детерминированное математическое описание объекта не удается, то для каждой построенной модели необходимо проверить, удовлетворяет ли она требованиям потребителя, т.е. работоспособна ли модель в качественном отношении. Для качественного заключения о работоспособности модели необходимо, прежде всего, на интервале изменения количественного показателя найти граничное значение, разделяющее модели на работоспособные и неработоспособные.

Определение граничных значений усложняется тем, что они используются до решения задачи потребителя и, следовательно, могут быть выбраны только по результатам решения сопоставимых задач потребителя. Это требование диктует условие использования сопоставимых (сравнимых) показателей точности различных математических моделей. Для проверки работоспособности моделей, особенно для новых приложений, необходимо [2, стр. 83]: выбрать сравнимый показатель работоспособности статистической модели; накопить опыт классификации задач потребителя и определить для каждого класса задач граничные значения используемых показателей работоспособности.

На практике можно использовать шкалу допустимых значений для работоспособных моделей, приведенную в таблице 1.1, например градации «высокая» и «очень высокая».

Понятие адекватности моделей в смысле, указанном в разделе 2.1, является качественным понятием. Адекватность не совпадает с работоспособностью. Эти понятия носят совершенно различное содержание и используются поразному. Модель может быть работоспособной, но не адекватной, она может быть адекватной, но не работоспособной [2, стр. 84].

Задание 1.5. Вычислить показатель R^2 для индивидуального варианта (**Задание 1.4**) и решить вопрос о работоспособности эмпирической модели. Как анализом невязок проверить адекватность модели, полученной по данным индивидуального задания?

Вопросы по теме 1.4:

1. Как вычислить значения показателей работоспособности, приведенных в работе [2, стр. 82 – 83]. Пояснить для своего индивидуального задания.

2. Поясните утверждение: «Модель может быть работоспособной, но не адекватной, она может быть адекватной, но не работоспособной» [2, стр. 84].

ТЕМА 1.6. Модели экспериментов. Правильные (идеальные) и реальные схемы наблюдений

Схема правильных наблюдений описана в теме 2, и реализована при выполнении индивидуального задания (**Задание 1.4**).

Наблюдения, в процессе выполнения которых имеются отклонения (или эти *отклонения возможно присутствуют*), назовем **неправильными наблюдениями**.

Общая схема наблюдения объектов аналогов при построении математической модели и схемы наблюдений при ее использовании представлена на рисунке 1.2.1 работы [2, стр. 30], в том числе при использованиях для прогноза, для оценки параметров, для обоснования оптимальных решений.

Наиболее очевидными в регрессионном анализе для неправильных наблюдений выступают следующие ситуации, которые приводят к нарушениям в реальности схемы правильных наблюдений:

1. Наличие ошибок измерения входных переменных.

2. Наличие ошибок измерения выходной переменной в совокупности с внутренними шумами.

3. Наличие аномалий в организации эксперимента, при измерениях и «выбросов» при регистрации данных. Процент выбросов может быть известен или неизвестен, но должен быть относительно не большим, до 5-10%.

4. Использование плана эксперимента, близкого к классу вырожденных планов.

Учитывая, что нарушение некоторых исходных предположений приводит к необходимости модификации процедур обработки, мы к схемам неправильных наблюдений можем относить и грубые несоответствия в реальности условий корректного применения выбранного метода моделирования процессов. К данной группе можно отнести:

1. Нарушение предположений о составе и области изменения значений параметров моделируемого процесса и характере случайных ошибок наблюдения.

2. Нарушение предположения о структуре модели процесса.

Пример дополнительной обработки данных в схеме эмпирического моделирования приведен в разделе «Обработка информации в случае неправильных наблюдений» [2, стр. 141].

Задание 1.6. Получить в среде MS Excel таблицы данных при не стандартных ошибках наблюдений переменных (нарушение п.1) линейного объекта [3, стр. 170, 171]. Для удобства это задачу приведем ниже.

В работе [3] рассмотрены результаты вычислительных экспериментов с модельным линейным процессом

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 \qquad (1.9)$$

и известными параметрами $a_0 = a_1 = 1;\ a_2 = -2.$

При наблюдениях за переменными x_1, x_2 и y вносятся аддитивные случайные ошибки измерения. При имитации процесса моделирования рассмотрены две схемы наблюдения.

Схема П. Ошибки 7 наблюдений независимы, одинаково распределены, а операторы Microsoft Excel имеют следующий вид:

$$\varepsilon_{x_1} = 0{,}1 \cdot \text{СЛЧИС}() - 0{,}5; \qquad (1.10)$$

$$\varepsilon_{x_2} = 02 \cdot \text{СЛЧИС}() - 0{,}1; \qquad (1.11)$$

$$\varepsilon_{y} = 0{,}05 \cdot \text{СЛЧИС}() - 0{,}025. \qquad (1.12)$$

Пример псевдослучайных переменных процесса приведен в таблице 1.2.

Таблица 1.2.

Исходные данные для исследования схемы *П* моделирования процессов

№ п/п	x_{1i}	x_{2i}	y_i	$\varepsilon_{x_{1i}}$	$\varepsilon_{x_{2i}}$	ε_{y_i}	x_{1i}^p	x_{2i}^p	y_i^p
1	1	8	–14	–0,05	0,01	0,010	0,95	8,01	–13,99
2	3	–3	10	–0,02	0,09	0,021	2,98	–2,91	10,02
3	5	6	–6	–0,04	–0,08	–0,016	4,96	5,92	–6,02
4	7	1	6	–0,02	0,03	–0,007	6,98	1,03	5,99
5	9	3	4	0,02	0,10	–0,017	9,02	3,10	3,98
6	11	5	2	0,00	–0,07	–0,002	11,00	4,93	2,00
7	13	2	10	0,02	–0,03	–0,023	13,02	1,97	9,98

Схема Э. Ошибки измерения переменных x_1, x_2 и y принадлежат эллипсу, оси которого параллельны исходным осям координат. Операторы Microsoft Excel (1.10) и (1.12) сохраняются, а ε_{x_2} генерируется как случайное число в интервале $[-\Delta_{2i},\ \Delta_{2i}\ ]$, где

$$\Delta_{2i} = 0{,}2 \cdot \sqrt{1 - \frac{\varepsilon_{1i}^2}{0{,}1^2}}. \qquad (1.13)$$

Исходные данные приведены в таблице 1.3.

Таблица 1.3.

Исходные данные для исследования схемы Э моделирования процессов

№ п/п	x_{1i}	x_{2i}	y_i	$\varepsilon_{x_{1i}}$	$\varepsilon_{x_{2i}}$	ε_{y_i}	x_{1i}^p	x_{2i}^p	y_i^p
1	1	8	–14	–0,02	0,06	0,008	0,98	8,06	–13,99
2	3	–3	10	0,01	–0,05	0,001	3,01	–3,05	10,00
3	5	6	–6	0,03	0,05	–0,022	5,03	6,05	–6,02
4	7	1	6	0,00	0,05	–0,013	7,00	1,05	5,99
5	9	3	4	0,03	-0,01	–0,012	9,03	2,99	3,99
6	11	5	2	–0,04	0,04	0,020	10,96	5,04	2,02
7	13	2	10	–0,03	0,07	–0,023	12,97	2,07	9,98

Приведенные в таблицах 1.2 и 1.3 базы данных и знаний могут служить в качестве примеров для изучения задач анализа данных как статистическими, так и интервальными моделями и методами.

Задание 1.7. Вопросы по разделу 1 [2 –6].

1. Построение и анализ эмпирических зависимостей.

2. Метод наименьших квадратов: историческая справка.

3. Теоретические основы регрессионного анализа. Метод максимума правдоподобия.

4. Критерии работоспособности и адекватности эмпирических моделей.

5. Планирование экспериментов: описание проблемы, примеры оптимальных факторных экспериментов.

Вопросы по теме 1.5:

1. Дать определение правильных наблюдений в задачах анализа данных.

2. Дать определение неправильных наблюдений в задачах анализа данных.

3. Указать основные причины, когда таблицу наблюдений в задачах анализа данных нельзя назвать правильной (идеальной).

4. Поясните утверждение: «…к схемам неправильных наблюдений можем относить и грубые несоответствия в реальности условий корректного применения выбранного метода моделирования процессов».

Библиографический список к разделу 1

1. Оскорбин Н.М., Журавлева В.В. Аналитические методы и модели в экономике: Учебное пособие. – Барнаул: Изд-во Алт. ун-та, 2015. – 30 с.

2. Максимов А.В., Оскорбин Н.М. Многопользовательские информационные системы: основы теории и методы исследования: монография. – 2-е изд. испр. и доп. – Барнаул : Изд-во Алт. ун-та, 2013. – 264 с.

3. Суханов В.А. Исследование эмпирических зависимостей: нестатистический подход: сб. науч. ст. / под. ред. Н.М. Оскорбина, П.И. Кузьмина. – Барнаул: Изд-во Алт. ун-та, 2007. – 290 с.
4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Множественная регрессия. – М.: «Диалектика», 2007. – 912 с.
5. Радченко С.Г. Устойчивые методы оценивания статистических моделей. – К.: ПП «Санспарель», 2005. – 504 с.
6. Радченко С.Г. Методология регрессионного анализа. – К.: «Корнийчук», 2011. – 376 с.

РАЗДЕЛ 2. Классический вариант интервального анализа данных

ТЕМА 2.1. Основы интервального анализа данных. Информационное множество.

2.1.1. Теоретические основы интервального анализа данных

Методы интервального анализа данных относят к нестатистическим методам теоретического или эмпирического моделирования процессов. Единственное (но существенное) отличие нестатистических методов от вероятностных методов состоит в том, что математические модели внутреннего «шума» (величины ε_y на рисунке 1.1.) и погрешностей измерения переменных моделируемого процесса при формировании базы данных не являются вероятностными моделями. В случае, когда некоторые параметры модели или ошибки измерения являются случайными величинами (случайными процессами или событиями), то при использовании нестатистических методов обработки данных эта иногда, безусловно, достоверная, информация не учитывается.

Как отмечено во введении начало интервального анализа причинно-следственных связей положено Канторовичем Л.В. в работе [1]. Это направление прикладного интервального анализа следует отнести к классическому разделу теории. Следует заметить, что его освоение в научном и прикладном аспектах началось с двадцатилетней задержкой. Существенное развитие методов и моделей прикладного интервального анализа получило при анализе данных с интервальными погрешностями измерения всех переменных моделируемых процессов и при анализе задач эмпирического моделирования процессов. Указанный раздел теории и практики прикладного интервального анализа следует назвать современным. Соответствующие модели, методы и результаты этого направления рассмотрим в следующих разделах учебного пособия.

В методах интервального анализа данных используются интервальные оценки для указанных «шума» и погрешностей измерения. Пример интервальных данных представлен в таблицах 1.2, 1.3 **Задания 1.6** при условии, что в процессе обработки данных вероятностные свойства псевдослучайных чисел при их генерации не учитываются при их анализе.

Корректное применение нестатистических методов анализа данных возможно в рамках теоретического моделирования и в рамках эмпирического моделирования процессов. Если исследователь располагает достоверной базой знаний, а база данных сформирована при правильных наблюдениях (см. рисунок 1.2) то полученная модель процесса в виде выражения (1.1) достоверна и относится к классу теоретических моделей процессов. Иначе необходимо использовать методы эмпирического моделирования процессов, а полученная модель является эмпирической.

Рассмотрим простейший случай построения теоретической модели процесса, представленного на рисунок 1.1, для которого зависимость y_0 от входных переменных $x = (x_1, \dots, x_n)$ является линейной, а внутренний шум отсутствует $\left(\varepsilon_y = 0\right)$. При формировании базы наблюдений переменные $(x_1, \dots, x_n)$ измеряются точно (без ошибочно), а измеренное значение выходной переменной Y_j является интервальной оценкой переменной y_{0j} в каждом из проведенных опытов $j = 1, \dots, N$. Эту интервальную переменную можно задать двумя способами:

$$Y_j = \left[y_j^H,\ y_j^V\right];\ Y_j = y_j^R \pm \varepsilon_j^0;\ y_j^H = y_j^R - \varepsilon_j^0;\ y_j^V = y_j^R + \varepsilon_j^0; y_{0j} \in \left[y_j^H,\ y_j^V\right].$$

Здесь y_j^R – результат измерения при проведении опыта с номером j, а ε_j^0 – оценка модуля ошибки измерения в этом опыте.

Тогда неизвестные значения истинных коэффициентов линейной модели:

$$y_0 = a_1 \cdot x_1 + \dots + a_n \cdot x_n \tag{2.1}$$

удовлетворяют системе N двухсторонних неравенств:

$$y_j^R - \varepsilon_j^0 \le\ a_1 \cdot x_{1j} + \dots + a_n \cdot x_{nj} \le y_j^R + \varepsilon_j^0; j = 1, \dots, N. \tag{2.2}$$

Заметим, что модель (2.1) соответствует и случаю, когда искомая зависимость содержит свободный член. В этом случае входная переменная x_1 во всех наблюдениях полается равной единице. Кроме того, в частных случаях оценки ε_j^0 модуля погрешностей измерения могут быть одинаковыми во всех или части наблюдений.

2.1.2. Информационное множество

Обозначим $A(N)$ – множество значений вектора $a = (a_1, \dots, a_n)$, удовлетворяющих системе неравенств (2.2). Это множество в прикладном интервальном анализе играет фундаментальную роль. Во-первых, оно содержит всю полученную информацию об искомой зависимости (2.1), т.е. является теоретическим результатом ее оценивания. Во-вторых, в ситуациях эмпирического моделирования это множество выступает основой оценки множества эмпирических моделей $B(x)$, представленных на рисунке 1.2. И, наконец, все задачи теоретического моделирования процессов решаются с использованием этого множества, в том числе и проверка работоспособности полученной модели. Учитывая указанное фундаментальное значение множества $A(N)$, в работах по нестатистическим методам моделирования процессов оно носит особое название: **«множество неопределенности»** или «**информационное множество**» [2, 6, 7, 11].

Нормальным результатом анализа данных считается случай, когда множество $A(N)$ не пусто и ограниченно. В случае пустого множества $A(N)$ информация баз знаний и баз данных противоречива и требуются их корректировка. Случай, когда множество $A(N)$ не ограниченно, свидетельствует о не достаточности информации для анализа и требуется расширить состав баз данных и/или знаний.

Вопросы по теме 2.1:

1. Сформулируйте основную задачу моделирования процессов и поясните отличие теоретических и эмпирических моделей (использовать рисунок 1.1 и выражение (1.1)).

2. Поясните отличие статистических и нестатистических методов анализа данных.

3. При каких условиях данные таблиц **Задания 1.6** можно анализировать статистическими методами, и при каких условиях эти данные анализируются нестатистическими методами?

4. Приведите графическую интерпретацию системы неравенств (2.2) при n=2.

5. Дайте определение информационного множества и множества неопределенности.

6. Какой элемент этого множества претендует отображать искомую оценку неизвестных коэффициентов в уравнении (2.1)?

7. Поясните состояние проблемы анализа данных в случаях, когда множество $A(N)$ является пустым или неограниченным.

ТЕМА 2.2. Интервальные оценки моделируемого процесса (теоретические модели процессов)

2.2.1. Точечные и интервальные оценки моделируемого процесса

Рассмотрим основные прикладные задачи, которые решаются при моделировании процессов.

1. Задача прогноза выходной переменной в точке факторного пространства $x^P = (x_1^P, \dots, x_n^P)$.

Интервальную оценку $[y^H(x^P),\ y^V(x^P)]$ получаем решением двух задач ЛП:

$$y^H(x^P) = \min_{a \in A(N)}(a_1 x_1^P + \dots + a_n x_n^P);\ y^V(x^P) = \max_{a \in A(N)}(a_1 x_1^P + \dots + a_n x_n^P). \quad (2.3)$$

Качество прогноза можно судить по показателю приведенной ошибки прогноза:

$$\gamma(x^P) = \frac{0.5 \cdot \left(y^V(x^P) - y^H(x^P)\right)}{\max_{j=1,\dots,N} y_j^R - \min_{j=1,\dots,N} y_j^R} \cdot 100\%. \quad (2.4)$$

Для работоспособных моделей в экономике и социологии допустимая погрешность прогноза не должна превышать от 5% до 15%. В технических приложениях этот порог допустимости обычно принимается на уровне 4%. Таким образом, выражение (2.4) может служить показателем для проверки работоспособности полученной модели процесса.

2. Задача интервальной оценки параметров искомой зависимости.

В качестве интервальной оценки всех параметров линейной модели (2.1) достаточно решить *2n* задач линейного программирования (ЗЛП). Например, гарантированная оценка коэффициента a_1 принадлежит интервалу $[a_1^H,\ a_1^V]$:

$$a_1^H = \min_{a \in A(N)} a_1; \qquad a_1^V = \max_{a \in A(N)} a_1. \quad (2.5)$$

4. Точечные оценки прогноза и параметров моделируемого процесса.

Для указанных точечных оценок можно принять середины интервалов, полученных выражениями (2.3), (2.5).

5. Проекция точки a^P на множество $A(N)$.

Точка a^P принадлежит множеству $A(N)$ тогда и только тогда, когда δ равно нулю, где δ – решение следующей задачи квадратичного программирования:

$$\delta = \min_{a \in A(N)} \|a^P - a\|. \tag{2.6}$$

Задача (2.6) принадлежности заданной точки множеству $A(N)$ используется в компьютерных моделях исследования гарантированных оценок исследуемых параметров. Пусть, например, a^P – истинное значение параметров моделируемого линейного процесса (2.1), а $A(N)$ – найденное информационное множество. Для выполнения условий гарантированного оценивания необходимо в проведенных вычислительных экспериментах решать задачу (2.6). Условия выполнены при нулевых минимальных значениях целевой функции в (2.6).

6. Задача исключения не значимых или мало значимых факторов моделируемого процесса.

Один из способов решения данной задачи в прикладном интервальном анализе основан на использовании интервальной оценки коэффициентов модели согласно (2.5). Если нулевое значение исследуемого параметра принадлежит найденному интервалу, то соответствующую входную переменную можно считать не значимой и сократить факторное пространство. На практике существуют и другие приемы решения этой задачи с использованием, например, результатов анализа полных и сокращенных баз данных.

Следует отметить, что рассмотренные математические постановки задач анализа данных не меняются в общем случае наличия ошибок измерения всех переменных. Изменения касаются системы неравенств, которые задают информационное множество, в том числе задание информационных множеств решениями интервальных систем линейных алгебраических уравнений (ИСЛАУ).

2.2.2. Косвенная проверка выполнимости условий теоретического моделирования процессов

Следует отметить, что при теоретическом моделировании процессов задачи разрешимы (множество неопределенности $A(N)$ не пусто и ограниченно). Из (2.2) видим, что множество $A(N)$ компактно.

Кроме того, уровень погрешности согласно (2.4) должен быть по нижней границе соизмерим с числом наблюдений и с оценками погрешностей измерения ε_j^0 [2, стр. 50 – 55].

Если эти условия не выполняются, то следует поставить под сомнение истинность предположений теоретического моделирования.

Исследователь вправе приложить усилия по обеспечению выполнимости предположений теоретического моделирования. Так, если множество $A(N)$ неограниченно, то этот факт указывает **на неполноту данных**. Она возникает при вырожденности плана эксперимента, либо при недостаточном числе наблюдений. При моделировании процессов данная проблема устраняется, как правило, без нарушений исходных предположений.

Если такого обоснования провести не удается, то исследователь обязан использовать методы **эмпирического моделирования процессов**.

