УДК 531.761

Согласование базы данных в прикладном интервальном анализе

Е.К. Ергалиев[1], М.Н. Мадияров[1], Н.М. Оскорбин[2], Л.Л. Смолякова[2]

[1]Восточно-Казахстанский университет им. С. Аманжолова (Усть-Каменогорск, Казахстан)
[2]Алтайский государственный университет (Барнаул, Россия)

Database Reconciliation in Applied Interval Analysis

E.K. Ergaliev[1], M.N. Madiyarov[1], N.M. Oskorbin[2], L.L. Smolyakova[2]

[1]Sarsen Amanzholov East Kazakhstan University (Ust-Kamenogorsk, Kazakhstan)
[2]Altai State University (Barnaul, Russia)

Рассматривается проблема согласования результатов наблюдений, которая возникает в задачах прикладного интервального анализа. Установлено, что значения совокупности входных переменных и выходной переменной согласованы, если график искомой зависимости расположен во всех внутренних точках интервального гиперпрямоугольника для каждого наблюдения. В этом случае при анализе данных линейных процессов в литературе предложено использовать допусковое множество решений интервальных систем линейных алгебраических уравнений (ИСЛАУ). Однако в реальных и модельных условиях указанное согласование базы данных априори не всегда выполняется. Авторы статьи предлагают использовать принцип робастного оценивания: несогласованные наблюдения следует либо исключить из выборки, либо скорректировать. В настоящей работе 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.

Ключевые слова: моделирование процессов, согласование базы данных, интервальные системы линейных алгебраических уравнений, допусковое множество решений ИСЛАУ, объединенное множество решений ИСЛАУ, компьютерное моделирование.

The article deals with the problem of the reconciliation of observation results, which arises when solving problems of interval analysis of a database. It is found that the values of the set of input variables and the output variable are consistent if the graph of the desired dependence is located at the inner points of the interval hyper-rectangle in each observation. In this case, it is proposed to use special solutions of interval systems of linear algebraic equations (ISLAU) to analyze the data of linear processes. However, in real and model conditions, the specified property of the database is not always fulfilled a priori. In these cases, it is proposed to use the principle of robust estimation: inconsistent observations should either be excluded from the sample or adjusted. This paper presents the results of the study of these methods of matching the used experimental database on model linear processes under conditions when the basic assumptions of interval estimation of dependencies are fulfilled. In addition, variant computational experiments have been investigated to reveal the possibility of increasing the accuracy of interval analysis due to preliminary correction of observations, including the possibility of guaranteed estimation of the sought dependences.

Keywords: process modeling, database reconciliation, interval systems of linear algebraic equations, ISLAU Solutions for a Consistent Database, united solution set, computer modelling.

DOI: 10.14258/izvasu(2022)1-14

Введение

Рассматривается предметная область прикладного интервального анализа данных при идентификации по экспериментальным данным параметров линейных детерминированных процессов, которая описана в работе [1]. Математическая модель и теоретиче-

ские основы анализа данных базируются на методах интервальных систем линейных алгебраических уравнений (ИСЛАУ). 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 работах [2–5]. В работе [6] затронут аспект анализа данных, содержащих выбросы, а в [7] обсуждены возможности гарантированного оценивания параметров исследуемой зависимости переменных моделируемого процесса.

Следует отметить, что исходная идея использования интервальной математики и линейного программирования (ЛП) в задачах анализа данных высказана Л. Канторовичем в работе [2]. Теоретические исследования в области прикладного интервального анализа данных представлены, например, в работах [3, 4]. 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.

Первую проблему поясним на примере оценки погрешности суммы большого числа интервальных данных. Гарантированная оценка суммы определяется 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 С.П. Шарый, с общим названием «сильное согласование» [8, 9, 10].

Метод основан на использовании допускового множества решений ИСЛАУ и характеризуется тем, что точечные оценки восстановления линейных зависимостей являются эффективными. Принцип сильного согласования базируется на следующей гипотезе [8]: «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 в базовом варианте оказывается пустым. В этом случае для точечного оценивания параметров в работе [8] предложен метод уширения брусов неопределенности данных (далее метод уширения), который можно рассматривать как один из методов согласования базы данных. Кроме того, допусковое множество решений ИСЛАУ в общем случае может не содержать истинных значений оцениваемых параметров [1], т.е. не выполняется принцип гарантированного оценивания. Таким образом, проблема согласования базы данных в прикладном интервальном анализе требует дополнительных исследований.

В данной работе гипотеза согласования интервальных наблюдений рассматривается как дополнение к базовым предположениям прикладного интервального анализа. Мы считаем, что сформулированная в [8] 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 [11, 12]: несогласованные наблюдения следует либо исключить из выборки, либо скорректировать. Нами представлены 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 [2].

В статье представлены математическое моделирование процессов коррекции базы данных и результаты исследования точности анализа данных, которое проводится методами вычислительной математики.

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

Описание методов и методик проводимого исследования

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: $\boldsymbol{A}x = \boldsymbol{B}$. 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:

$$b = x_1 a_1 + \cdots + x_n a_n. \qquad (1)$$

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

$$\boldsymbol{a}_{ij} = [a_{ij}^H,\ a_{ij}^V];\ a_{ij}^H = a_{ij}^M - \varepsilon_{ij}^0;\ a_{ij}^V = a_{ij}^M + \varepsilon_{ij}^0;\ i = 1,\ \ldots\ n;\ j = 1,\ \ldots,\ N; \tag{2}$$

$$\boldsymbol{b}_j = [b_j^H, b_j^V];\ b_j^H = b_j^M - \varepsilon_{bj}^0;\ b_j^V = b_j^M + \varepsilon_{bj}^0;\ j = 1,\ \ldots,\ N, \tag{3}$$

где a_{ij}^M, b_j^M — экспериментальные данные входов и выхода в каждом из N наблюдений; $\varepsilon_{ij}^0, \varepsilon_{bj}^0$ неотрицательные 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 представленной работе мы ограничимся исследованием объединенного и допускового множеств решений ИСЛАУ, считая, что элементы этих множеств неотрицательны. Объединенное множество решений задается системой линейных неравенств, которые запишем в следующем виде [1]:

$$\Xi_{uni}(\boldsymbol{A},\boldsymbol{B}) = \{x \in R_+^n : A^V x \geq B^H; A^H x \leq B^V\}. \tag{4}$$

Допусковое множество решений ИСЛАУ задается так [1]:

$$\Xi_{tol}(\boldsymbol{A},\boldsymbol{B}) = \{x \in R_+^n : A^V x \leq B^V; A^H x \geq B^H\}. \tag{5}$$

Корректирование всех наблюдений проводим в следующем виде:

$$\widehat{a}_{ij}^M = a_{ij}^M + e_{ij}\varepsilon_j^0;\ \widehat{b}_j^M = b_j^M + e_{bj}\varepsilon_b^0;\ |e_{ij}| \leq 1;\ |e_{bj}| \leq 1. \tag{6}$$

Скорректированные согласно (6) экспериментальные принадлежат интервалам в выражениях (2), (3), 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), (5) для базы данных (6).

Рассмотрим задачи исследования множеств решений ИСЛАУ, которые возникают при моделировании процессов. Обозначим X — одно из введенных множеств решения ИСЛАУ и запишем выражения для прикладных задач анализа данных.

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

$$b^H(a^P) = \min_{x \in X}(x_1 a_1^p + \cdots + x_n a_n^p);\ b^V(a^P) = \max_{x \in X}(x_1 a_1^p + \cdots + x_n a_n^p). \tag{7}$$

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

3. Проекция точки x^P на множество X. Точка x^P принадлежит многограннику X тогда и только тогда, когда δ равно нулю, где δ — решение следующей задачи квадратичного программирования: $\delta = \min_{x \in X} \| x^P - x \|$. Эта задача решается при тестировании процедур анализа данных в модельных условиях на возможность получения гарантированных оценок.

4. Пример задачи коррекции базы данных. Рассматриваем ИСЛАУ, для которой допусковое множество решений является пустым. Это обстоятельство — достаточное основание считать отдельные наблюдения базы данных несогласованными. Необходимо найти коэффициенты корректирования $e^* \in E$ и оценки вектора $x^* \in \Xi_{tol}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E)$ в перечисленных выше задачах прикладного интервального анализа. В следующем разделе рассмотрим решение данной задачи в один и в два этапа. В качестве примера рассмотрим одноэтапную оценку $b^H(a^P)$ при решении задачи прогноза (7).

$$b^H(a^P) = \min_{\substack{\Xi_{tol}(\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}, E) \\ b \in E}} \left((x_1 a_1^p + \cdots + x_n a_n^p) + L \| e \| \right). \tag{8}$$

В данном случае для коррекции базы данных и оценки параметров записана двухкритериальная задача нелинейного программирования. При достаточно большом значении параметра L обеспечиваются согласование базы данных $\Xi_{tol}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},E)\neq\phi$ и минимальная коррекция значений экспериментальных данных. Задачи типа (8) имеют большое число переменных и ограничений и требуют на практике применения специальных методов их решения [14].