Вопросы по теме 2.2:

1. Поясните задачу интервального прогноза выходной переменной для процесса (1.1) графически при $n = 2$.

2. Значение компонент вектора переменной $x^P = (x_1^P, \dots, x_n^P)$ в задаче прогноза задано с интервальными ошибками. Уточните постановки задач прогноза в выражении (2.3).

3. Запишите задачи поиска интервальной оценки значений коэффициента a_0.

4. В интервальном анализе данных активно используются два пространства для отображения множества допустимых прямых (2.1) и множества $A(N)$ решений неравенств (2.2). Поясните графически эти пространства при $n = 2$ и установите их основные свойства.

5. Поясните графически интервальные оценки коэффициентов модели (2.1) при $n = 2$ (Приложение 2).

6. Задача проверка работоспособности построенной модели процесса. Поясните порядок ее решения.

7. В статистических методах используются понятия для искомых оценок: несмещенность, состоятельность и эффективность. Можно ли привести аналоги этих понятий для характеристики множества $A(N)$.

8. Множества $A(N)$ компактны? При каких условиях? Множество $A(N)$ содержит бесконечно удаленные точки. Это нормально для задач анализа данных?

9. Поясните утверждение: «уровень погрешности согласно (2.4) должен быть по нижней границе соизмерим с числом наблюдений и с оценками погрешностей измерения ε_j^0» [2, стр. 50 – 55]. Покажите связь показателя, вычисленного согласно выражения (2.4), с множеством $A(N)$.

ТЕМА 2.3. Интервальный анализ данных в задачах эмпирического моделирования процессов. Метод центра неопределенности

2.3.1. Метод центра неопределенности

Рассматриваем случай, когда при решении задач моделирования (2.3), (2.5) информационное множество оказалось пустым, т.е. система неравенств (2.2) – противоречива. В данном случае условия теоретического моделирования процессов не выполнены, но мы не знаем, какие из условий оказываются нарушенными. Допустим, что мы убеждены в достоверности базы знаний. Как проверить, что база данных не достоверна? Предположим, что одно из N наблюдений с неизвестным номером J является выбросом. Как найти этот номер?

Перечисленные задачи решаются при нахождении точечной оценки параметров линейных моделей методом центра неопределенности (МЦН) [3].

В ряде случаев требуется проверить выполнимость исходных предположений (например, в случае, когда $A(N)$ – пустое множество) или получить точечные оценки параметров модели. Один из способов решения этой задачи связан с «уширением» или с «сужением» множества $A(N)$. Зададим информационное множество $A(N, k)$ следующей системой неравенств:

$$y_j^{И} - k\varepsilon_j^0 \le a_1 x_{1j} + \dots + a_n x_{nj} \le y_j^{И} + k\varepsilon_j^0; \quad j = 1, \dots, N; \quad k > 0. \qquad (2.7)$$

Система неравенств (2.7) совпадает с (2.2) при $k = 1$. Этот параметр называем коэффициентом расширения ($k > 1$) или сужения ($k < 1$). Следующая задача на минимум носит название задачей поиска центра неопределенности [3]:

$$k^* = \min_{a \in (N,k)} k\,. \qquad (2.8)$$

Как показывает практика решения задачи (2.8) на реальных и модельных данных минимум достигается в единственной точке, которая может рассматриваться в качестве точечной оценки параметров моделируемого процесса. Значение k^* выступает индикатором выполнения исходных предположений интервального анализа данных. Так индексы активных наблюдений при $k^* > 1$ выделяют порцию наблюдений базы данных, среди которых возможны грубые ошибки регистрации данных или заниженные оценки ошибок измерения переменных. Подобная информация может быть использована для корректировки базы данных и базы знаний. Здесь и далее под порцией наблюдений понимаются совокупность наблюдений (или их номеров), которые являются активными при решении задач математического программирования в процессе анализа данных.

Решением задачи (2.8) выступают точечная оценка $\hat{a}$ вектора a и значение коэффициента $k^*(N)$. Это решение соответствует граничному минимальному значению параметра k, при котором множество неопределенности еще не является пустым.

Определение. Методы анализа данных, в которых используется задача (2.8) называются **методами центра неопределенности** (МЦН) [4].

2.3.2. Интервальный анализ данных в задачах эмпирического моделирования процессов

Решение задачи (2.8) позволяет выявить выбросы (наблюдения с номером *J*) в случае, когда $k^*(N) > 1$. Такие наблюдения нужно искать среди активных ограничений в задаче (2.8). Далее указанные наблюдения необходимо проверять на достоверность совместно с экспериментаторами. Исследователь (математик) не имеет право самостоятельно исключать эти наблюдения, но может рассматривать различные варианты нарушений исходных предположений и высказывать соответствующие гипотезы.

Некоторые приемы эмпирического моделирования процессов методом центра неопределенности рассмотрены в работе [5] и в главе 3 работы [2, стр. 141 – 145].

Задание 2.1. (индивидуальное задание). Для объекта на рисунке 1.1. сгенерировать выборку линейного объекта с двумя входами $x_1 \in [-10, 10]$; $x_2 \in [-10, 10]$; $a_0 = 1$; $a_1 = 1$; $a_2 = 2$.

Принять для равномерно распределенного случайного внутреннего шума следующий интервал значений [-1.5, +1.5]. Размер выборки принять равной 100 ($N = 100$). Входные переменные в заданных границах сгенерировать с использованием оператора **СЛЧИС().**

Получить модель процесса в виде выражения (1.1). Результаты сравнить с заданием 1.4. При защите результатов сравнить полученные оценки с расчетными данными других магистрантов.

Задание 2.2. Для объекта на рисунке 1.1. сгенерировать выборку линейного объекта с двумя входами $x_1 \in [-10, 10]$; $x_2 \in [-10, 10]$; $a_0 = 1$; $a_1 = 1$; $a_2 = 2$.

Получить интервальные оценки выходной переменной при значениях входных переменных, согласно таблице полного факторного эксперимента. Провести сравнение интервальных оценок с оценками МНК и МЦН. (Задание 2.2. выполняется коллективно на практических занятиях).

Для изучения методов и моделей прикладного интервального анализа следует познакомиться с практикой его применения на реальных данных. Приложения и проблемы методов интервального анализа данных рассмотрены в литературе достаточно подробно [2, 3 – 7].

Задание 2.3. Темы самостоятельных работ:

1. Задачи дисперсионного и ковариационного анализов [2, стр. 133 – 136].

2. Временные тренды в совокупности динамических данных: таблица 3.2.5 [2, стр. 133 – 139], таблица 3.2.6 [2, стр. 139 – 141].

3. Модельный пример 1 Суханова С.И. [6, стр. 34 – 36]. Выполнить анализ данных.

4. Модельный пример 2 Суханова С.И. [6, стр. 37 – 46]. Выполнить анализ данных.

Вопросы по теме 2.3:

1. Утверждается: информационное множество пусто, следовательно, условия теоретического моделирования не выполнены. Поясните это утверждение. Оно верно?

2. Информационное множество не пусто, следовательно, условия теоретического моделирования выполнены. Поясните это утверждение. Оно верно?

3. Допустим, что мы убеждены в достоверности базы знаний. Как проверить, что база данных не достоверна? Предположим, что одно из N наблюдений с неизвестным номером J является выбросом. Как найти этот номер?

4. Поясните графически метод центра неопределенности при поиске оценки коэффициентов модели (2.1) при n=2.

5. Дайте определение метода центра неопределенности. Какие задачи анализа данных решаются этим методом?

Библиографический список к разделу 2

1. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. – 1962. – Т. 3. – №5. – С. 701–709.

2. Максимов А.В., Оскорбин Н.М. Многопользовательские информационные системы: основы теории и методы исследования : монография. – 2-е изд. испр. и доп. – Барнаул : Изд-во Алт. ун-та, 2013. – 264 с.

3. Суханов В.А. Исследование эмпирических зависимостей: нестатистический подход: сб. науч. ст. / под. ред. Н.М. Оскорбина, П.И. Кузьмина. – Барнаул: Изд-во Алт. ун-та, 2007. – 290 с.

4. Оскорбин Н.М, Жилин С.И., Максимов А.В. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия АГУ. – 1998. – № 1. – С. 35–38.

5. Жилин С.И. Нестатистические методы и модели построения и анализа зависимостей // Дисс. … канд. физ.-мат. наук. – Барнаул, 2004. – 119 с.

6. Суханов С.И. Математическое моделирование пространственного положения геообъектов и интервальное оценивание его точности // Дисс. канд. тех. наук. – Барнаул, 2011. – 102 с.

7. Шарый С.П. Конечномерный интервальный анализ. – Новосибирск: Изд-во «*XYZ*». 2016. – 611 с.

РАЗДЕЛ 3. Численные методы интервального анализа данных.

ТЕМА 3.1. Интервальные системы линейных алгебраических уравнений (ИСЛАУ)

Как показано в разделе 2, интервальный анализ данных базируется на теоретических результатах интервальной математики [1]. Используемые на практике численные методы в простых случаях связаны с решением интервальных систем линейных алгебраических уравнений. В данном разделе мы будем использовать обозначения, принятые в линейной алгебре, и литературе по ИСЛАУ и традиционные обозначения линейной алгебры и линейного программирования (ЛП).

В базовых предположениях прикладного интервального анализа ИСЛАУ в матричной форме записывается интервальной ($N \times n$) матрицей коэффициентов и интервальным ($N \times 1$) вектором правой части в следующем виде:

$$\boldsymbol{A}x = \boldsymbol{B}\,. \tag{3.1}$$

Элементы матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ являются интервальными оценками результатов измерения входных и выходной переменной в N наблюдениях и задаются неравенствами: $A^H \le \boldsymbol{A} \le A^V;\ B^H \le \boldsymbol{B} \le B^V$.

Значения вектора $x \in R^n$ в ИСЛАУ соответствует оценками параметров искомой линейной зависимости:

$$b = x_1 a_1 + \dots + x_n a_n \tag{3.2}$$

Интервальные наблюдения $\boldsymbol{a}_{ij}, \boldsymbol{b}_j$ за переменными моделируемого процесса обозначим так:

$$\boldsymbol{a}_{ij} = \left[a_{ij}^H,\ a_{ij}^V\right];\ a_{ij}^H = a_{ij}^M - \varepsilon_{ij}^0;\ \ a_{ij}^V = a_{ij}^M + \varepsilon_{ij}^0;\ i = 1, \dots n;\ j = 1, \dots, N;$$

$$\boldsymbol{b}_j = \left[b_j^H,\ b_j^V\right];\ b_j^H = b_j^M - \varepsilon_{bj}^0;\ \ b_j^V = b_j^M + \varepsilon_{bj}^0;\ j = 1, \dots, N, \tag{3.3}$$

где $a_{ij}^M,\ b_j^M$ – экспериментальные данные входов и выхода в каждом из N наблюдений; $\varepsilon_{ij}^0,\ \varepsilon_{bj}^0$ не отрицательные оценки ошибок наблюдения (ниже индекс j будет опущен, т.е. интервальные ошибки приняты одинаковыми для всех наблюдений); a_{ij}^H, a_{ij}^V, b_j^H, b_j^V – элементы матриц A^H, A^V, B^H, B^V интервальных оценок коэффициентов ИСЛАУ и правой части (3.1) соответственно.

Рассмотрим обычную (не интервальную) систему уравнений, которую запишем так: $Ax = B$. Если набор матриц A, B имеет размерность ИСЛАУ, а сами матрицы включены в заданные соответствующие интервальные матрицы $\boldsymbol{A}, \boldsymbol{B}$ ($A \in \boldsymbol{A}$; $B \in \boldsymbol{B}$) и, если эта система имеет решение, то можно исследовать это решение на принадлежность к решению ИСЛАУ $\boldsymbol{A}x = \boldsymbol{B}$.

Для интервальных уравнений обычное понятие решения утрачивает смысл и чаще приходится иметь дело с теми или иными множествами решений. Например, известно, так называемое, *объединенное множество решений* $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$ для (3.1) [3], предикатная формула которого записывается так:

$$\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B}) = \{x \in R^n : (\exists A \in \boldsymbol{A})(\exists B \in \boldsymbol{B})(Ax = B)\}. \tag{3.2}$$

Можно показать, что множество неопределенности, введенное в разделе 2, в случае не интервальной матрицы коэффициентов является объединенным множеством решений системы неравенств (2.2). Если систему уравнений (3.1) рассматривать в контексте задачи анализа данных, то введенное множество решений (3.2) следует называть множеством неопределенности или информационным множеством.

ТЕМА 3.2. Использование линейного программирования для внешней аппроксимации множеств решений ИСЛАУ

Пусть в ИСЛАУ (3.1) матрица коэффициентов не является интервальной. Тогда объединенное множество решений системы неравенств (2.2) запишется так:

$$\Xi_{uni}(A, \boldsymbol{B}) = \{x \in R^n : (\exists B \in \boldsymbol{B})(Ax = B)\}. \tag{3.3}$$

Рассмотрим следующую систему линейных неравенств:

$$\begin{cases} Ax \le B^V; \\ Ax \ge B^H. \end{cases} \tag{3.4}$$

Задание 3.1. Показать, что множество $\Xi_{uni}(A, \boldsymbol{B})$, заданное выражением (3.4) совпадает с решением системы неравенств (2.2) раздела 2.

Учитывая, что множество $\Xi_{uni}(A, \boldsymbol{B})$ согласно выражения (3.4) является многогранным, возможно использование методов линейного программирования для него исследования.

В работе [4, с. 323] утверждается, что «вычисление для объединённого множества решений внешних покоординатных оценок с любой заданной абсолютной или относительной точностью есть NP-трудная задача». В частном случае положительных компонент решения ИСЛАУ объединенное множество решений задается системой линейных неравенств, которые следуя [1, с. 112] запишем ниже.

Задание 3.2. Показать, что задачи интервального оценивания решения ИСЛАУ (3.1) по типу задач (2.5) раздела 2 обеспечивают внешнюю аппроксимацию множества $\Xi_{uni}(A, \boldsymbol{B})$. Записать эти задачи в обозначениях данного раздела и системы неравенств (3.4).

Рассмотрим общий случай ИСЛАУ заданной выражением (3.1), и предположим, что объединенное множество решений $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$ содержится в R_{+}^{n}, т.е. содержит только строго положительные векторы x.

Тогда это множество является решением следующей системы линейных неравенств, аналогичных (3.4) [1, стр. 113.]:

$$\begin{cases} A^{H}x \leq B^{V}; \\ A^{V}x \geq B^{H}. \end{cases} \tag{3.5}$$

Задание 3.3. Показать, что множество $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$, определенное предикатной формулой (3.2) совпадает с решением системы неравенств (3.5) (эти определения и задание множества эквивалентны, см. работу [1, стр. 112-113.]). На рисунке 3.1 приведена схема наблюдений за переменными процесса, зависимость выходной переменной от входной является линейной, а ошибки наблюдения – интервальными: $y_0 = a_0 + a_1 \cdot x_1$; $a_0 > 0$; $a_1 > 0$. Анализ данных в этом случае сводится к исследованию объединенного множества решений ИСЛАУ (3.1). На рисунке 3.1 видно, что множество допустимых прямых $y = \beta_0 + \beta_1 \cdot x_1$, которые могут «разместиться» в точках прямоугольников описывается системой неравенств (3.5).

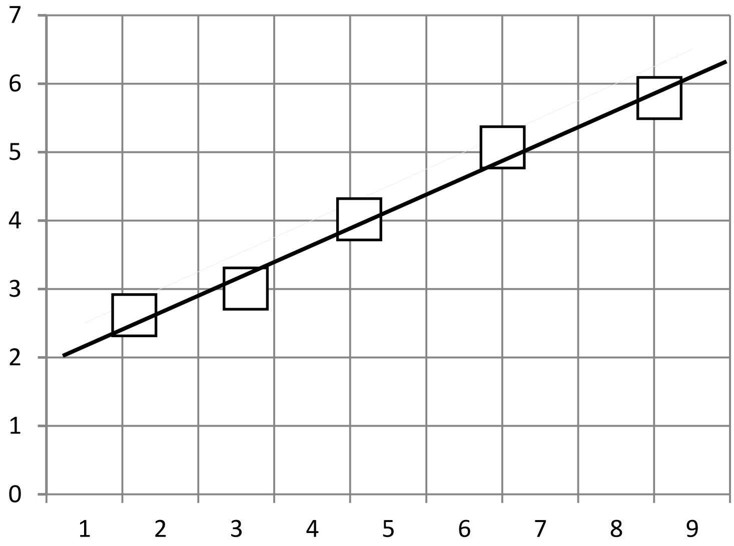


Рис. 3.1. Положение интервальных наблюдений за переменными процесса

Задание 3.4. Показать, что задачи интервального оценивания решения ИСЛАУ (3.1) по типу (2.5) раздела 2 обеспечивают внешнюю аппроксимацию множества $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$. Записать эти задачи в обозначениях данного раздела с использованием системы неравенств (3.5).

ТЕМА 3.3. Решение интервальной линейной задачи о допусках

Следуя, например [4], определим множество решений ИСЛАУ (3.1) (допусковое множество решений), которое образовано всеми такими векторами $x \in R^n$, что произведение Ax принадлежит интервальному вектору $\boldsymbol{B}$ для любого $A \in \boldsymbol{A}$. Указанное множество решений определяется так [3]:

$$\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B}) = \{x \in R^n: (\forall A \in \boldsymbol{A})(\exists B \in \boldsymbol{B})(Ax = B)\}. \qquad (3.6)$$

В задачах анализа данных указанное множество (если оно не пусто) является внутренней оценкой объединенного множества решений ИСЛАУ, а его построение носит название задачи о допуске [3].

В [3] приведен пример ИСЛАУ (рисунок 3.2) и даны графические изображения множеств $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$ и $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ ее решений (рисунок 3.3).

Теорема И. Рона [4]. Точка $x \in R^n$ принадлежит допусковому множеству решений $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ тогда и только тогда, когда $x = x' - x''$, где векторы $x', x'' \in R^n$ образуют решение системы линейных неравенств:

$$
\begin{cases}
A^V x' - A^H x'' \le B^V; \\
-A^H x' + A^V x'' \le -B^H; \\
x', x'' \ge 0.
\end{cases}
\tag{3.7}
$$

$$
\begin{pmatrix}
[1,2] & \left[-\frac{2}{3}, \frac{1}{2}\right] \\
\left[-\frac{2}{3}, \frac{1}{2}\right] & [1,2]
\end{pmatrix} x = \begin{pmatrix} [-1,1] \\ [-1,1] \end{pmatrix}.
$$

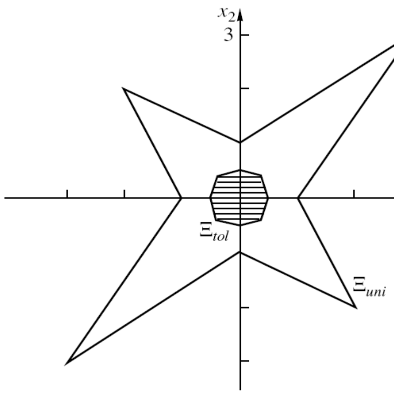


Рис. 3.2. Пример интервальной СЛАУ Рис. 3.3. Решения ИСЛАУ

Из теоремы И. Рона непосредственно следует, что множество решений $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ образовано пересечением конечного числа полупространств, является выпуклым многогранным множеством.

Для его характеристики можно использовать интервальные оценки по типу интервальных оценок раздела 2 (выражение (2.5)) и точечную оценку по типу МЦН (выражение (2.8)).

Более точные методы исследования множества $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ представлены в [4] и в ссылках в этой работе. Там же представлены методы внутренней аппроксимации максимальным брусом множества $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$. Прикладные аспекты допускового множества решений рассмотрим в следующем разделе.

ТЕМА 3.4. Управляемое множество решений ИСЛАУ

Данное решение ИСЛАУ необходимо в случаях, когда точно заданный входной вектор гарантированно дает оценку выходной переменной в заданном интервале. Предикатной формулой это решение определяется так [4]:

$$
\Xi_{ctl}(\boldsymbol{A}, \boldsymbol{B}) = \{x \in R^n : (\forall B \in \boldsymbol{B})(\exists A \in \boldsymbol{A})(Ax = B)\}. \tag{3.8}
$$

Теорема Лакеева-Носкова [4]. Точка $x \in R^n$ принадлежит управляемому множеству решений $\Xi_{ctl}(\boldsymbol{A},\boldsymbol{B})$ тогда и только тогда, когда $x = x' - x''$, где векторы $x',\ x'' \in R^n$ образуют решение системы линейных неравенств:

$$\begin{cases} A^V x' - A^H x'' \geq B^V; \\ -A^H x' + A^V x'' \geq -B^H; \\ x', x'' \geq 0. \end{cases} \tag{3.9}$$

а также условию дополнительности $(x', x'') = 0$, где (x', x'') – скалярное произведение соответствующих векторов.

Заметим, что при малых размерностях ИСЛАУ по числу неизвестных условие дополнительности в теореме Лакеева-Носкова можно учитывать путем априорного задания исследуемого квадранта декартовой системы координат. Так в случае, когда априори известно, что решение ИСЛАУ не отрицательно $(x \in R^n_+)$ в условиях системы неравенств (3.9) можно требовать $x'' = 0$. В этом случае допусковое и управляемое множества решений можно записать так:

$$\Xi_{tol}(\boldsymbol{A},\boldsymbol{B}) = \{x \in R^n_+ : A^V x \leq B^V;\ A^H x \geq B^H\}. \tag{3.10}$$

$$\Xi_{ctl}(\boldsymbol{A},\boldsymbol{B}) = \{x \in R^n_+ : A^V x \geq B^V;\ A^H x \leq B^H\}. \tag{3.11}$$

В этих случаях задачи анализа данных решаются методами линейного программирования, а выполнение условий гарантированного оценивания в вычислительных экспериментах можно контролировать решение задачи (2.6). Сравнительные исследования эффективности анализа данных с использованием рассмотренных множеств решений ИСЛАУ рассмотрим в следующем разделе.

Задание 3.5. Темы самостоятельных расчетных работ по разделу 3.

1. Для введенных множеств решений ИСЛАУ (объединенное, допусковое и управляемое множества решений) привести примеры их использования на практике и записать математические задачи анализа данных.

2. Разработать компьютерные программы генерации таблиц наблюдений переменных с ошибками для детерминированного линейного процесса, записать соответствующие ИСЛАУ и провести анализ множеств решений методами линейного программирования.

3. Пусть $x^P \in R^n_+$ – истинное значение искомого решения ИСЛАУ в задаче анализа данных, $X_s \subset R^n_+$ – одно из трех введенных множеств. Точка x^P принадлежит множеству X_s тогда и только тогда, когда $\delta_s = 0$, где δ_s – определяется решением следующей задачи квадратичного программирования: $\delta_s = \min_{x \in X_s} \|x^P - x\|$. Провести вычислительные экспери-

менты для проверки выполнения (или не выполнения) условия $\delta_s = 0$ для всех исследуемых множеств решений ИСЛАУ.

Вопросы по разделу 3.

1. Пояснить обозначения элементов матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ в ИСЛАУ.

2. Записать предикатные формулы объединённого, допускового и управляемого множеств решений ИСЛАУ.

3. Постановка и свойства задачи оценок минимальных брусов, содержащих множества решений ИСЛАУ.

4. В случае, когда решение ИСЛАУ не отрицательно, исследование объединенного множества решений ИСЛАУ можно проводить методами линейного программирования. Записать условия задач оценки минимального бруса, содержащего это решение.

5. Структура и средства имитационного моделирования для вычислительных экспериментов задачи оценки параметров линейных процессов с использованием решений ИСЛАУ.

Библиографический список к разделу 3

1. Жолен Л. и др., Прикладной интервальный анализ. – М.-Ижевск: Институт компьютерных исследований, 2005. – 468 с.

2. Шарый С.П. Конечномерный интервальный анализ. Новосибирск: ИВТ СО РАН, 2011 [Электронный ресурс]. URL:http://www.nsc.ru/interval/Library/ InteBooks/SharyBook.pdf

3. Шарый С.П. Решение интервальной линейной задачи о допусках // Автоматика и телемеханика, № 10, 2004. – С. 147 – 162.

4.Шарый С.П. Конечномерный интервальный анализ. – Новосибирск: Изд-во «XYZ». 2017. – 618 с.

РАЗДЕЛ 4. Исследование моделей и методов прикладного интервального анализа на примерах линейных процессов

ТЕМА 4.1. Задачи и инструменты исследования решений ИСЛАУ с использованием вычислительных экспериментов

В данном разделе проводится исследование возможностей применения теории интервальных систем линейных алгебраических уравнений (ИСЛАУ) для решения задач математического моделирования процессов по экспериментальным данным. Предполагается, что моделируемый процесс описывается выходной переменной и совокупностью входных переменных, детерминированная связь которых описывается линейным уравнением. При моделировании процесса используется теоретический подход, при котором исходные предположения относительно структуры модели и границы интервалов ошибок измерения всех переменных являются достоверными и не требуют проверки их выполнимости методами разведочного анализа.

Указанная постановка задачи моделирования процессов позволяет сосредоточить внимание на исследование прикладной ценности основных множеств решений ИСЛАУ: объединенного, допускового и управляемого. В целях достижения наглядности методов и полученных результатов исследование проводится на конкретном примере процесса, для которого параллельно можно проводить визуальный анализ данных.

В данном случае считается, что внутренний шум в описании объекта отсутствует, а неопределенности таблицы наблюдений связаны с ошибками измерения входных и выходной переменной, которые принадлежат известным симметричным относительно нулевого значения интервалам.

Модельным примером процесса выступает корпоративная организация, для которой проводится анализ вкладов дочерних предприятий, и прогноз суммарной прибыли [1]. Считаем, что корпорация состоит из n предприятий и заинтересована в получении прибыли. Оценка величины прибыли на очередной период времени (месяц, квартал) проводится на основе планов ее получения каждым предприятием. Управляющая компания объединения не производит продукцию и не осуществляет финансовые (инвестиционные) услуги. В случае такой возможности в корпорации создается самостоятельная компания, которая включена в состав рассматриваемых предприятий.

Пусть $A_i^{N+1}, i = 1, \dots, n$ – оценки прибыли предприятий и B^{N+1} – расчетное значение прибыли корпорации, полученное аналитиками на плановый период времени. Предполагаем, что управляющая компания располагает статистикой плановых оценок прибыли предприятий и корпорации за N временных периодов в прошлом:

$$\left(A_1^j, A_2^j, \dots, A_n^j, B^j\right), j = 1, \dots, N. \tag{4.1}$$

Кроме того, полагаем выполнение балансового уравнения по прибыли корпорации в каждом периоде наблюдения, которое запишем в следующем виде. Пусть A_i' – значения истинных величин прибыли предприятий, которые априори неизвестны как предприятиям, так и управляющей компании. Тогда действительная прибыль корпорации B' равна сумме прибылей предприятий (налоги и транзакции при переходе прав собственности на прибыль не учитываем): $B' = \sum_{i=1}^{n} A_i'$.

По результатам финансового анализа управляющая компания располагает достоверными оценками модулей предельных ошибок, которые имели место в прошлом при обосновании базы данных (4.1). Эти оценки обозначим так:

$$\left(\varepsilon_1^j, \varepsilon_2^j, \dots, \varepsilon_n^j, \varepsilon_B^j\right), j = 1, \dots, N. \tag{4.2}$$

Экономическая ситуация и существующие технологии финансового планирования корпорации позволяют ее аналитикам обосновать оценку будущей прибыли корпорации B^{N+1} оценить границы абсолютных значений погрешностей планирования прибыли предприятий и корпорации в целом:

$$(\varepsilon_1^{N+1}, \varepsilon_2^{N+1}, \dots, \varepsilon_n^{N+1}, \varepsilon_B^{N+1}), j = 1, \dots, N. \tag{4.3}$$

Далее учитывается, что приведенные данные соответствуют правильным наблюдениям [7 – 10], например, не содержат выбросов. Методами прикладного анализа данных предлагается выполнить комплекс аналитических исследований:

1. Получить интервальную оценку $\boldsymbol{B}^{N+1}$ прибыли корпорации при существующей технологии финансового планирования, которая является достоверной, т.е. содержит истинное значение будущей прибыли корпорации B'^{N+1} .

2. Проверить согласованность существующей технологии расчета прибыли корпорации B^{N+1} его интервальной оценке, т.е. проверить условие $B^{N+1} \in \boldsymbol{B}^{N+1}$.

3. Качественно оценить эффективность решения задач математического моделирования процессов с использованием множеств решений ИСЛАУ: объединенного $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$, допускового $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ и управляемого $\Xi_{ctl}(\boldsymbol{A}, \boldsymbol{B})$ [3, 5].

Исследование выполняется методами вычислительного эксперимента с использованием математических и программных средств раздела 3.

ТЕМА 4.2. Имитационное моделирование оценки прибыли корпорации

Имитация условий анализа прибыли корпорации проводилась в среде Excel при следующих исходных данных: $n = 3$; $N = 12$. Истинные значения прибылей в каждом из временных периодов принимались как равномерно распределенные псевдослучайные числа на интервалах [0, 100]. Соответствующие значения прибыли корпорации определялись на основе балансного уравнения. Ошибки измерения во всех испытаниях принимались случайными величинами, равномерно распределенными на одинаковых симметричных относительно нуля интервалах:$[-5,\ 5]\text{x}[-5,\ 5]\text{x}[-5,\ 5]\text{x}[-\varepsilon_b, \varepsilon_b]$.

Здесь верхние оценки ошибки выходной переменной задавались переменными для обеспечения условия, при котором исследуемое множество решений ИСЛАУ не пусто и сравнимо с двумя другими. Пример точных значений прибылей корпорации и их оценок с заданными погрешностями измерений ($\varepsilon_b = 10$) представлен в таблицах 4.1 и 4.2.

Ниже представлены результаты многовариантных вычислительных экспериментов исследования эффективности оценок последовательно для всех трех множеств решений ИСЛАУ. Итоговые сравнения и выводы представим в следующей теме.

Таблица 4.1. – Данные без ошибок измерения

	A_1'	A_1'	A_1'	B'
1	23,93	19,84	97,14	140,91
2	13,84	91,08	65,11	170,03
3	35,27	18,48	70,67	124,42
4	68,21	39,87	84,02	192,09
5	76,78	22,20	13,35	112,33
6	24,28	88,86	64,10	177,23
7	11,60	6,28	26,29	99,18
8	12,92	49,01	83,69	145,62
9	91,99	26,19	21,89	140,07
10	29,43	61,70	16,77	107,90
11	58,73	80,64	95,76	235,12
12	83,98	89,17	5,45	178,60
13	**17,01**	**73,08**	**51,91**	**142,00**

Таблица 4.2. – Данные с ошибками измерения, при $\varepsilon_b = 10$

	A_1	A_1	A_1	B
1	26,99	19,99	98,11	149,74
2	17,80	93,8	68,41	160,7
3	38,41	14,64	67,47	127,32
4	67,48	38,84	88,28	191,76
5	75,53	23,05	11,26	113,07
6	19,37	83,87	64,64	182,70
7	16,57	57,07	30,77	94,54
8	17,00	47,84	79,19	150,69
9	89,93	27,74	22,38	147,71
10	24,47	65,00	18,49	107,10
11	57,86	84,12	95,31	240,01
12	79,54	89,72	4,98	169,15
13	**15,94**	**73,27**	**47,94**	**135,37**

Рассматриваем анализ данных по прибыли корпорации с использование множества $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$. Объединенное множество решений ИСЛАУ содержит точку $x^d = (1, 1, 1)$ по определению этого множества для правильных измерений, что следует и по результатам численного решения задачи (2.6).

Для данных таблицы 4.2 решение задач (2.3) является следующим:

$$\hat{B}^{H,N+1} = 109{,}7;\ \hat{B}^{V,N+1} = 160{,}3;\ B^{C,N+1} = 135;\ \Delta = 17{,}8\%. \qquad (4.4)$$

В выражении (4.4) представлены прогнозные значения прибыли корпорации на 13 (N+1=13) период времени, в том числе: $\hat{B}^{H,N+1}$, $\hat{B}^{V,N+1}$, $B^{C,N+1}$ – нижнее, верхнее и среднее значения прогнозной прибыли соответственно, где $B^{C,N+1} = 0{,}5 \cdot \left(\hat{B}^{H,N+1} + \hat{B}^{V,N+1}\right)$; Δ – погрешность прогноза в процентах.

Для сравнения приведем значения и погрешности прогноза по оценкам аналитиков. Эти данные в порядке выражения (4.4.) имеют следующие значения: 125,4; 145,4; 135,4; 7,0%. Приведем соответствующие значения, полученные с использование балансового уравнения по нижним и верхним значениям прибылей предприятий: 122,2; 152,2; 137,2; 10,6%. Приведенные числовые данные не противоречат визуальному анализу и свойствам объединенного множества решений.

Следует отметить, что использование объединенного множества решений ИСЛАУ в качестве инструмента прогноза прибыли корпорации в рассмотренном случае не позволяет улучшить оценки, полученные визуальным анализом данных.

Оценки параметров балансного уравнения в предположении, что точка $x^d = (1, 1, 1)$ аналитикам неизвестна, получим решением задач (2.6) (таблица 4.3).

Таблица 4.3 – Оценки объединенного множества решений ИСЛАУ

Оценки $\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B})$	Индекс параметров		
	$i = 1$	$i = 2$	$i = 3$
Нижнее значение	0,77	0,57	0,72
Верхнее значение	1,45	1,26	1,39
Погрешность, %	34,1%	34,2%	33,5%

Рассматриваем оценки прибыли корпорации с использованием множества $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$. Допусковое множество решений ИСЛАУ для данных таблицы 4.2 и заданных предельных значений погрешностей измерения является пустым. Этот результат вполне согласуется с исследованиями автора работ [2, 3], т.е. в данном случае произведение $\boldsymbol{Ax}$ получает «большой размах» в сравнении с размахом вектора $\boldsymbol{B}$.

В нашем случае для данных таблицы 4.1 и принятых оценок погрешностей измерения, в которых $\varepsilon_b = 20$ с дополнительным коэффициентом уширения $k_b = 2$ размаха вектора $\boldsymbol{B}$ получена ИСЛАУ с подходящими свойствами. Ее исследование для 15 вариантов выборок независимых ошибок измерения показало, что в 5 вариантах допусковое множество не пусто и содержит точку $x^d = (1, 1, 1)$; в 8 вариантах оно не пусто, но необходимое для корректности модели условие $x^d \in \Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ не выполнено; в 2 вариантах множество $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ оказалось пустым.

Для первого варианта расчетных выборок приведем решение задач (2.3):

$$\hat{B}^{H,N+1} = 121{,}1;\ \hat{B}^{V,N+1} = 162{,}5;\ B^{C,N+1} = 141{,}8;\ \Delta = 14{,}6\%. \qquad (4.5)$$

Для сравнения, как и ранее, приведем анализ погрешностей прогноза оценок параметров балансного уравнения (таблица 4.4) аналогичные результатам для $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$. Эти данные имеют следующие значения: оценки аналитиков – 124,8; 144,8; 134,8; 7,0% при $\varepsilon_b = 10$; оценки по балансу –122,4; 152,4; 137,4; 10,6%.

Таблица 4.4 – Оценки допускового множества решений ИСЛАУ

Оценки $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$	Индекс параметров		
	$i=1$	$i=2$	$i=3$
Нижнее значение	0,85	1,00	0,93
Верхнее значение	1,03	1,16	1,07
Погрешность, %	8,2%	8,7%	6,9%

Сравнение данных таблиц 4.3 и 4.4 показывает безусловную эффективность процедур моделирования процессов с использованием допускового множества решений. Данный вывод согласуется с результатами работы [6] и полностью оправдывает его метод сильной совместимости. Однако, следует учитывать, что в ряде задач идентификации процессов необходимо выполнение условия принадлежности истинных значений параметров процесса допусковому множеству решений [1, 4, 7].

Заметим, что прямая проверка этого включения встречает существенные трудности. Сходная проблема возникает при оценке информационного множества в задачах моделирования процессов при неправильных наблюдениях [7, с. 53].

Рассмотрим формально по аналогии с вышеизложенным свойства управляемого множества решений ИСЛАУ для данных таблицы 4.2 и заданных предельных значений погрешностей. Следует отметить, что примеры использования множества $\Xi_{ctl}(\boldsymbol{A}, \boldsymbol{B})$ в задачах моделирования процессов авторам данного учебного пособия неизвестны.

Как и следовало ожидать, это множество для рассматриваемой таблицы измерений является пустым. Поступая зеркально схеме исследования множества $\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B})$ зададим $\varepsilon_b = 5$ с дополнительным коэффициентом сжатия размаха вектора $\boldsymbol{B}$ равным 0,6 ($k_b = 0{,}6$) и получим ИСЛАУ с подходящими свойства. Ее исследование для 15 вариантов выборок с независимыми ошибками измерений показало, что в 11 вариантах управляемое множество не пусто и содержит точку $x^d = (1, 1, 1)$; в 4 вариантах оно не пусто, но необходимое для корректности модели условие $x^d \in \Xi_{ctl}(\boldsymbol{A}, \boldsymbol{B})$ не выполнено.

Для одного из первой группы вариантов приведем решение задач (2.3):

$$\hat{B}^{H,N+1} = 119{,}1; \hat{B}^{V,N+1} = 160{,}95; B^{C,N+1} = 140; \Delta = 14{,}7\%. \qquad (4.6)$$

Как видим, эти данные с учетом точности их вычисления совпадают с (4.5), как и оценки погрешностей с использованием визуальных методов прогнозирования.

Равнозначность по эффективности моделирования процессов показывают оценки параметров балансового уравнения, полученные как характеристика бруса, содержащего управляемое множество решений (таблица 4.5). Мы считаем различия данных таблиц 4.4 и 4.5. несущественными, поскольку задание управляемого множества решений оказалось более свободным.

Таблица 4.5 – Оценки управляемого множества решений ИСЛАУ

Оценки $\Xi_{ctl}(\boldsymbol{A}, \boldsymbol{B})$	Индекс параметров		
	$i = 1$	$i = 2$	$i = 3$
Нижнее значение	0,87	0,83	0,89
Верхнее значение	1,16	1,17	1,14
Погрешность, %	14,6%	16,8%	12,4%

ТЕМА 4.3. Итоговые результаты сравнения оценок прибыли корпорации методами интервального анализа данных

Следует отметить, что результаты вычислительного эксперимента показывают, что в данном случае полезность итоговых интервальных оценок для практики является низкой, как в силу значительных ошибок, так и доминирование результатов визуального анализа. Это обстоятельство вызвано тем, что все погрешности хаотичны, а при анализе данных не учитывается априорная информация. Реально для рассматриваемого примера в качестве такой информации могут выступать, во-первых, множество X_0 содержащее точку x^d и, во-вторых, оценки интервалов прибыли корпорации, полученные с учетом прогноза прибыли предприятий.