Имитационное моделирование процессов согласования базы данных

Методическое и программное обеспечение имитационного моделирования прикладного интервального анализа данных описано в работе [1]. 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 (1) с параметрами $x_1 = x_2 = 1$; $a_1\, a_2 \in$ [10, 100]. Задача прогноза (7) исследовалась для a^p = (55, 55). Многовариантность вычислительных экспериментов достигалась изменением интервалов $\varepsilon_1^0, \varepsilon_2^0, \varepsilon_b^0$, 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 несогласованных наблюдений (метод исключения); согласование базы данных по критерию минимума нормы $||e||$ (метод минимальной коррекции); согласование по типу задачи (8) (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 оценках ошибок наблюдения (табл. 1, 2). Все исходные базы данных содержали несогласованные 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 1 и 2. Исключение несогласованных наблюдений проводилось до единичного *Кр,* и их число *Nuc* изменялось от 2 до 10. Для исходной и скорректированной баз данных получены интервальные оценки прогноза (7) 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 1

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

№ опыта	Метод исключения				ОМР	
	Кр	*Nuc*	*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

Таблица 2

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

№ опыта	Метод исключения				ОМР	
	Кр	*Nuc*	*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

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 $x_1 = x_2 = 1$.

2. Метод минимальной коррекции. В данном методе коэффициенты корректирования $e^* \in E$ и оценки вектора $x^* \in \Xi_{tol}(\widehat{\boldsymbol{A}},\widehat{\boldsymbol{B}},E)$ 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 несогласованных баз данных. Среднее число скорректированных наблюдений оказалось равным 7,3, погрешность прогноза по среднему значению равна 0,14 % и соответствует точности методу уширения [8]. Метод минимальной коррекции можно рекомендовать на практике для согласования наблюдений при проведении натурных экспериментов.

3. Метод совместной коррекции. Исследования проведены для задачи (7) с параметрами процесса, как в методе п. 2. Результаты испытаний приведены в таблице 3, где *Nк* — число скорректированных наблюдений.

Таблица 3

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, обозначение переменных в которой аналогично обозначениям таблицы 3.

Таблица 4

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.

Заключение

Для решения проблемы согласования базы данных в прикладном интервальном анализе, возможно впервые, реализован принцип робастного оценивания: несогласованные 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.

Полученные результаты позволяют уточнить методические подходы применения теоретических результатов ИСЛАУ в задачах анализа данных и математического моделирования реальных процессов.

Библиографический список

1. Мадияров М.Н., Оскорбин Н.М., Суханов С.И. Примеры интервального анализа данных в задачах моделирования процессов // Известия Алт. гос. ун-та. 2018. № 1 (99). DOI: 10.14258/izvasu(2018)1-20.

2. Канторович Л.В. О некоторых новых подходах к вычислительным методам и обработке наблюдений // Сибирский математический журнал. 1962. Т. 3. № 5.

3. Шарый С.П. Конечномерный интервальный анализ. Новосибирск, 2017.

4. Жолен Л. Прикладной интервальный анализ. М. : Ижевск, 2005.

5. Gutowski M.W. Interval experimental data fitting. In: Liu, J.P. (ed.): Focus on 6. Numerical Analysis. Nova Science, New York, NY, USA (2006). https://doi.org/10.13140/2.1.5156.3520.

6. Zhilin S.I. Simple method for outlier detection in fitting experimental data under interval error // Chemometrics and Intellectual Laboratory Systems. 2007. Vol. 88. № 1.

7. Шелудько А.С. Гарантированное оценивание параметров дискретных моделей хаотических процессов // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2018. Т. 7. № 1. DOI: 10.14529/cmse180103.

8. Шарый С.П. Сильная согласованность в задаче восстановления зависимостей при интервальной неопределенности данных // Вычислительные технологии. 2017. Т. 22. № 2.

9. Shary S.P. Maximum consistency method for data fitting under interval uncertainty. Journal of Global Optimization, 2016. № 66 (1). https://doi.org/10.1007/s10898-015-0340-1.

10. Shary S.P. Weak and strong compatibility in data fitting problems under interval uncertainty. Advances in Data Science and Adaptive Analysis. 2020. № 12 (1). https://doi.org/10.1142/S2424922X20500023.

11. Хьюбер П. Робастность в статистике. М., 1984.

12. Miller B.M., Kolosov K.S. Robust Estimation Based on the Least Absolute Deviations Method and the Kalman Filter // Automation and Remote Control. 2020. Vol. 81. № 11. DOI: 10.1134/S0005117920110041.

13. Максимов А.В., Оскорбин Н.М. Многопользовательские информационные системы: основы теории и методы исследования ; 2-е изд. испр. и доп. Барнаул, 2013.

14. Оскорбин Н.М. Вычислительные технологии анализа больших данных методами линейного программирования // Высокопроизводительные вычислительные системы и технологии. 2021. Т. 5. № 1.