Выполнено исследование множеств решения ИСЛАУ, в которой интервалы правой части получены на основе балансового уравнения. Как и следовало ожидать, допусковое и управляемое множества решений ИСЛАУ не пусто и одноэлементное, т.е. содержит одну точку $x^d = (1, 1, 1)$. Характеристика объединенного множества решений качественно соответствует таблице 4.5. Прогнозные значения выходной переменной соответствуют результатам визуального анализа данных. Данный пример не имеет прикладной ценности и носит чисто теоретический характер.

В статье рассмотрены задачи моделирования процессов по экспериментальным данным с интервальными ошибками измерения как входных, так и выходной переменных. При статистическом подходе математическое моделирование процессов выполняется методами конфлюэнтного анализа.

Для процесса без внутренних шумов в условиях правильных наблюдений методами вычислительного эксперимента проведено качественное сравнение точностей прогнозирования выходной переменной и оценивания коэффициентов линейной связи для трех множеств решений ИСЛАУ: объединенного, допускового и управляемого.

При решении задачи интервального анализа коэффициентов линейной зависимости подтверждена эффективность использования принципа сильного согласования С.П. Шарого [2, 6] при условии корректной оценки их истинных значений, которое в статистических методах обозначается понятиями несмещенности и состоятельности. В процессе исследованиях затронуты проблемы интервального анализа зависимых измерений (измерений, в которых ошибки взаимно зависимы), учета априорной информации при записи ИСЛАУ и поиске ее решения.

Полученные результаты позволяют уточнить особенности применения теоретических результатов ИСЛАУ в задачах анализа данных, в том числе перспективы оценивания параметров с использованием допускового множества решения. Эти вопросы рассматриваются в следующих темах данного раздела.

ТЕМА 4.4. Согласование базы данных и условий допускового множества решений ИСЛАУ

Рассматриваем возможность согласования результатов наблюдений с целью корректности оценок параметров линейных процессов с использованием допускового множества решений ИСЛАУ. В работе [12] показано, что значения совокупности входных переменных и выходной переменной согласованы, если график искомой зависимости расположен во всех внутренних точках интервального бруса неопределенности для каждого наблюдения (Приложение 2). Однако, в реальных и модельных условиях указанное согласование базы данных априори не всегда выполняется. В этом случае предлагается использовать принцип робастного оценивания: не согласованные наблюдения следует либо исключить из выборки, либо скорректировать [12, 14]. В данной работе далее представлены результаты исследования этих способов согласования используемой экспериментальной базы данных на модельных линейных процессах в условиях, когда базовые предположения интервального оценивания зависимостей выполняются. Многовариантные вычислительные эксперименты показали возможность повышения точности интервального анализа за счет предварительной корректировки наблюдений, в том числе возможность гарантированного оценивания искомых зависимостей [12].

Предполагается, что моделируемый процесс описывается выходной переменной и совокупностью входных переменных, измерения которых представлены в базе данных интервальными величинами, а структура модели и границы интервалов ошибок измерения всех переменных являются достоверными. Этот базовый вариант интервального анализа данных обоснован в разделе 3 данного учебного пособия, а в [13] обсуждены возможности гарантированного оценивания параметров исследуемой зависимости переменных моделируемого процесса.

В настоящее время прикладной интервальный анализ экспериментальных данных не встречает существенной критики, однако его применение на практике сопряжено с проблемами, которые исследованы в данном разделе. Главные из этих проблем, во-первых, высокая погрешность гарантированных оценок, и, во-вторых, высокие требования к исходным предположениям.

Первую проблему поясним на примере оценки погрешности суммы большого числа интервальных данных. Гарантированная оценка суммы определится интервалом сумм нижних и верхних оценок слагаемых и уже при нескольких десятках слагаемых теряет прикладное значение. При анализе экспериментальных данных обоснованы некоторые методы «усечения» итоговых интервалов, один из которых предложил Шарый С.П. с общим названием «сильная совместимость» [6].

Метод основан на использовании допускового множества решений ИСЛАУ и характеризуется тем, что точечные оценки восстановления линейных зависимостей являются эффективными. Принцип сильной совместимости базируется на следующей гипотезе [6]: «Если процесс измерения входа и выхода разделен во времени …, то более адекватно … понимание «согласования», при котором ограничение на выходе должно выполняться равномерно при любых значениях входов».

Однако, в ряде случаев допусковое множество решений ИСЛАУ даже базовом варианте оказывается пустым. В этом случае для точечного оценивания параметров в работе [3] используется метод уширения брусов неопределенности данных (далее метод уширения), который можно рассматривать как один из методов согласования базы данных. Исследования показали что, допусковое множество решений ИСЛАУ, в общем случае может не содержать истинных значений оцениваемых параметров, т.е. не выполняется принцип гарантированного оценивания. Таким образом, проблема согласования базы данных в прикладном интервальном анализе требует дополнительных исследований.

В данной работе используется подход, при котором гипотеза согласования интервальных наблюдений рассматривается как дополнение к базовым предположениям прикладного интервального анализа. Мы считаем, что сформулированная в [6] гипотеза согласования очевидно **верна для истинных значений входных переменных** в каждом из N опытов, при условии, что исходные предположения метода выполнены, т.е. искомая зависимость линейна, внутренние шумы отсутствуют, а погрешности измерений оценены правильно.

Однако в реальных и модельных условиях указанное согласование базы данных априори не всегда выполняется. В этих случаях предлагается использовать принцип робастного оценивания [12, 14]: не согласованные наблюдения следует либо исключить из выборки, либо скорректировать. Ниже мы представим результаты исследования этих способов согласования используемой экспериментальной базы данных на модельных линейных процессах в условиях, когда базовые предположения интервального оценивания зависимостей выполняются. Методической основой коррекции наблюдений является возможность использования априорной информации при анализе данных, примеры и приемы которой описаны в работах [7 стр. 51].

Многовариантные вычислительные эксперименты показали возможность повышения эффективности интервального анализа за счет предварительной корректировки наблюдений, в том числе возможность гарантированного оценивания параметров искомых зависимостей.

ТЕМА 4.5. Описание методов и методик планируемого исследования

Как отмечены выше в разделе 3 ИСЛАУ в матричной форме записывается интервальной ($N \times n$) матрицей коэффициентов и интервальным ($N \times 1$) вектором правой части в следующем виде: $\boldsymbol{A}x = \boldsymbol{B}$.

Элементы матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ являются интервальными оценками результатов измерения входных и выходной переменной в N наблюдениях и задаются неравенствами: $A^H \le \boldsymbol{A} \le A^V$; $B^H \le \boldsymbol{B} \le B^V$.

Значения вектора $x \in R^n$ в ИСЛАУ соответствует оценками параметров искомой линейной зависимости:

$$b = x_1 a_1 + \cdots + x_n a_n \qquad (4.7)$$

Интервальные наблюдения $\boldsymbol{a}_{ij}, \boldsymbol{b}_j$ за переменными моделируемого процесса обозначим так:

$$\boldsymbol{a}_{ij} = \left[a_{ij}^H,\ a_{ij}^V\right];\ a_{ij}^H = a_{ij}^M - \varepsilon_{ij}^0;\ \ a_{ij}^V = a_{ij}^M + \varepsilon_{ij}^0;\ i = 1, \dots n;\ j = 1, \dots, N; \quad (4.8)$$

$$\boldsymbol{b}_j = \left[b_j^H,\ b_j^V\right];\ b_j^H = b_j^M - \varepsilon_{bj}^0;\ \ b_j^V = b_j^M + \varepsilon_{bj}^0;\ j = 1, \dots, N, \quad (4.9)$$

где a_{ij}^M, b_j^M – экспериментальные данные входов и выхода в каждом из N наблюдений; ε_{ij}^0, ε_{bj}^0 не отрицательные оценки ошибок наблюдения (ниже индекс j будет опущен, т.е. интервальные ошибки приняты одинаковыми для всех наблюдений); a_{ij}^H, a_{ij}^V, b_j^H, b_j^V – элементы матриц A^H, A^V, B^H, B^V соответственно.

В данной работе мы ограничимся сравнение оценок объединенного и допускового множеств решений ИСЛАУ, считая, что элементы этих множеств не отрицательны. Объединенное множество решений зададим системой линейных неравенств, которые запишем в следующем виде:

$$\Xi_{uni}(\boldsymbol{A}, \boldsymbol{B}) = \{x \in R_+^n\colon A^V x\ \geq B^H;\ A^H x\ \leq B^V\}. \quad (4.10)$$

Допусковое множество решений ИСЛАУ зададим так:

$$\Xi_{tol}(\boldsymbol{A}, \boldsymbol{B}) = \{x \in R_+^n\colon A^V x\ \leq B^V;\ A^H x\ \geq B^H\}. \quad (4.11)$$

Корректирование всех наблюдений проводим в следующем виде:

$$\hat{a}_{ij}^M = a_{ij}^M + e_{ij}\varepsilon_j^0;\ \hat{b}_j^M = b_j^M + e_{bj}\varepsilon_b^0;\ \left|e_{ij}\right| \leq 1;\ \left|e_{bj}\right| \leq 1. \quad (4.12)$$

Скорректированные согласно (4.12) экспериментальные принадлежат интервалам в выражениях (4.8), (4.9) и мы считаем, что оценки ошибок их измерений не меняются. Обозначим как E множество значений $N(n+1)$ коэффициентов корректирования, а через $\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}$, $\Xi_{uni}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E)$, $\Xi_{tol}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E)$ интервальные матрицы ИСЛАУ и множества их решений согласно выражениям, аналогичным (4.10), (4.11) для базы данных (4.12).

Рассмотрим для обсуждения подхода один из примеров задачи коррекции базы данных. Рассматриваем вариант ИСЛАУ, для которой допусковое множество решений является пустым. Это обстоятельство является достаточным основанием считать отдельные наблюдения базы данных не согласованными. Необходимо найти коэффициенты корректирования $e^* \epsilon E$ и оценки вектора

$x^*\epsilon\Xi_{tol}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},E)$ в перечисленных выше задачах прикладного интервального анализа. В следующем разделе мы рассмотрим решение данной задачи в один и в два этапа. В качестве примера рассмотрим одноэтапную оценку $b^H(a^P)$ при решении задачи прогноза (2.3):

$$b^H(a^p) = \min_{\substack{x\in\Xi_{tol}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},E)\\ e\in E}}\left(\left(x_1a_1^p + \cdots + x_na_n^p\right) + L\|e\|\right). \qquad (4.13)$$

В данном случае для коррекции базы данных и оценки параметров записана двухкритериальная задача нелинейного программирования. При достаточно большом значении параметра L обеспечивается согласование базы данных $\left(\Xi_{tol}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},E) \neq \emptyset\right)$ и минимальная коррекция значений экспериментальных данных. Задачи типа (4.13) имеют большое число переменных и ограничений и требуют на практике применение специальных методов их решения [15].

ТЕМА 4.6. Имитационное моделирование процессов согласования базы экспериментальных данных

Имитация коррекции данных проводилась в среде Excel при использовании инструмента «Поиск решения». Возможности этого инструмента позволили исследовать задачи и методы анализа данных линейных процессов с двумя входами ($n = 2$) при двадцати наблюдениях ($N = 20$).

Генерация значений входных переменных и ошибок наблюдений осуществлялась с использованием функции Excel СЛЧИС() в задаваемых пределах. Вычислительные эксперименты проведены для линейного процесса (4.7) с параметрами $x_1 = x_2 = 1$; $a_1, a_2 \in [10,\ 100]$. Задача прогноза (2.3) исследовалась для $a^p = (55,\ 55)$. Многовариантность вычислительных экспериментов достигалась изменением интервалов ε_1^0, ε_2^0, ε_b^0, обновлением значений входов и ошибок. Исследование каждого метода согласования базы данных проводилось с большим числом повторений модельных испытаний.

В описанной инструментальной среде проведены исследования следующих методов согласования баз данных: исключение не согласованных наблюдений (метод исключения); согласование базы данных по критерию минимума нормы $\|e\|$ (метод минимальной коррекции); согласование по типу задачи

(4.13) (метод совместной коррекции); согласование как самостоятельный этап анализа данных (метод предварительной коррекции).

Методики и критерии сравнительных исследований методов согласования индивидуальны и описаны ниже.

1. Метод исключения. Исследование метода проведено для базового процесса при двух интервальных оценок ошибок наблюдения (таблицы 4.6, 4.7). Все исходные базы данных содержали не согласованные наблюдения. Выделение этих наблюдений проводилось указанным во введении методом уширения оценки ε_b^0. Значение начального коэффициента уширения *Кр* для исходной базы данных приведены в таблицах 4.6 и 4.7. Исключение не согласованных наблюдений проводилось до единичного *Кр* и их число *Nис* изменялось от 2 до 10. Для исходной и скорректированной баз данных получены интервальные оценки прогноза (2.3) на объединенных множествах решений (ОМР) ИСЛАУ. Точность прогноза характеризуется относительной погрешностью оценки среднего значения и оценкой доверительного интервала для исходной (*Дп%, Дд%*) и скорректированной (*Дпс%, Ддс%)* баз данных соответственно.

Таблица 4.6.

Согласование данных методом исключения ($\varepsilon_1^0 = 0{,}1; \varepsilon_2^0 = 0.5; \varepsilon_b^0 = 2$)

№ опыта	Метод исключения				ОМР	
	Кр	*Nис*	*Дпс%*	*Ддс%*	*Дп%*	*Дд%*
1	1,2	5	0,31	1,41	0,18	0,92
2	1,18	2	0,38	1,27	0,04	0,93
3	1,14	2	0,42	1,39	0,18	0,96
4	1,29	4	0,23	0,98	0,09	0,47
Сред	1,20	3,25	0,34	1,26	0,12	0,82

Результаты позволяют рекомендовать для практики метод исключения при малых ошибках измерения входных переменных. Так для ошибок таблицы 1 в трех опытах из 4 обеспечивалось гарантированное оценивание допускового множества решений, что существенно повышает точность оценок. Для ошибок таблицы 2 все множества $\Xi_{tol}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E)$ не содержали точку $x_1 = x_2 = 1$.

Таблица 4.7.

Согласование данных методом исключения ($\varepsilon_1^0 = 1; \varepsilon_2^0 = 1; \varepsilon_b^0 = 4$)

№ опыта	Метод исключения				ОМР	
	Кр	*Nис*	*Дпс%*	*Ддс%*	*Дп%*	*Дд%*
1	1,41	6	1,86	2,6	0,2	1,62
2	1,61	9	1,21	3,43	0,57	1,07
3	1,45	8	1,13	3,19	0,23	2,07
4	1,53	10	1,32	2,95	0,01	1,48
Сред	1,50	8,25	1,38	3,04	0,25	1,56

2. Метод минимальной коррекции. В данном методе коэффициенты корректирования $e^* \epsilon E$ и оценки вектора $x^* \epsilon \Xi_{tol}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E)$ вычисляются решением задачи квадратичного программирования по критерию минимума нормы $\|e\|$. Полученная согласованная база данных может использоваться для точечной оценки коэффициентов искомой зависимости. Для базового процесса и ошибок как в таблице 2 проведена серия 10 испытаний не согласованных баз данных. Среднее число скорректированных наблюдений оказалось равным 7,3, погрешность прогноза по среднему значению равна 0,14% и соответствует точности методу уширения [1]. Метод минимальной коррекции можно рекомендовать на практике для согласования наблюдений при проведении натурных экспериментов.

3. Метод совместной коррекции. Исследования проведены для задачи (2.3) с параметрами процесса как в методе п.2. Результаты испытаний приведены в таблице 4.8, где *Nк* – число скорректированных наблюдений.

Таблица 4.8.

Исследование метода совместной коррекции ($\varepsilon_1^0 = 1; \varepsilon_2^0 = 1; \varepsilon_b^0 = 4$)

№ опыта	Метод совместной коррекции				ОМР	
	Кр	*Nк*	*Дпс%*	*Ддс%*	*Дп%*	*Дд%*
1	1,52	8	0,31	0,23	1,15	0,33
2	1,47	7	0,58	0,24	1,95	0,11
3	1,64	9	0,55	0,05	2,26	0,37
4	1,54	10	0,32	0,05	1,91	0,07
Сред	1,54	8,50	0,44	0,14	1,82	0,22

Во всех опытах выполнены условия гарантированного оценивания на допусковом множестве решений скорректированной базы данных. Точечные и интервальные оценки прогноза характеризуются высокой точностью в сравнении с оценками объединенного множества решений. Недостатком метода является то, что отсутствует итоговая скорректированная база данных. Следующий метод согласования устраняет этот недостаток.

4. Метод предварительной коррекции. Идея метода состоит в том, чтобы максимально раздвинуть границы матриц ИСЛАУ при минимуме нормы $\|e\|$, что требует постановки задачи оптимизации по двум критериям. В нашем случае скорректированная база данных обеспечивает максимум разности верхней и нижней оценок в задаче интервального прогноза. Результаты исследования скорректированной базы представлены в таблице 4.9, обозначение переменных в которой аналогично обозначениям таблицы 4.8.

Таблица 4.9.

Исследование метода предварительной коррекции ($\varepsilon_1^0 = 1; \varepsilon_2^0 = 1; \varepsilon_b^0 = 4$)

№ опыта	Метод предварительной коррекции				ОМР	
	Кр	*Nк*	*Дпс%*	*Ддс%*	*Дп%*	*Дд%*
1	1,31	11	0,01%	2,09%	0,42%	2,48%
2	1,53	9	1,32%	2,34%	1,55%	2,37%
3	1,56	7	0,45%	2,14%	0,26%	2,54%
4	1,61	10	0,03%	2,22%	0,30%	1,94%
Сред	1,50	9,25	0,45%	2,20%	0,63%	2,33%

Исследуемый метод коррекции сравним с методом совместной коррекции, а новая база данных может использоваться как исходная при решении задач прикладного интервального анализа.

Как итоговое обсуждение заметим, что для решения проблемы согласования базы данных в прикладном интервальном анализе, возможно впервые, реализован принцип робастного оценивания: не согласованные наблюдения следует либо исключить из выборки, либо скорректировать. Подходами вычислительной математики и компьютерного моделирования исследованы четыре метода согласования базы данных, которые обладают элементами научной новизны и имеют практическую направленность. Многовариантные вычислительные эксперименты показали возможность повышения точности интервального анализа за счет предварительной корректировки наблюдений, в том числе возможность гарантированного оценивания искомых зависимостей.

Вопросы по разделу 4.

1. Пояснить методы и задачи сравнения оценок параметров линейных процессов с использованием решений ИСЛАУ.

2. Записать модель формирования прибыли корпорации описать свойства базы данных и базы знаний при аналитических исследованиях прибыли на заданном временном интервале.

3. Описать компьютерную модель оценки прибыли корпорации с использованием множеств решений ИСЛАУ.

4. Пояснить условия гарантированного оценивая параметров на примере оценок прибыли корпорации.

5. Провести сравнительный анализ свойств оценок линейных процессов с использованием множеств решений ИСЛАУ на примере оценок прибыли корпорации.

6. Пояснить задачи согласования базы данных и теоретические основания их решения.

7. Записать модель коррекции базы данных и пояснить свойства экстремальных задач для ее вычислений на примере коррекции по критерии минимальной нормы.

8. Рассмотреть процедуры поиска и исключения не согласованных наблюдений при коррекции базы данных методом исключения.

9. Записать и пояснить математические задачи коррекции экспериментальных данных в трех описанных в данном разделе методах согласования баз данных.

10. Структура и средства имитационного моделирования для вычислительных экспериментов задач коррекции наблюдений линейных процессов с использованием допускового множества решений ИСЛАУ.

Библиографический список к разделу 4.

1. Мадияров М.Н., Оскорбин Н.М., Суханов С.И. Примеры интервального анализа данных в задачах моделирования процессов // Известия Алтайского государственного университета. 2019. № 1 (105). DOI 10.14258/izvasu(2018)1-20.

2. Шарый С.П. Решение интервальной линейной задачи о допусках // Автоматика и телемеханика, № 10, 2004. – С. 147-162.

3. Шарый С.П. Конечномерный интервальный анализ. – Новосибирск: Изд-во «*XYZ*». 2017. – 618 с.
4. Оскорбин Н.М, Жилин С.И., Максимов А.В. Построение и анализ эмпирических зависимостей методом центра неопределенности // Известия АГУ. 1998. № 1. – С. 35-38.
5. Жолен Л. и др., Прикладной интервальный анализ. – М.-Ижевск: Институт компьютерных исследований, 2005. – 468 с.
6. Shary, S.P.: Weak and strong compatibility in data fitting problems under interval uncertainty. Advances in Data Science and Adaptive Analysis, 12(1), 2050002 (2020). https://doi.org/10.1142/S2424922X20500023.
7. Максимов А.В., Оскорбин Н.М. Многопользовательские информационные системы: основы теории и методы исследования. – 2-е изд. испр. и доп. – Барнаул: Изд-во Алт. ун-та, 2013. – 264 с.
8. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: Облинформпечать, 1996. – 320 с.
9. Поляк Б.Т., Назин С.А. Оценивание параметров в линейных многомерных системах с интервальной неопределенностью // Проблемы управления и информатики. 2006. № 1. С. 103–115.
10. Zhilin, S.I. Simple method for outlier detection in źtting experimental data under interval error // Chemometrics and Intellectual Laboratory Systems. 2007. Vol. 88. No. 1. – P. 60–68.
11. Milanese M., Norton J., Piet-Lahanier H., Walter E. Bounding Approaches to Sys-tem IdentiƠcation (Eds). New York: Plenum Press, 1996. – 567 p.
12. Ергалиев Е.К., Мадияров М.Н., Оскорбин Н.М., Смолякова Л.Л. Согласование базы данных в прикладном интервальном анализе // Известия Алтайского государственного университета. 2022, № 1 (123). – С. 90-94. DOI: 10.14258/izvasu(2022)1-141.

13. Шелудько А.С. Гарантированное оценивание параметров дискретных моделей хаотических процессов // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2018. Т. 7, № 1. – С. 25-39. DOI: 10.14529/cmse180103.
14. Хьюбер П. Робастность в статистике. – М.: Мир, 1984. – 304 с.
15. Оскорбин Н.М. Вычислительные технологии анализа больших данных методами линейного программирования // Высокопроизводительные вычислительные системы и технологии. 2021. Т. 5, № 1. – С. 222-228.

РАЗДЕЛ 5. Прикладной интервальный анализ больших данных

ТЕМА 5.1. Обоснование моделей и методов анализа больших данных

Методы анализа больших данных получили в последние годы существенное развитие в связи с новыми возможностями сбора, хранения, передачи ситуационной информации и использованием технологии искусственного интеллекта для поддержки приятия решений на их основе [1]. Для многих приложений анализ больших данных требует применения схем распараллеливания в вычислительных алгоритмах и использования высокопроизводительных вычислительных систем [2].

В данной области исследований в настоящее время выделяют два этапа. Первый этап связан с решением многообразных задач структурирования информационного потока данных, включая выявление существенных факторов, параметризацией качественных оценок, тестов на естественном языке, изображений с целью получения цифровой таблицы входных и выходных переменных моделируемого процесса. На втором этапе проводится исследование причинно-следственных связей, оценки адекватности и работоспособности полученной эмпирической модели [2, 3]. Среди методов второго этапа, рассматриваемого в данной работе, используются как классические теоретико-вероятностные методы, так и относительно новые, в частности, прикладной интервальный анализ [4 – 6].

Исходная идея и первые приложения интервального подхода к анализу данных представлены в работе [4] для оценки параметров линейной зависимости выходной переменной, измеряемой с интервальной ошибкой, от n входных переменных точно измеренных в каждом из N испытаний. Обработка таблицы экспериментальных данных модели проведена решением специальных задач линейного программирования (ЗЛП).

В общем случае задача построения линейных по параметрам зависимостей с интервальными ошибками измерения всех переменных сводится к решению интервальных систем линейных алгебраических наблюдений (ИСЛАУ). Среди множества решений ИСЛАУ, известных в литературе [6], выбрано объединенное множество решений, которое обеспечивает гарантированное оценивание искомых зависимостей [7] и допускает применение методов линейного программирования (ЛП), в том числе для решения задач оптимизации большой размерности [8 – 11].

Обоснование технологий прикладного интервального анализа предварительно структурированных больших данных проводится в следующем порядке:

- рассматриваются теоретические основы прикладного интервального анализа при моделировании линейных стационарных процессов по экспериментальным данным при выполнении исходных предположений относительно структуры модели процесса и условий наблюдения его входных и выходной переменных;
- предлагаются способы последовательного анализа большой базы экспериментальных данных в одном компьютере и параллельного анализа данных в высокопроизводительных вычислительных системах;
- проводится компьютерное моделирование предложенных технологий.

С использованием рассмотренного подхода при анализе больших данных получены интервальные оценки выходной переменной моделируемого процесса при заданных значениях входных переменных и интервальные оценки параметров исследуемой зависимости. Предложены две вычислительные технологии последовательного считывания ограничений и параллельных компьютерных вычислений при решении задач линейного программирования большой размерности. Оптимальность вычислений обоснована с использованием приема релаксации ограничений при решении задач оптимизации большой размерности [8, 9, 11]. Проведено компьютерное моделирование предложенных технологий с целью изучения возможности их использования на практике и для оценки погрешностей их приближенных реализаций. Исследование проведено на модельных процессах в условиях выполнимости предположений интервального анализа данных.

ТЕМА 5.2. Вычислительные технологии интервального анализа больших данных

Рассматриваем последовательную обработку большой базы данных в одном компьютере. Требуем, чтобы вычислительная технология обеспечила оптимальное решение одной из задач ЛП, приведенных выше.

Введем следующие обозначения:

$J = \{1, \dots, N\}$ – индексы записей полной базы данных.

J_l – разбиение множества J – индексы l-ой порции данных; l=1,...,m; m – число выделенных порций при разбиении больших данных.

I_l – множество индексов наблюдений, которые активны при анализе l-ой порции данных.

M_l – информационное множество при анализе l-ой порции данных.

I_0 – индексы наблюдений, которые не формально выделены на начальном этапе вычислений из совокупности всех наблюдений и обеспечивают не пустоту и ограниченность соответствующего информационного множества.

Если из всей базы данных выделить таких наблюдений не удается, то совокупность данных не полна или противоречива и, следовательно, требуется ее корректировка. В частности, такая ситуация может возникнуть при $n>N$ или при не достаточном числе различных точек факторного пространства. Далее предполагаем, что такая ситуация при анализе данных не возникает и множество I_0 выделяет n линейно независимых ограничений решаемой задачи ЛП ($|I_0| = n$).

Алгоритм последовательной обработки большой базы запишем в следующем виде.

Шаг 0. Задаем разбиение множества J на подмножества $J_1, \dots, J_m$, выбираем задачу анализа данных и целевую функцию соответствующей задачи ЛП. Полагаем $\mu=0$.

Шаг 1. Формируем множество I_0 и полагаем l=1.

Шаг 2. Решаем задачу анализа для наблюдений с индексами $J_l \cup I_{l-1}$, которые определяют множество M_l. Если это множество пустое, то вычисления останавливаем с выдачей соответствующего сообщения.

Шаг 3. Выделяем индексы активных ограничений в задаче ЛП, решенной на шаге 2. Если $I_l \neq I_{l-1}$, то полагаем μ=1. Далее полагаем l=l+1. Если $l \leq m$ переходим к шагу 2.

Шаг 4. Если μ=*1*, полагаем $I_0 = I_k$; $l = 1$, идем на шаг 2. Иначе проводим анализ результатов решения ЗЛП и информационного множества базы данных, которое совпадает с множеством M_m.

Возможности реализации этой вычислительной технологии определяются допустимой для выбранного компьютера размерность задачи ЛП на шаге 2.

Рассмотрим алгоритм параллельной обработки больших данных, который является вариантом иерархических алгоритмов решения задач большой размерности, аналогичным алгоритмам в [11]. Задачей Центра является формирование множества I_0 и его последовательное уточнение до выполнения условия оптимальности решения выбранной ЗЛП для всей большой базы данных. Представление алгоритма проведем по аналогичной схеме, приведенной выше.

Шаг 0 (Задача Центра). Задаем разбиение множества J на подмножества $J_1, \dots, J_m$, выбираем задачу анализа данных и целевую функцию соответствующей задачи ЛП.

Шаг 1 (Задача Центра). Формируем множество I_0 и соответствующие строки матрицы наблюдений передаем в буфер обмена данными.

Шаг 2 (Задача блоков). Решаем в каждом компьютере l задачу анализа для наблюдений с индексами $J_l \cup I_0$, которые определяют множество M_l; $l = 1, \dots, m$. Если найдется компьютер, для которого это множество пустое, то вычисления останавливаем с выдачей соответствующего сообщения. Далее выделяем индексные множества I_l активных ограничений, и соответствующие строки матрицы наблюдений передаем в буфер обмена данными.

Шаг 3 (Задача Центра). Проводим сравнение индексных множеств I_0 и I_l; $l = 1, \dots, m$. Если найдется компьютер l для которого $I_l \neq I_0$, то с использованием всех активных наблюдений формируем новое множество I_0 и идем с ним к шагу 1. Иначе проводим анализ результатов обработки всей базы данных, состав которой представлен множеством I_0.

Конкретные реализации описанной вычислительной технологии определяются характером больших данных и программным обеспечением используемой вычислительной системы. Следует отметить, что на шаге 3 алгоритма Центр может уточнять состав активных ограничений решением ЗЛП с ограничениями, которые соответствуют индексным множествам I_0 и I_l; $l = 1, \dots, m$. В случае, когда ЗЛП окажется задачей большой размерности, Центр может воспользоваться первым алгоритмом последовательной обработки данных.

Кроме этого, следует подчеркнуть, что в рассмотренных вычислительных технологиях нет требований к количеству наблюдений в разбиении на порции большой базы данных.

ТЕМА 5.3. Компьютерное моделирование вычислительных технологий анализа больших данных

В данной теме рассматриваются вопросы реализации распределенных вычислений с учетом ограниченных возможностей выбранных программно-технических средств. Компьютерное моделирование процессов анализа больших данных проводится в среде Excel с использованием инструмента «Поиск решения». Предельные размеры базы данных по числу переменных модели определяются возможностями данного инструмента.

В нашем случае компьютерного моделирования учитываем, что для проверки оптимальности вычислений необходимо находить решение ЗЛП для всей базы данных, и ее размерность по числу переменных не должна быть более 200, а по числу главных ограничений – более 100.

Рассмотрим математические задачи вычислительных экспериментов для интервального анализа данных в общем случае наличия ошибок наблюдения всех переменных линейного процесса, которые касаются получения оценок объединенного множества решений ИСЛАУ. В данном разделе используются обозначения, принятые в литературе по теории ИСЛАУ и традиционные обозначения линейной алгебры и ЛП. Как описано в разделе 3 ИСЛАУ в матричной форме записывается интервальной ($N \times n$) матрицей коэффициентов и интервальным ($N \times 1$) вектором правой части в следующем виде:

$$\boldsymbol{A}x = \boldsymbol{B}\,. \tag{5.1}$$

Элементы матриц $\boldsymbol{A}$ и $\boldsymbol{B}$ являются интервальными оценками результатов измерения входных и выходной переменной в N наблюдениях и условно представляются неравенствами: $A^H \le \boldsymbol{A} \le A^V$; $B^H \le \boldsymbol{B} \le B^V$.

Значения вектора $x \in R^n$ в (5.1) соответствует оценками параметров линейной зависимости, а объединенное множество решений Ξ_{uni} соответствует множеству неопределенности описанному выше. В работе [6, с. 323] утверждается, что «вычисление для объединенного множества решений внешних покоординатных оценок с любой заданной абсолютной или относительной точностью есть *NP*-трудная задача». В частном случае положительных компонент решения ИСЛАУ объединенное множество решений задается системой линейных неравенств, которые запишем в следующем виде:

$$\Xi_{uni} = \{x \in R_+^n \colon A^V x \ge B^H;\ A^H x \le B^V\}. \tag{5.2}$$

Запишем ЗЛП для интервальной оценки значения выходной переменной b в заданной точке $a_p \in R^n$ факторного пространства:

$$b^H(a_p) = \min\nolimits_{x \in \Xi_{uni}} a_p x;\ \ b^V(a_p) = \max\nolimits_{x \in \Xi_{uni}} a_p x. \tag{5.3}$$

Таким образом, задачи ЛП в выбранном варианте прикладного интервального анализа данных имеют n переменных, а число ограничений равно удвоенному числу наблюдений анализируемой порции данных.

Рассматриваем компьютерную реализацию вычислительных технологий. В общем случае число m и размеры порций данных без учета условия целочисленности при допустимом n должны удовлетворять неравенству $2(n+N/m) < D$, где D – максимальное число ограничений задачи ЛП, используемого программного комплекса.

В нашем случае (D=100) для основного варианта экспериментальной базы выбрано n=20 и допустимые 50 наблюдений «разрезаны» на 5 порций. Заметим, что минимальное число порций в нашем случае может быть три, например, с числами наблюдений 17, 17, 16 соответственно. В вычислительных экспериментах исследовались и другие варианты размеров таблиц наблюдений, в том числе значительно больше (50x20), для различных по параметрам линейных процессов.

Во всех вариантах таблица наблюдений заполнялась в выбранных интервалах для входных переменных и для ошибок измерений равномерными псевдослучайными числами функцией Excel СЛЧИС(). Значения выходной переменной для заданных параметров линейной зависимости моделировалось с интервальной ошибкой. Для основного варианта базы данных интервалы для входных переменных были равны [5, 100], для их ошибок – [-1, 1], для ошибки наблюдения выходной переменной – [-2, 2]. Значение коэффициентов зависимости для всех переменных равнялось 10.

Вычислительные эксперименты проводились в одной книге Excel и схема программы для двух вычислительных технологий выбрана одинаковой: часть листов занимали генераторы базы данных, далее на отдельных листах в диалоге решалась одна из ЗЛП (10), задача Центра (шаг 1) и 5 задач блоков (шаг 2). Отличия алгоритмов последовательного и параллельного анализа порций данных состояло в разных схемах передачи формируемого и корректируемого Центром индексного множества *Io* для наблюдений, которые определяют матрицу базисных переменных ЗЛП.

Для основного варианта базы данных при ее неоднократных обновлениях показано, что предложенные вычислительные технологии позволяют получать оптимальное решение ЗЛП при анализе больших данных. При прочих равных условиях параллельный анализ является более предпочтительным по числу подзадач ЛП для получения точного решения. Число задач ЛП, решаемых на шаге 2 алгоритмов, существенно зависит от начального множества *Io* всей ЗЛП. Соответствующая матрица ограничений, кроме отсутствия линейно зависимых строк, должна быть хорошо обусловленной. В этом случае, показано, что в среднем за m решенных задач ЛП на шаге 2, можно получить оптимальное или близкое к нему значение искомой оценки. По опыту решения больших задач

ЛП [8, 9] на практике следует ожидать быстрое приближение (за 1, 2 прогона) в окрестность оптимальной оценки и медленное движение к ее точному значению.

При компьютерном моделировании рассмотрены особенности применения методов линейного программирования для выявления переменных, влияние которых на выходную переменную процесса отсутствует или не является существенным. Для решения этой задачи анализа данных проверялась гипотеза принадлежности нулевого значения исследуемого коэффициента его интервальной оценке решением задач ЛП (5.1). В данном случае можно воспользоваться решением задач нелинейного программирования (5.2) для выбранной совокупности переменных. Состав переменных для проверки их значимости получается методом центра неопределенности решением задачи (5.3), вариант которого в нашем случае сводится к двухкритериальной задаче нелинейного программирования.

Следует отметить, что эффективность решения проблемы анализа больших данных на практике может быть повышена за счет модификации схем распараллеливания в предложенных алгоритмах, применением дополнительных приемов организации вычислительного процесса и использования высокопроизводительных вычислительных систем.

ТЕМА 5.4. Обсуждение итоговых результатов и возможностей анализа больших данных

В данном разделе учебного пособия для решения проблемы анализа данных с большим числом наблюдений, превышающим возможности компьютера, предложено использовать интервальный подход, который при построении линейных по параметрам зависимостей сводится к решению ИСЛАУ. Среди множества решений ИСЛАУ, известных в литературе, выбрано объединенное множество решений, которое обеспечивает гарантированное оценивание искомых зависимостей и допускает в отдельных случаях применение методов линейного программирования. С использованием рассмотренного подхода при анализе больших данных получены интервальные оценки выходной переменной моделируемого процесса при заданных значениях входных переменных и интервальные оценки параметров исследуемой зависимости. Предложены два способа последовательного считывания ограничений и параллельных компьютерных вычислений при решении задач линейного программирования большой размерности. Оптимальность вычислений обоснована с использованием известного приема релаксации ограничений при решении задач оптимизации большой размерности. Проведено компьютерное моделирование вычислительных технологий с целью изучения возможности их использования на практике и для оценки погрешностей их приближенных реализаций. Исследование проведено на модельных процессах в условиях выполнимости предположений интервального анализа данных. Рассмотрены особенности применения методов линейного программирования для выявления переменных, влияние которых на выходную переменную процесса отсутствует или не является существенным.

Вопросы по разделу 5.

1. Рассмотреть метод релаксации ограничений для решения задач линейного программирования большой размерности.

2. Пояснить вычислительную технологию последовательного анализа большой базы данных и распределение вычислительных задач, решаемых Центром и Блоками.

3. Пояснить вычислительную технологию параллельного анализа большой базы данных и распределение вычислительных задач, решаемых Центром и Блоками.

4. Описать компьютерные модели для исследования методов и моделей прикладного интервального анализа больших данных.

Библиографический список к разделу 5.

1. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. – М.: Издательство Вильямс. 2017. – 480 с.

2. Оскорбин Н.М. Вычислительные технологии анализа больших данных методами линейного программирования // Высокопроизводительные вычислительные системы и технологии. 2021. Т. 5, № 1. С. 222-228.

3. Oskorbin N.M. Applied Interval Analysis of Big Data Using Linear Programming Methods // «Communications in Computer and Information Science» (CCIS), 2022, том 1526. Springer, Cham. https://doi.org/10.1007/978-3-030-94141-3_16.
4. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сиб. мат. журнал. 1962. Т. 3. №5. – С. 701-709.
5. Shary S.P. A New Technique in Systems Analysis under Interval Uncertainty and Ambiguity. Reliable Computing. 2002. Vol. 8, No. 5. pp. 321–418. DOI: 10.1023/A:1020505620702.
6. Шарый С.П. Конечномерный интервальный анализ. – Новосибирск: Изд-во «XYZ». 2017. – 618 с.
7. Шелудько А.С. Гарантированное оценивание параметров дискретных моделей хаотических процессов // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2018. Т. 7, № 1. С. 25-39. DOI: 10.14529/cmse180103.
8. Лэсдон Л.С. Оптимизация больших систем. – М.: Наука, 1975. – 432 с.
9. Цурков В.И. Декомпозиция в задачах большой размерности. – М.: Наука, 1981. – 352 с.
10. Geoffrion A. Reducing Concave Programs with Some Linear Constraints, SIAM J. Appl. Math., 15, 1967, pp. 653-664.
11. Oskorbin N.M. Computational technologies for the synthesis of decentralized control systems for multistage technological processes // Journal of Physics: Conference Series. 2020. Vol. 1615. Article ID 012020. DOI:10.1088/1742-6596/1615/1/012020.

ПРИЛОЖЕНИЯ

Приложение 1 Пример компьютерной модели анализа данных с использованием ИСЛАУ

Лист Excel программы генерации таблицы наблюдений без ошибок

Генерирование таблицы наблюдения без ошибок

Модельный процесс: в=a1*x1+a2*x2

Низ 0 <=a1<=100; 0 <=a2<=100

a1=	10	100
a2=	10	100

x1	x2
1	1

Таблица 1.1

a1 & a2 без ошибок

№ п/п	a1	a2
1	44,92	36,02
2	34,91	70,21
3	90,96	84,52
4	18,22	36,57
5	91,97	85,15
6	13,91	14,58
7	63,48	83,28
8	77,83	32,55
9	87,34	44,50
10	62,85	12,12
11	87,67	19,27
12	65,82	16,78
13	68,31	39,05
14	16,30	99,19
15	19,75	40,05
16	85,39	49,65
17	53,81	62,39
18	59,58	17,48
19	11,27	22,24
20	13,09	21,91

КОПИР ТАБ 1.2

Таблица 1.2

Значения переменных без ошибок

№ п/п	a1	a2	b
1	48,37	86,73	135,10
2	36,48	92,47	128,95
3	87,19	90,78	177,97
4	73,73	95,55	169,28
5	99,94	37,79	137,73
6	99,72	74,62	174,34
7	78,73	35,99	114,72
8	37,93	65,95	103,88
9	93,58	70,95	164,54
10	94,01	47,90	141,91
11	93,29	75,28	168,56
12	44,90	60,81	105,71
13	67,71	46,35	114,07
14	52,76	40,90	93,67
15	58,31	14,67	72,98
16	43,35	68,09	111,44
17	27,35	85,06	112,41
18	22,51	49,13	71,64
19	11,13	93,30	104,43
20	76,70	99,59	176,29

Данные Без_Ошиб | Генерация_Ошибок | Данне_С_Ошиб | ОМР-ИСХ | ДМР-ИСХ

Приложение 1 (продолжение)

Лист Excel программы генерации ошибок наблюдения

Генерирование псевдослучайных ошибок

Ea1	Ea2	Eb
1	1	4

Генератор ошибок — Таблица 2.1

№ п	Ea1	Ea2	Eb
1	0,904	0,550	3,651
2	0,904	-0,762	-2,255
3	0,171	0,806	-0,030
4	0,963	0,719	1,182
5	-0,509	0,814	1,092
6	-0,414	-0,894	3,430
7	0,368	0,076	-1,548
8	-0,614	-0,349	0,037
9	0,896	0,457	0,625
10	0,900	0,160	-0,557
11	0,862	0,279	-1,872
12	0,349	-0,747	3,431
13	0,411	-0,692	3,113
14	-0,198	-0,429	3,723
15	-0,876	-0,149	0,764
16	-0,391	-0,946	3,098
17	-0,055	0,085	-2,074
18	-0,375	0,347	2,839
19	0,998	-0,594	-0,035
20	0,819	-0,862	2,729

Спец копия значений таб 2.1 — Таблица 2.2

№ п	Ea1	Ea2	Eb
1	-0,190	-0,652	-1,455
2	-0,194	0,708	1,888
3	-0,232	-0,857	3,995
4	-0,933	-0,281	-1,766
5	-0,030	-0,059	0,227
6	0,550	-0,662	-3,032
7	-0,148	0,978	-2,991
8	0,915	-0,950	2,481
9	0,131	-0,522	3,441
10	-0,243	0,924	-0,981
11	0,053	0,928	-1,016
12	0,617	-0,517	0,455
13	-0,542	-0,150	-1,356
14	0,489	-0,225	3,046
15	-0,525	0,379	1,146
16	-0,968	-0,208	-0,293
17	0,403	0,432	3,736
18	-0,128	-0,870	-3,702
19	-0,083	-0,204	2,802
20	-0,042	0,434	3,411

Приложение 1 (продолжение)

Лист Excel генерация таблицы наблюдений моделируемого процесса

Таблица наблюдения с ошибками

Таблица 3.1

Итог наблюдений с ошибками

№ п	a1	a2	b
1	48,183	86,076	133,646
2	36,283	93,180	130,838
3	86,960	89,925	181,968
4	72,795	95,273	167,517
5	99,910	37,726	137,953
6	100,271	73,956	171,307
7	78,582	36,963	111,724
8	38,847	65,000	106,363
9	93,716	70,431	167,978
10	93,768	48,827	140,932
11	93,339	76,206	167,548
12	45,517	60,294	106,166
13	67,170	46,203	112,709
14	53,254	40,676	96,712
15	57,783	15,049	74,124
16	42,387	67,879	111,149
17	27,751	85,495	116,147
18	22,383	48,262	67,941
19	11,049	93,098	107,236
20	76,658	100,026	179,704

Данные Без_Ошиб | Генерация_Ошибок | Данне_С_Ошиб | ОМР-ИСХ

Приложение 1 (продолжение)

Лист Excel: Пример программы оценки параметров на примере ОМР

Объединенное множество решений ИСЛАУ

$$\begin{cases} A^{Г} x \geq B^{Н} \\ A^{Н} x \leq B^{Г} \end{cases}$$

Оцеки ошибок измерения

Ea1	Ea2	Eb
1	1	4

Таб 4.5 Av*X An*X — Таблица 4.4 B верх, B низ

№ п/п	Av*X	An*X	№ п/п	B верх	B низ
1	137,36	133,38	1	137,65	129,65
2	133,55	129,58	2	134,84	126,84
3	177,97	173,99	3	185,97	177,97
4	170,19	166,21	4	171,52	163,52
5	135,66	131,68	5	141,95	133,95
6	173,85	169,87	6	175,31	167,31
7	114,74	110,76	7	115,72	107,72
8	106,54	102,56	8	110,36	102,36
9	163,98	160,00	9	171,98	163,98
10	141,46	137,48	10	144,93	136,93
11	169,66	165,68	11	171,55	163,55
12	107,92	103,94	12	110,17	102,17
13	113,63	109,65	13	116,71	108,71
14	94,72	90,75	14	100,71	92,71
15	72,23	68,25	15	78,12	70,12
16	112,88	108,91	16	115,15	107,15
17	117,48	113,50	17	120,15	112,15
18	73,52	69,54	18	71,94	63,94
19	109,66	105,68	19	111,24	103,24
20	178,80	174,82	20	183,70	175,70

Таблица 4.5 Av — Таблица 4.4 An

№ п/п	A1V	A2V	№ п/п	A1h	A2h
1	49,18	87,08	1	47,18	85,08
2	37,28	94,18	2	35,28	92,18
3	87,96	90,93	3	85,96	88,93
4	73,80	96,27	4	71,80	94,27
5	100,91	38,73	5	98,91	36,73
6	101,27	74,96	6	99,27	72,96
7	79,58	37,96	7	77,58	35,96
8	39,85	66,00	8	37,85	64,00
9	94,72	71,43	9	92,72	69,43
10	94,77	49,83	10	92,77	47,83
11	94,34	77,21	11	92,34	75,21
12	46,52	61,29	12	44,52	59,29
13	68,17	47,20	13	66,17	45,20
14	54,25	41,68	14	52,25	39,68
15	58,78	16,05	15	56,78	14,05
16	43,39	68,88	16	41,39	66,88
17	28,75	86,49	17	26,75	84,49
18	23,38	49,26	18	21,38	47,26
19	12,05	94,10	19	10,05	92,10
20	77,66	101,03	20	75,66	99,03

Оценки	X1	X2	Прогноз в точке	B1	B2	
	0,94	1,04		55	55	110,00
Интервальные оценки решений ИСЛАУ					109,34	
MAX	1,02	1,02		MAX	112,84	
МИН	0,99	0,99		МИН	109,34	
Среднее	1,00	1,00		Среднее	111,09	
Дов инт	0,01	0,02		Дов инт	1,59%	

Приложение 1 (продолжение)

Лист Excel: Пример программы оценки параметров на примере ДМР

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Допусковое множество решений ИСЛАУ											AvX<=Bv	
2	Ea1	Ea2	Eb										AhX>=Bh	
3	1	1	4											
4	Таб 4.5 Av*X An*X				Таблица 4.4 В верх, В низ				Таблица 4.5 Av			Таблица 4.4 An		
5	№ п/п	Av*X	An*X		№ п/п	В верх	В низ		№ п/п	A1V	A2V	№ п/п	A1h	A2h
6	1	138,04	134,01		1	139,42	127,87		1	49,18	87,08	1	47,18	85,08
7	2	133,61	129,58		2	136,61	125,07		2	37,28	94,18	2	35,28	92,18
8	3	180,22	176,20		3	187,74	176,20		3	87,96	90,93	3	85,96	88,93
9	4	171,76	167,73		4	173,29	161,74		4	73,80	96,27	4	71,80	94,27
10	5	139,31	135,28		5	143,73	132,18		5	100,91	38,73	5	98,91	36,73
11	6	176,92	172,89		6	177,08	165,53		6	101,27	74,96	6	99,27	72,96
12	7	117,50	113,47		7	117,50	105,95		7	79,58	37,96	7	77,58	35,96
13	8	107,16	103,13		8	112,14	100,59		8	39,85	66,00	8	37,85	64,00
14	9	166,83	162,81		9	173,75	162,21		9	94,72	71,43	9	92,72	69,43
15	10	144,67	140,64		10	146,70	135,16		10	94,77	49,83	10	92,77	47,83
16	11	172,40	168,37		11	173,32	161,78		11	94,34	77,21	11	92,34	75,21
17	12	108,89	104,87		12	111,94	100,39		12	46,52	61,29	12	44,52	59,29
18	13	115,75	111,72		13	118,48	106,94		13	68,17	47,20	13	66,17	45,20
19	14	96,35	92,32		14	102,48	90,94		14	54,25	41,68	14	52,25	39,68
20	15	74,46	70,43		15	79,90	68,35		15	58,78	16,05	15	56,78	14,05
21	16	113,61	109,58		16	116,92	105,38		16	43,39	68,88	16	41,39	66,88
22	17	117,30	113,27		17	121,92	110,37		17	28,75	86,49	17	26,75	84,49
23	18	73,71	69,69		18	73,71	62,17		18	23,38	49,26	18	21,38	47,26
24	19	108,65	104,62		19	113,01	101,46		19	12,05	94,10	19	10,05	92,10
25	20	180,46	176,43		20	185,48	173,93		20	77,66	101,03	20	75,66	99,03
26														
27	Красш	X1	X2			Прогноз в точке		B1	B2		Принадлежность точки			
28	1,44	0,99	1,03					55	55		допусковому множеству			
29	Интервальные оценки решений ИСЛАУ								110,78		E	E*E		
30	MAX	2,05	1,01					MAX	120,82		-0,01	0,00		
31	МИН	1,99	1,00					МИН	120,14		-0,03	0,00		
32	Среднее	2,02	1,00					Среднее	120,48		-0,04	0,00		
33	Дов инт	0,03	0,00					Дов инт	0,34		90000	90,76		

Приложение 2. ***Прикладной интервальный анализ в иллюстрациях***

Список иллюстраций

Иллюстрация 1. Задачи анализа данных

На данной иллюстрации перечислены итоговые задачи анализа данных: оценки параметров моделируемого процесса и прогнозы значений выходной переменной. База данных это таблица наблюдений (как то относящихся к процессу), а база знаний, все что известно о наблюдениях и о свойствах процесса для анализа данных. Таким образом, анализ данных это 2 начальные стадии триады: «Знать – Предвидеть – Управлять».

ПРИКЛАДНЫЕ ЗАДАЧИ АНАЛИЗА ДАННЫХ

Кибернетическая модель процессов «Черный ящик»

$$\xrightarrow{x \in R^n} \boxed{\;y = f(a,x)\;} \xrightarrow{y \in R}$$

Информация о процессе: база данных и база знаний.

База данных: $\{x_j, y_j\}, j = 1, \dots, N; N$ **– число наблюдений.**

ЗАДАЧА 1. **Точечные и интервальные оценки параметра $a \in R^m$ (Задача оценки).**

ЗАДАЧА 2. **Точечная и интервальная оценка $y(x_0)$ в точке $x_0 \in R^n$ (Задача прогноза).**

Иллюстрация 2. История анализа данных

МНК. Гауссу (1795) принадлежит первое применение метода, а Лежандр (1805) опубликовал его под современным названием (фр. Méthode des moindres quarrés). Лаплас связал метод с теорией вероятностей, а американский математик Эдрейн (1808) рассмотрел его теоретико-вероятностные приложения.

Математическая статистика. Теория оценивания. Эффективность. Состоятельность. Несмещенность. Неравенство Рао-Крамера. Неравенство Крамера–Рао – неравенство, которое при некоторых условиях на статистическую модель даёт нижнюю границу для дисперсии оценки неизвестного параметра.

ПИА. Все согласны с приоритетом Канторовича Леонида Витальевича, математика, Нобелевского лауреата по экономике.

Нейронные сети, машинное обучение. С 1943 года разработка первой нейронной сети головного мозга. Далее изобретение перцетрона, идея которого развита до современных моделей нейронных сетей. Вклад Горбаня А.Н.

Один из результатов Горбаня А.Н. любую непрерывную функцию в *n*-мерном пространстве можно представить точно нейронной сетью.

Александр Николаевич Горбань (род. 1952, Омск), д.ф.-м.н, профессор, работает в областях фундаментальной и прикладной науки, включая статистическую физику, неравновесную термодинамику, теорию машинного обучения и математическую биологию. В настоящее время живет в Великобритании, но остается сотрудником ИВМ СО РАН (Красноярский научный центр.)

Научные и прикладные исследования нейронных сетей, включая гибридные, проводятся в Алтайском техническом университете под научным руководством Пятковского Олега Ивановича, д.т.н., профессора.

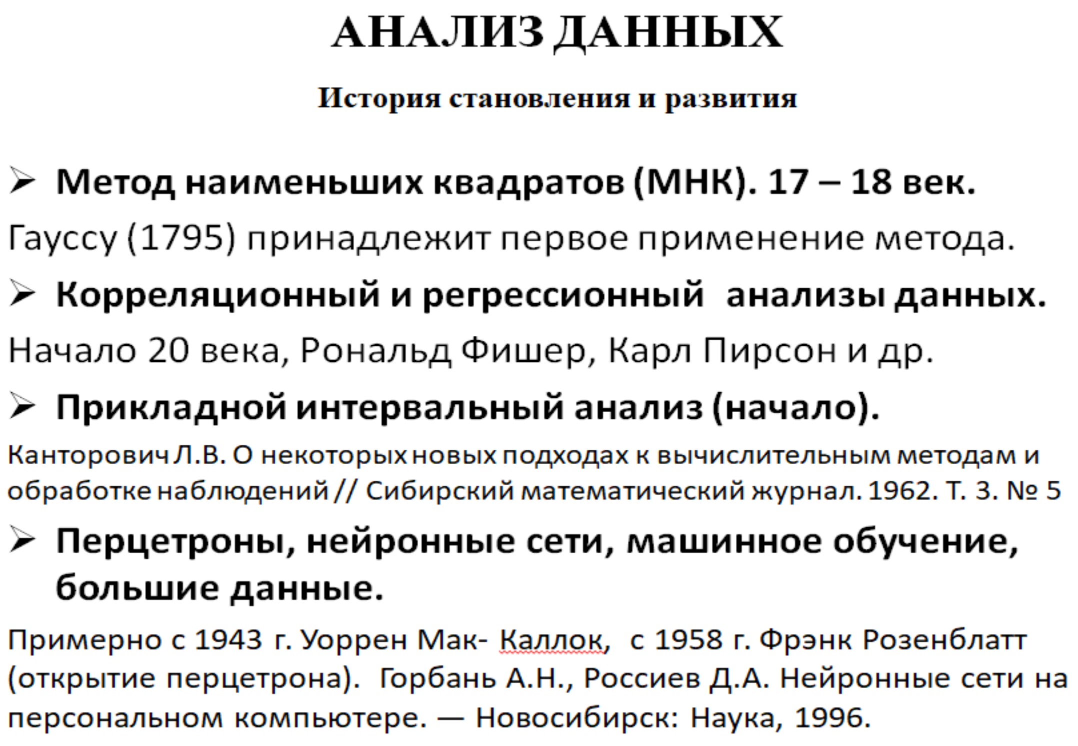


Иллюстрация 3. Прикладной интервальный анализ (разделы)

На этой иллюстрации разделы прикладного интервального анализа данных (ПИА) в соответствии с современной классификацией. В ПИА рассматривается анализа данных при идентификации по экспериментальным данным параметров линейных детерминированных процессов. Математическая модель и теоретические основы анализа данных базируются на методах интервальных систем линейных алгебраических уравнений (ИСЛАУ).

Предполагается, что моделируемый процесс описывается выходной переменной и n входных переменных, измерения которых представлены в базе данных интервальными величинами, а структура модели и границы интервалов ошибок измерения всех переменных являются достоверными.

Следует отметить, что исходная идея использования интервальной математики и линейного программирования (ЛП) для анализа данных высказана Л. Канторовичем в 1962 г.

Теоретические исследования в области прикладного интервального анализа данных представлены в литературе. В настоящее время прикладной интервальный анализ экспериментальных данных не встречает существенной критики. Однако его применение на практике сопряжено с проблемами, главные из которых, во-первых, высокая погрешность гарантированных оценок, и, во-вторых, высокие требования к исходным предположениям.

ПРИКЛАДНОЙ ИНТЕРВАЛЬНЫЙ АНАЛИЗ

РАЗДЕЛЫ СЛАЙДОВ

- **КЛАССИЧЕСКИЙ РАЗДЕЛ ПИА (Канторович Л.В.).**

  Входные переменные процесса измеряется без ошибок, а выходная переменная – с известной интервальной ошибкой.

- **СОВРЕМЕННЫЙ РАЗДЕЛ ПИА (Шарый С.П., Жолен Л. и др.).**

  Допускает наличие интервальных ошибок измерения всех переменных моделируемого линейного процесса.

- **ЗАДАЧИ И МЕТОДЫ СОГЛАСОВАНИЯ БАЗЫ ДАННЫХ.**

  На примерах классического и современного разделов ПИА.

Иллюстрация 4. История анализа данных

Рассматривается классика ПИА, когда в измерениях отсутствуют ошибки измерения входных переменных при формировании базы данных.

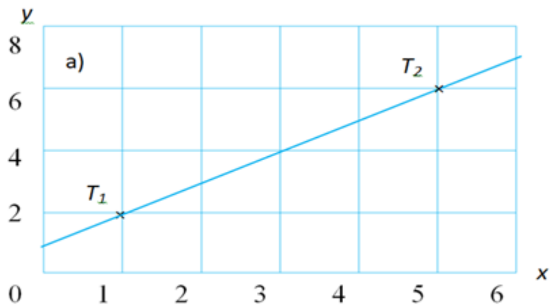
В простых случаях двух наблюдений все элементарно. Рисунок а) это как построить уравнение прямой по 2 заданным точкам. Справа от рисунка заданы условия информационного множества *A(2)*. Строка выше по слайду описывает моделируемый процесс. Здесь и по всему докладу приводятся примеры компьютерного моделирования процессов, свойства которых точно известны.

Смотрим рисунок б). Средняя прямая – моделируемый процесс, крайние выше и ниже линии – границы оценок прямых. Еще 2 границы не показаны, они проходят по противоположным крайним точкам.

Обращаем внимание на условия 4.2. Число строк может быть любым, и даже большим. Чисто столбцов, т.е. размерность пространства n обычно не более нескольких десятков. Анализ проводим методами с использованием ЛП.

КЛАССИЧЕСКИЙ РАЗДЕЛ ПИА

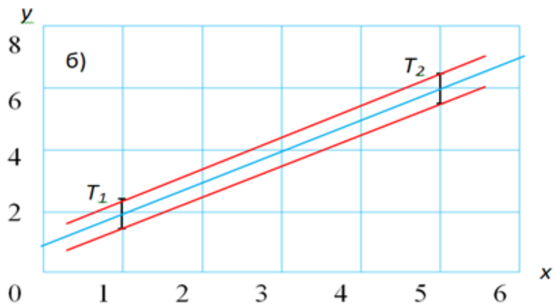
$\boldsymbol{n = 1;\ N = 2;\ m = 2;\quad y = a_0 + a_1 {*} x;\quad a_0 = a_1 = 1}$



а) База данных: $\boldsymbol{T_1}$: $(x_1, y_1) = (1, 2)$;
$\boldsymbol{T_2}$: $(x_2, y_2) = (5, 6)$.

Информационное множество A(N)

$$\left.\begin{aligned} 2 \le a_0 + a_1 \le 2 \\ 6 \le a_0 + 5a_1 \le 6 \end{aligned}\right\} \qquad (4.1)$$



б) Анализ данных с ошибкой измерения выходной переменной.

Оценка ошибок $|\varepsilon_y| \le 0.5$

$$\left.\begin{aligned} 1{,}5 \le a_0 + a_1 \le 2{,}5 \\ 5{,}5 \le a_0 + 5a_1 \le 6{,}5 \end{aligned}\right\} \qquad (4.2)$$

Иллюстрация 5. КЛАССИКА ПИА: оценка параметров

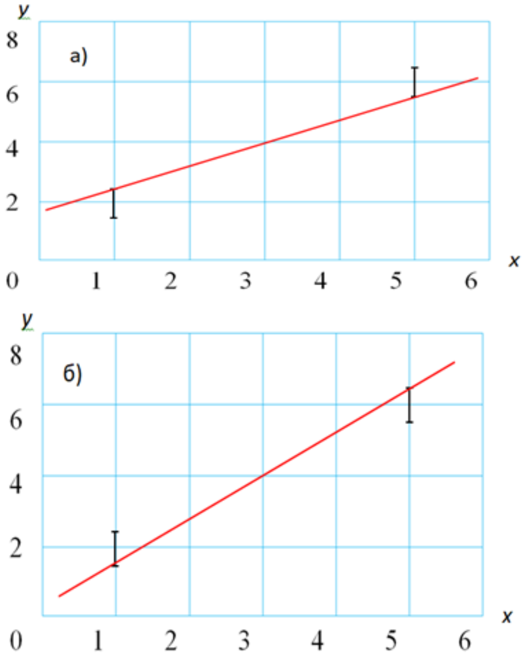
Пример решения задачи 1 – оценки параметров процесса. Тот же процесс, та же база данных (4.2). На иллюстрации 5 приведены точечные и интервальные оценки параметров a_0 и a_1 вычисленные в Excel.

Задача оценки значений коэффициентов искомой зависимости. Оценки получаем решением $2n$ задач ЛП. Например, коэффициент x_1 принадлежит интервалу $[x_1^H, x_1^V]$: $x_1^H = min_{x \in X}\, x_1$; $x_1^V = max_{x \in X}\, x_1$.

В общем случае оценки параметров получаем решение задач линейного программирования в среде Excel.

КЛАССИКА ПИА: ОЦЕНКА ПАРАМЕТРОВ

$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 * x;\ a_0 = a_1 = 1}$



База данных: T_1: $(x_1, y_1) = (1, 2)$;
T_2: $(x_2, y_2) = (5, 6)$.

Информационное множество A(2)

$$\left.\begin{array}{l} 1{,}5 \le a_0 + a_1 \le 2{,}5 \\ 5{,}5 \le a_0 + 5a_1 \le 6{,}5 \end{array}\right\} \quad (4.2)$$

Решаем 2m задач ЛП. Оценки a_0, a_1

а) $a_0^V = \max_{a \in A(N)} a_0$ (5.1)

б) $a_0^H = \min_{a \in A(N)} a_0$ (5.2)

а) $a_1^V = \max_{a \in A(N)} a_1$ (5.3)

б) $a_1^H = \min_{a \in A(N)} a_1$ (5.4)

Оценки: $a_0 \in [0.25; 1.75], a_1 \in [0.75; 1.25]$

Точечная: $\hat{a}_0 = 0.5(a_0^V + a_0^H) = 1, \hat{a}_1 = 1$

ДОВ: $Da_0 = 0.5(a_0^V - a_0^H) = 0.75, Da_1 = 0.25$

Иллюстрация 6. Классика ПИА: прогноз выхода

Иллюстрация 6 показывает получение точечных и интервальных оценок прогноза выходной переменной в 2 точках: x_0=3 – рисунок а) *и* x_0=6 рисунок б).

В общем случае задача прогноза выходной переменной моделируемого процесса в заданной точке факторного пространства (пространства входных переменных) $a^P = (a_1^P, \dots, a_n^P)$. Интервальную оценку $[b^H(a^P),\ b^V(a^P)]$ получаем решением двух задач ЛП:

$b^H(a^P) = \min_{x \in X}(x_1 a_1^P + \dots + x_n a_n^P);\ b^V(a^P) = max_{x \in X}(x_1 a_1^P + \dots + x_n a_n^P).$

Иллюстрация 7. Классика ПИА: расчеты в Excel

На слайде лист Excel, где запрограммированы формулы и данные для всех расчетов иллюстраций 5 и 6. В программе используется инструмент «Поиск решения» для любых N<100 и любым n<200. Эти ограничения следуют для Excel версии 2010.

Обозначены: Задача 1 – интервальные оценки коэффициентов линейной зависимости; Задача 2, вариант 1 – интервальная оценка прогноза выходной переменной в точке $x_1 = 3$; вариант 2 – интервальная оценка прогноза выходной переменной в точке $x_1 = 6$.

На слайде приведены расчетные средние значения и полуинтервалы искомых оценок.

КЛАССИКА ПИА: РАСЧЕТЫ В EXCEL

$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1{*}x;\ a_0{=}a_1{=}1}$

	A	B	C	D	E	F
1		Ex=	0	Ey=	0,5	
2	№	X	У	Ур	Ун	Ув
3	1	1	2	2,5	1,5	2,5
4	2	5	6	5,5	5,5	6,5
5						
6	ЗАДАЧА 1				ЗАДАЧА 2 (2 вар)	
7	A0	A1			x1=	x2=
8	1,75	0,75			3	6
9					4	6,25
10		A0	A1		У1	У2
11	Max	1,75	1,25	Max	4,5	7,75
12	Мин	0,25	0,75	Мин	3,5	6,25
13	СР	1,00	1,00	СР	4,00	7,00
14	ДОВ	0,75	0,25	ДОВ	0,50	0,75

Иллюстрация 8. Классика ПИА: согласование данных

Рассматриваем случай, когда при решении задач моделирования (оценки параметров или прогноза значения выходной переменной) информационное множество оказалось пустым, т.е. система неравенств, аналогичная (4.2) – противоречива. В данном случае условия теоретического моделирования процессов не выполнены, но мы не знаем, какие из условий оказываются нарушенными. Допустим, что мы убеждены в достоверности базы знаний. Как проверить, что база данных не достоверна? Предположим, что одно из N наблюдений с неизвестным номером J является выбросом. Как найти этот номер?

Перечисленные задачи решаются при нахождении точечной оценки параметров линейных моделей методом центра неопределенности (МЦН).

В ряде случаев требуется проверить выполнимость исходных предположений (например, в случае, когда A(N) – пустое множество) или получить точечные оценки параметров модели.

Рассматриваемый раздел прикладного интервального анализа назван как согласование базы данных, а в некоторых случаях и базы знаний.

ПРИКЛАДНОЙ ИНТЕРВАЛЬНЫЙ АНАЛИЗ

РАЗДЕЛЫ СЛАЙДОВ НИЖЕ

- КЛАССИЧЕСКИЙ РАЗДЕЛ ПИА (ПО Канторовичу Л.В.).
  Входные переменные процесса измеряется без ошибок, а выходная переменная – с известной интервальной ошибкой.
- **ЗАДАЧИ И МЕТОДЫ СОГЛАСОВАНИЯ БАЗЫ ДАННЫХ.**
  На примерах классического раздела ПИА.

Иллюстрация 9. Классика ПИА: реальная база данных

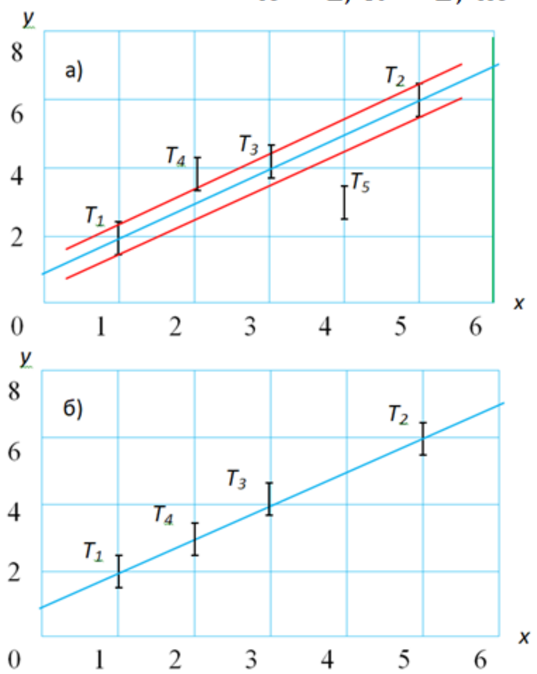
На этой иллюстрации приведен вариант реальной базы данных. К данным ранее добавлены 3 наблюдения, координаты и доверительные интервалы ошибок приведены на рисунке а).

Методы ПИА без визуализации данных должны обнаружить не согласованность данных и выполнить согласование базы: либо выбросить точки, либо исправить.

Согласованная база данных приведена на рисунке б) слайда 7. Все вычисления в Excel на слайдах 10 – 14.

КЛАССИКА ПИА: РЕАЛЬНАЯ БАЗА ДАННЫХ

$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 {*} x;\ a_0 = a_1 = 1}$



База данных: T_1: $(x_1, y_1) = (1, 2)$;
T_2: $(x_2, y_2) = (5, 6)$;
T_3: $(x_3, y_3) = (3, 4.2)$;
T_4: $(x_4, y_4) = (2, 3.5)$
T_5: $(x_5, y_5) = (4, 3)$.

Наблюдения T_1, T_2, T_3 **согласованы.**
Точка T_5 **- явный выброс.**
Точка T_4 **внешне согласована, но оценка ее информационной ценности явно завышена. Корректируется по прогнозу:** T_4: $(x_4, y_4) = (2, 3.05)$

Иллюстрация 10. Классика ПИА: этапы коррекции-1

На этом слайде методом уширения раздвигаются равномерно вверх и вниз все границы по переменной *y* пока не выполнится условие *A(5)* не пусто. Здесь *Кр* коэффициент расширения.

Для согласованной базы *Кр*=1. Параметр *W* в столбце *G* исключает наблюдение для нас при *W*=10. Выбросы находим по решению двойственной ЗЛП. Вычисления проводим пока *Кр* не больше единицы.

КЛАССИКА ПИА: ЭТАПЫ КОРРЕКЦИИ-1

$$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 * x;\ a_0 = a_1 = 1}$$

	A	B	C	D	E	F	G
1		Ex=	0	Ey=	0,5		
2	№	X	У	Ур	Ун	Ув	W
3	T1	1	2	2,167	0,833	3,167	1
4	T3	3	4,2	3,500	3,033	5,367	1
5	T4	2	4,0	2,833	2,833	5,167	1
6	T5	4	3	4,167	1,833	4,167	1
7	T2	5	6	4,833	4,833	7,167	1
8							
9	Метод уширения						
10	Кр	A0	A1				
11	2,333	1,500	0,667				

Иллюстрация 10. Классика ПИА: этапы коррекции-2

На слайде 10 уже нашли и исключили выброс точку 5. При поиске можно кроме визуального анализа использовать данные отчета инструмента Excel «Поиск решения». Выбросам соответствуют не нулевые значения двойственных переменных. Исключение из базы данных для проверки выброса или для его окончательного удаления используется параметр W в столбце G.

Значение коэффициента уширения Кр меньше или равное единице является критерием отсутствия аномальных наблюдений в базе данных. Этот критерий верен только в случае, когда база знаний сформирована правильно.

КЛАССИКА ПИА: ЭТАПЫ КОРРЕКЦИИ-2

$$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 * x;\ a_0 = a_1 = 1}$$

	A	B	C	D	E	F	G
1		Ex=	0	Ey=	0,5		
2	№	X	У	Ур	Ун	Ув	W
3	T1	1	2	2,500	1,500	2,500	1
4	T3	3	4,2	4,500	3,700	4,700	1
5	T4	2	4,0	3,500	3,500	4,500	1
6	T5	4	3	5,500	-2,000	8,000	10
7	T2	5	6	6,500	5,500	6,500	1
8							
9	Метод уширения						
10	Кр	A0	A1				
11	1,000	1,500	1,000				

Иллюстрация 11. Классика ПИА: этапы коррекции-3

На сладе 11 решены задачи 1 и 2 без точки 5 и обнаружено, что ценность базы данных недопустимо большая. Кроме того, не выполнено условие гарантированного оценивания (невязка – минимальное расстояние от истинной точки коэффициентов линейной зависимости до допускового множества решений – равно 0,25).

Нулевые доверительные интервалы не реальны. Приводим пример. Вы едете в Москву искать друга. Вы ходите из аэропорта и вот он. Это случайно? Нет. В жизни так не бывает, если специально не подстроено.

Напомним, опыт Майкельсона-Морли повторялся не менее десятков раз. Не верили в ошибочность теории эфира. Но как найти противоречие и поправить базу данных в нашем случае?

КЛАССИКА ПИА: ЭТАПЫ КОРРЕКЦИИ-3

$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 {*} x;\ a_0 = a_1 = 1}$

	A	B	C	D	E	F	G
1		Ex=	0	Ey=	0,5		
2	№	X	У	Ур	Ун	Ув	W
3	T1	1	2	2,500	1,500	2,500	1
4	T3	3	4,2	4,500	3,700	4,700	1
5	T4	2	4,0	3,500	3,500	4,500	1
6	T5	4	3	5,500	-2,000	8,000	10
7	T2	5	6	6,500	5,500	6,500	1
8							
9	Метод уширения				Гарантированное оценивание		
10	Кр	A0	A1		A0ис	A1ис	Невяз
11	1,000	1,500	1,000		1	1	0,25
12					3	6	
13		A0	A1		4,5	7,5	
14	Max	1,5	1	Max	4,5	7,5	
15	Мин	1,5	1	Мин	4,5	7,5	
16	СР	1,50	1,00	СР	4,50	7,50	
17	ДОВ	0,00	0,00	ДОВ	0,00	0,00	

Иллюстрация 12. Классика ПИА: этапы коррекции-4

На этом слайде нашли по двойственным оценкам точку Т4 и проверяем ее на как подозрительную на наличие выброса. В строке 5 столбца G полагаем значение индикатора W равное 10 (достаточно для исключения точки Т4 из базы данных). С новой укороченной базой данных видим, что условие гарантированного оценивания выполнено. Значение клетки G11 нулевое.

Решаем задачу прогноза среднего значения выходной переменной в точке Т4 с укороченной базой данных. В клетке F16 это среднее равно 3,05. Далее необходимо принять решение об исключении точки Т4 из базы данных или скорректировать эту точку по прогнозному значению. Второе решение предпочтительно для случая, когда база данных будет пополняться в последующих задачах анализа данных. Это обеспечит ее устойчивость.

КЛАССИКА ПИА: ЭТАПЫ КОРРЕКЦИИ-4

$$n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 {*} x;\ a_0 = a_1 = 1$$

	A	B	C	D	E	F	G
1		Ex=	0	Ey=	0,5		
2	№	X	У	Ур	Ун	Ув	W
3	Т1	1	2	2,000	1,500	2,500	1
4	Т3	3	4,2	4,000	3,700	4,700	1
5	Т4	2	4,0	3,000	-1,000	9,000	10
6	Т5	4	3	5,000	-2,000	8,000	10
7	Т2	5	6	6,000	5,500	6,500	1
8							
9	Метод уширения				Гарантированное оценивание		
10	Кр	А0	А1		А0ис	А1ис	Невяз
11	1,000	1,000	1,000		1	1	1,81E-14
12					3	2	
13		А0	А1		4	3	
14	Max	1,75	1,25	Max	4,5	3,5	
15	Мин	0,25	0,75	Мин	3,7	2,6	
16	СР	1,00	1,00	СР	4,10	3,05	
17	ДОВ	0,75	0,25	ДОВ	0,40	0,45	

Иллюстрация 13. Классика ПИА: этапы коррекции-5

На этом слайде все исправлено и согласовано. По 4 точкам оценки параметров линейной зависимости стали гарантированными и лучше (по прогнозу при значении входной переменной в точке $x = 3$) в сравнении с аналогичной оценкой по базе данных с двумя наблюдениями.

КЛАССИКА ПИА: ЭТАПЫ КОРРЕКЦИИ-5

$$n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 * x;\ a_0 = a_1 = 1$$

	A	B	C	D	E	F	G
1		Ex=	0	Ey=	0,5		
2	№	X	У	Ур	Ун	Ув	W
3	T1	1	2	2,000	1,500	2,500	1
4	T3	3	4,2	4,000	3,700	4,700	1
5	T4	2	3,05	3,000	2,550	3,550	1
6	T5	4	3	5,000	-2,000	8,000	10
7	T2	5	6	6,000	5,500	6,500	1
8							
9	Метод уширения				Гарантированное оценивание		
10	Кр	A0	A1		A0ис	A1ис	Невяз
11	1,000	1,000	1,000		1	1	7,19E-13
12					3	2	
13		A0	A1		4,0	3,0	
14	Max	1,75	1,25	Max	4,5	3,5	
15	Мин	0,25	0,75	Мин	3,7	2,6	
16	СР	1,00	1,00	СР	4,10	3,05	
17	ДОВ	0,75	0,25	ДОВ	0,40	0,45	
18		Оценки по базе данных с двумя наблюдениями					
19		A0	A1		У1	У2	
20	Max	1,75	1,25	Max	4,5	7,75	
21	Мин	0,25	0,75	Мин	3,5	6,25	
22	СР	1,00	1,00	СР	4,00	7,00	
23	ДОВ	0,75	0,25	ДОВ	0,50	0,75	

Иллюстрация 15. ПИА: ошибки измерения всех переменных

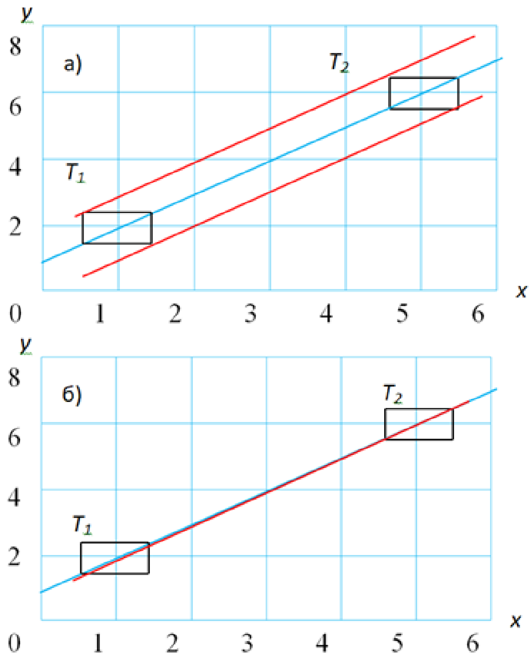
Возвращаемся к базе данных с 2 наблюдениями с тем же линейным процессом, но с ошибками измерения всех переменных.

Реально они принимают любые значения в заданных интервалах, но неизвестно какие. В современном прикладном анализе в известной литературе применяют 3 определения информационных множеств для всех возможных искомых прямых. Это объединенное множество, допусковое множество и управляемое множество. Условие первого (15.1) на рисунке а). Второго (15.2) на рисунке б). Пояснить отличие условий.

Экспериментальные точки лежат в прямоугольниках в любых местах. И множество прямых соединяют любые точки первого с любой точкой второго прямоугольников. На рисунке а) для наглядности все возможные линии не приведены. При согласованной базе данных и знаний объединенное множество не пусто и содержит допусковое множество. Но допусковое множество даже при 2 наблюдениях может быть пустым. На рисунке б) допусковое множество выделяет только одну прямую. И при увеличении ошибки входа или уменьшении ошибки выхода становится пустым. По нашим требованиям база данных для допускового множества уже не согласована.

ПИА: ОШИБКИ ИЗМЕРЕНИЯ ВСЕХ ПЕРЕМЕННЫХ

$$\boldsymbol{n} = \mathbf{1};\ \boldsymbol{N} = \mathbf{2};\ \boldsymbol{m} = \mathbf{2};\ \boldsymbol{y} = \boldsymbol{a}_0 + \boldsymbol{a}_1{*}\boldsymbol{x};\ \boldsymbol{a}_0{=}\boldsymbol{a}_1{=}\mathbf{1};\ \boldsymbol{\varepsilon}_x^0 = \boldsymbol{\varepsilon}_y^0 = \mathbf{0.5}$$



База данных: $\boldsymbol{T_1}:\ (\boldsymbol{x_1}, \boldsymbol{y_1}) = (\mathbf{1}, \mathbf{2});$

$\boldsymbol{T_2}:\ (\boldsymbol{x_2}, \boldsymbol{y_2}) = (\mathbf{5}, \mathbf{6}).$

$$\boldsymbol{x_i} \in \left[\boldsymbol{x_i^H};\ \boldsymbol{x_i^V}\right]; \boldsymbol{y_i} \in \left[\boldsymbol{y_i^H};\ \boldsymbol{y_i^V}\right]$$

а) Объединенное множество Auni(2)

$$\left.\begin{aligned} a_0 + a_1 x_i^H \le y_i^V \\ a_0 + a_1 x_i^V \ge y_i^H \end{aligned}\right\}, i = \overline{1, N} \qquad \mathbf{(15.1)}$$

б) Допусковое множество Atol(2)

$$\left.\begin{aligned} a_0 + a_1 x_i^H \ge y_i^H \\ a_0 + a_1 x_i^V \le y_i^V \end{aligned}\right\}, i = \overline{1, N} \qquad \mathbf{(15.2)}$$

Если $\boldsymbol{x_i^H} = \boldsymbol{x_i^V}, i = \overline{1, N}$ **то Auni(N)=Atol(N)**

Иллюстрация 16. ПИА: объединенное множество решений Auni(2)

На слайде приведены оценки параметров и прогнозы для нашего процесса с использованием объединенного множества решений.

Условие гарантированных оценок выполнено. Проверяется в модельных исследованиях методами вычислительных экспериментов. Инструментом «Поиск решения» определяется расстояние заданной точки (истинных значений параметров искомой зависимости) до информационного множества. На слайде это расстояние вычислено в ячейке G9.

Недостаток этого метода – низкая точность оценок. Оценка минимального значения свободного члена даже отрицательна (ячейка B13)

	A	B	C	D	E	F	G	H	I	J
1		Объединенное множество решений								
2		Ex=	0,5	Ey=	0,5					
3	№	X	Хн	Хв	У	Урн	Ун	Ув	Урв	
4	T1	1	0,5	1,5	2	1,500	1,500	2,500	2,500	
5	T2	5	4,5	5,5	6	5,500	5,500	6,500	6,500	
6										
7	Метод уширения				Гарантированное оценивание					
8	Кр	A0	A1		A0ис	A1ис	Невяз			
9	1,000	1,000	1,000		1	1	2,08E-12			
10					3	6		Результаты решения задачи прогноза при Ex=0 совпадают с результатами классики ПИА (Блок D10:F15)		
11		A0	A1		4,0	7,0				
12	Max	2,20	1,44	Max	4,5	7,75				
13	Мин	-1,00	0,60	Мин	3,5	6,25				
14	СР	0,60	1,02	СР	4,00	7,00				
15	ДОВ	1,60	0,42	ДОВ	0,50	0,75				

Иллюстрация 17. ПИА: допусковое множество решений Atol(2)

Аналогично предыдущему на данном слайде приведены оценки для допускового множества решений.

В данном случае для допускового множества решений ИСЛАУ выполнено условие гарантированного оценивания (расстояние истинной точки до информационного множества нулевое (ячейка G9)).

Видна не согласованность базы данных из-за нулевых значений доверительных интервалов оценок коэффициентов искомой зависимости (блок ячеек B15: C15). В данном случае это противоречие при решении задач анализа данных: по двум ошибочным наблюдениям невозможно точное оценивание зависимости.

Тогда для согласования наблюдений размах интервала оценок должен быть согласован с уровнем ошибок экспериментальных данных. Это согласование можно провести только на принципах эмпирического моделирования процессов, т.е. на основании опыта экспериментаторов и аналитиков.

ПИА: ДОПУСКОВОЕ МНОЖЕСТВО РЕШЕНИЙ Atol(2)

$$n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 {*} x;\ a_0 = a_1 = 1;\ \varepsilon_x^0 = \varepsilon_y^0 = 0.5$$

	A	B	C	D	E	F	G	H	I
1		Допусковое множество решений							
2	Слайд 16	Ex=	0,5	Ey=	0,5				
3	№	X	Xн	Xв	У	Урн	Ун	Ув	Урв
4	T1	1	0,5	1,5	2	1,500	1,500	2,500	2,500
5	T2	5	4,5	5,5	6	5,500	5,500	6,500	6,500
6									
7	Метод уширения				Гарантированное оценивание				
8	Кр	A0	A1		A0ис	A1ис	Невяз		
9	1,000	1,000	1,000		1	1	0,000		
10	1,000				3	6		Результаты	
11		A0	A1		4,0	7,0		решения задачи	
12	Max	1,00	1,00	Max	4,5	7,75		прогноза при Ex=0	
13	Мин	1,00	1,00	Мин	3,5	6,25		совпадают с	
14	СР	1,00	1,00	СР	4,00	7,00		результатами	
15	ДОВ	0,00	0,00	ДОВ	0,50	0,75		классики ПИА	

Иллюстрация 18. ПИА: согласование баз данных и знаний

Рассматриваем методы и задачи согласования базы данных на примерах современного раздела ПИА. Теоретически трудно обосновать уровень доверительного интервала оценок параметров (смотри предыдущий слайд). По опыту примем согласование по допусковому решению выполним путем уширения на 20% интервала ошибок выхода. Далее полагаем Ey=0,6. В программе коррекция выполнена заданием Кр равным 1,2 (ячейка А9).

Рассматриваем только коррекцию базы данных для допускового множества решений. В качестве реальной базы данных примем базу иллюстрации 19.

Результаты расчетов приведены в блоках ячеек В12:С15 – оценки параметров допусковым множеством решений ИСЛАУ; Е12:F15 – оценки прогнозов в точках (3 и 6) допусковым множеством решений; В16:С17 – оценки параметров объединенным множеством решений ИСЛАУ.

ПИА: СОГЛАСОВАНИЕ БАЗ ДАННЫХ И ЗНАНИЙ

$$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1 {*} x;\ a_0 = a_1 = 1;\ \varepsilon_x^0 = 0,5;\ \varepsilon_y^0 = 0.6}$$

	A	B	C	D	E	F	G	H	I
1		Допусковое множество решений-2							
2		Ex=	0,5	Ey=	0,5				
3	№	X	Хн	Хв	У	Урн	Ун	Ув	Урв
4	T1	1	0,5	1,5	2	1,50	1,40	2,60	2,50
5	T2	5	4,5	5,5	6	5,50	5,40	6,60	6,50
6									
7	Метод уширения				Гарантированное оценивание				
8	Кр	A0	A1		A0ис	A1ис	Невяз		
9	1,200	1,000	1,000		1	1	0,000		
10	1,200				3	6	Точки прогноза		
11		A0	A1		4,0	7,0			
12	Max	1,20	1,10	Max	4,1	7,12			
13	Мин	0,88	0,83	Мин	3,9	6,80			
14	СР	1,04	0,97	СР	4,00	6,96			
15	ДОВ	0,16	0,13	ДОВ	0,10	0,16			
16	СР	1,10	1,02		Слева оценки				
17	ДОВ	1,10	0,42		для ОМР				

Иллюстрация 18. ПИА: реальная база данных

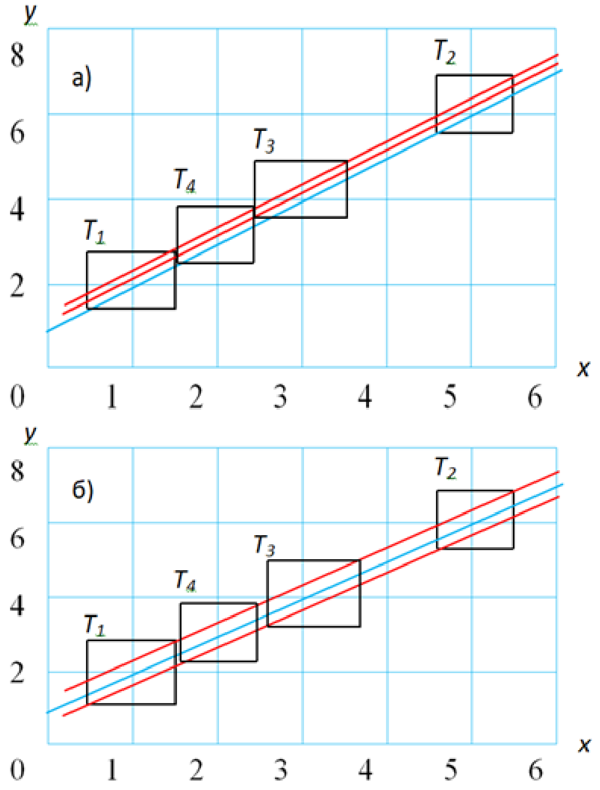
На этом слайде рассматриваем пример реальной базы данных. Проводим визуализацию допускового множества решений. По рисунку а) подозреваем, что не обеспечивается гарантированное оценивание.

Коррекцию точек наблюдений при заданных интервалах ошибок предлагаю выполнить путем расширения граничных линий в задаче прогноза. При этом требуем обеспечить условия эффективных и гарантированных оценок в задачах анализа данных.

Скорректированная база данных приведена на рисунке б) и в таблице 18.2. Все вычисления проводим в Excel, расширяя расчетную программу предыдущего слайда. Поэтапные результаты на следующих трех слайдах.

ПИА: РЕАЛЬНАЯ БАЗА ДАННЫХ

$$n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1{*}x;\ a_0{=}a_1{=}1;\ \varepsilon_x^0 = 0,5;\ \varepsilon_y^0 = 0.6$$



Исходная база данных:

	A	B	C
29		Таблица 18-1	
30	№	X	У
31	T1	1,0	2,0
32	T3	3,0	4,2
33	T4	2,0	3,1
34	T2	5,0	6,0

Скорректированная база данных

	A	B	C
36		Таблица 18-2	
37	№	X	У
38	T1	0,98	2,03
39	T3	3,02	4,16
40	T4	2,00	3,05
41	T2	5,00	6,01

Коррекция базы обеспечила гарантирование оценивание и эффективные оценки

Иллюстрация 19. ПИА: коррекция базы данных в среде Excel-1

На этом слайде доказательство того, что допусковое множество не содержит точку a=(1, 1), т.е. не обеспечивается условие гарантированных оценок.

Здесь B11 – проекция точки $a_0 = 1$ на допусковое множество решений не равна 1, а невязка G11 – не равна нулю.

ПИА: КОРРЕКЦИЯ БАЗЫ ДАННЫХ В СРЕДЕ EXCEL-1

$$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1{*}x;\ a_0{=}a_1{=}1;\ \varepsilon_x^0{=}\ 0,5;\ \varepsilon_y^0 = 0.6}$$

	A	B	C	D	E	F	G	H	I
1		Допусковое множество решений-3							
2		Ex=	0,5	Ey=	0,6				
3	№	X	Xн	Xв	У	Урн	Ун	Ув	Урв
4	T1	1	0,5	1,5	2	1,600	1,400	2,600	2,600
5	T3	3	2,5	3,5	4,2	3,600	3,600	4,800	4,600
6	T4	2	1,5	2,5	3,05	2,600	2,450	3,650	3,600
7	T2	5	4,5	5,5	6	5,600	5,400	6,600	6,600
8									
9	Метод уширения				Гарантированное оценивание				
10	Кр	A0	A1		A0ис	A1ис	Невяз		
11	1,000	1,100	1,000		1	1	0,010		
12	1,000				3	6			
13		A0	A1		4,1	7,1			
14	Max	1,20	1,10	Max	4,1	7,12			
15	Мин	0,88	0,83	Мин	3,9	6,80			
16	СР	1,04	0,97	СР	4,00	6,96			
17	ДОВ	0,16	0,13	ДОВ	0,10	0,16			
18	СР	1,10	1,02		Слева оценки для				
19	ДОВ	1,10	0,42		ОМР				

Иллюстрация 20. ПИА: коррекция базы данных в среде Excel-2

На данном слайде приведены таблицы Excel после коррекции базы данных методом предварительной коррекции (см. стр. 53). Идея метода состоит в том, чтобы максимально раздвинуть границы матриц ИСЛАУ при минимуме нормы $\|e\|$, что требует постановки задачи оптимизации по двум критериям. В нашем случае, скорректированная база данных обеспечивает максимум разности верхней и нижней оценок в задаче интервального прогноза.

Изменяемые ячейки блок (B10:G11), параметры уравнения и корректирующие коэффициенты. Решена задача нелинейного программирования методом ОПГ с 12 переменными и 20 формульными ограничениями. Целевая ячейка I11. Ее значение определяется разностью верхней и нижней оценок прогноза выходной переменной в точке x= 3. Норма $\|e\|$ вычислена в ячейке H14. Свертка критериев вычислена с весом, значение которого задается в ячейке I13.

ПИА: КОРРЕКЦИЯ БАЗЫ ДАННЫХ В СРЕДЕ EXCEL-2

$$\boldsymbol{n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1{*}x;\ a_0{=}a_1{=}1;\ \varepsilon_x^0 = 0,5;\ \varepsilon_y^0 = 0.6}$$

	A	B	C	D	E	F	G	H	I
1		Коррекция базы данных. Допусковое множество решений							
2	Слайд 20	Ex=	0,5	Ey=	0,6				
3	№	X	Xн	Xв	У	Урн	Ун	Ув	Урв
4	T1	0,98	0,48	1,48	2,03	1,64	1,43	2,63	2,63
5	T3	3,02	2,52	3,52	4,16	3,56	3,56	4,76	4,66
6	T4	2,00	1,50	2,50	3,05	2,60	2,45	3,65	3,64
7	T2	5,00	4,50	5,50	6,01	5,41	5,41	6,61	6,61
8									
9	Кр	A0	A1	T1	T3	T4	T2	Разность	MAX
10	1,000	1,167	0,990	-0,04	0,05	0,00	-0,01	В-Н	0,131
11	1,000	1,195	0,937	0,05	-0,07	0,00	0,01	ЦФ	0,091
12	Прозн-В	3	4,137	0,002	0,002	0,000	0,000	0,004	Вес
13	Прозн-Н	3	4,005	0,003	0,005	0,000	0,000	0,007	3,5
14				1,000	1,000	1,000	1,000	0,012	0,040

Иллюстрация 21. ПИА: коррекция базы данных в среде Excel-3

На этом слайде представлены оценки скорректированной базы данных. Множество неопределенности – допусковое множество решений. Оценки гарантированные и сравнимы с оценками ОМР при Ex=0. Оценки ОМР для данной базы вычисленные значительно хуже.

Для анализа результатов слайда следует обратиться к Приложению 1 к листу программы, на которой заданы условия допускового множества решений ИСЛАУ. Это блок ячеек на строках 5 – 8. Проверка условия гарантированного оценивания выполняется по значению расстояния истинного значения коэффициентов уравнения до исследуемого множества. Эта величина приведена в ячейке G12. Учитываем значение параметра уширения (ячейка A12).

ПИА: КОРРЕКЦИЯ БАЗЫ ДАННЫХ В СРЕДЕ EXCEL-3

$$n = 1;\ N = 2;\ m = 2;\ y = a_0 + a_1{*}x;\ a_0{=}a_1{=}1;\ \varepsilon_x^0 = 0,5;\ \varepsilon_y^0 = 0.6$$

	A	B	C	D	E	F	G	H	I
1	Допусковое множество решений-5. Согласованная база данных								
2									
3	Слайд 21	Ex=	0,5	Ey=	0,6				
4	№	X	Хн	Хв	У	Урн	Ун	Ув	Урв
5	Т1	0,98	0,48	1,48	2,03	1,48	1,43	2,63	2,48
6	Т3	3,02	2,52	3,52	4,16	3,52	3,56	4,76	4,52
7	Т4	2,00	1,50	2,50	3,05	2,50	2,45	3,65	3,50
8	Т2	5,00	4,50	5,50	6,01	5,50	5,41	6,61	6,50
9									
10	Метод уширения				Гарантированное оценивание				
11	Кр	А0	А1		А0ис	А1ис	Невяз		
12	1,000	1,000	1,000		1	1	0,000		
13	1,000				3	6			
14		А0	А1		4,0	7,0			
15	Max	1,31	1,03	Max	4,16	7,13			
16	Мин	0,97	0,91	Мин	3,97	6,79			
17	СР	1,14	0,97	СР	4,06	6,96			
18	ДОВ	0,17	0,06	ДОВ	0,09	0,17			
19	СР	1,10	1,02		Слева оценки для ОМР				
20	ДОВ	1,10	0,42		при Ex=0				

Учебное издание

Оскорбин Николай Михайлович

Алгазин Геннадий Иванович

Суханов Сергей Иванович

МЕТОДЫ И МОДЕЛИ ИНТЕРВАЛЬНОГО АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Издано в авторской редакции

Дизайн обложки Ю.В. Луценко

Издательская лицензия ЛР 020261 от 14.01.1997.
Подписано в печать 19.05.2023.
Дата выхода в свет 26.05.2023.
Формат 60x84 1/16. Бумага офсетная.
Усл.-печ. л. 5,46. Тираж 100. Заказ 338.

Типография Алтайского государственного университета:
656049, Барнаул, ул. Димитрова, 66