УДК 681.51:57

С.Я. МАЙСТРЕНКО

ОРИЕНТИРОВОЧНЫЕ ОЦЕНКИ ТОЧНОСТИ АГРЕГИРОВАННЫХ ПОКАЗАТЕЛЕЙ В МНОГОУРОВНЕВЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Abstract: *The model is constructed and the dependences determining an expected declination of aggregated (summarized) quantitative parameters are received the part of which is deformed by mistakes, from their real meaning. The received relations and estimations can be useful in systems of support of making the decisions oriented on various problem areas (armed forces, economy, sociology etc.).*
Key words: *aggregated parameters, quantitative estimations, multilevel information systems.*

Анотація: *Побудовано модель та отримано залежності, що визначають відхилення агрегованих (шляхом додавання) показників, частина з яких спотворена помилками, від їхнього істинного значення. Отримані співвідношення і оцінки можуть бути корисними в системах підтримки рішень, орієнтованих на різні проблемні області (збройні сили, економіка, соціологія та ін.).*
Ключові слова: *агреговані показники, кількісні оцінки, багаторівневі інформаційні системи.*

Аннотация: *Построена модель и получены зависимости, определяющие ожидаемое отклонение агрегированных (суммированных) количественных показателей, часть из которых искажена ошибками, от их истинного значения. Полученные соотношения и оценки могут быть полезны в системах поддержки принятия решений, ориентированных на различные проблемные области (вооруженные силы, экономика, социология и др.).*
Ключевые слова: *агрегированные показатели, количественные оценки, многоуровневые информационные системы.*

1. Введение

В теории и практике обработки данных достоверность элементов информации обычно оценивается вероятностью их искажения [1, 2]. По отношению к атрибутам, отражающим качественные характеристики объектов предметной области, такая оценка вполне адекватна, т.к. ошибка даже в одном символе атрибута приводит к неправильной интерпретации записи (кортежа) в целом. Для количественных показателей существенным является не только факт их искажения, но и степень отклонения от истинного значения, т.е. точность представления.

Существуют две основные причины возможной неточности количественных показателей. Первая из них носит случайных характер и связана с возможной приблизительностью некоторых оценок при формировании и регистрации первичной информации. В этом случае наиболее вероятно, что могут быть искажены младшие разряды количественного показателя, т.е. результатная неточность не должна быть высокой. Для оценки случайных искажений в конкретных случаях возможно применение аппарата математической статистики, а именно теории ошибок в измеряемых величинах [3, 4]. Вторая причина связана с возможными «грубыми» ошибками при подготовке и вводе первичной информации в компьютер. В этом случае пользователь может сделать ошибку в любой цифре, в том числе и самой старшей, практически с одинаковой вероятностью, в результате чего неточность представления отдельных показателей может оказаться значительной.

Процесс решения многих задач в многоуровневых информационных системах связан со сбором первичных количественных показателей, их последовательной агрегацией (в частности, суммированием), накоплением и последующей обработкой. В связи с возможностью появления

грубых ошибок в отдельных первичных показателях представляет интерес оценка ожидаемой точности агрегированных показателей.

Процесс накопления ошибок в агрегированном показателе с истинным значением $S_0^{(m)}$ схематически показан на рис.1.

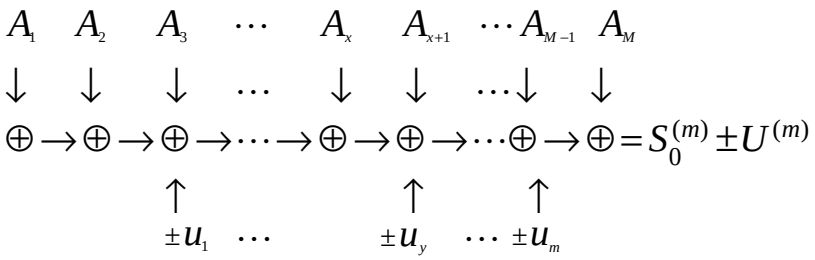

Рис. 1. Схема процесса накопления ошибок

Обозначим через q основание системы счисления показателя, а через n – количество разрядов. Задача оценки ожидаемой точности агрегированного показателя (будем говорить «оценки ожидаемого смещения значения агрегата») ставится следующим образом. Суммируется M независимых n-разрядных случайных величин, m из которых искажены статистически одинаковыми грубыми ошибками с математическим ожиданием u_y (единичным смещением) и дисперсией d_y (единичной дисперсией). Задача заключается в построении модели и определении:

1. Модуля математического ожидания суммы m ошибок $U^{(m)} = \left| \sum_{y=1}^{m} u_y \right|$

и дисперсии суммы m ошибок $D^{(m)} = \sum_{y=1}^{m} d_y$;

2. Значений u_y, d_y для наиболее типичных ошибок пользователя.

Поставленная задача, адекватно отражающая процесс искажения грубыми ошибками агрегируемых данных, не типична для теории измерений [3, 4] и решить ее на основе известных результатов [3] не представляется возможным. В связи с этим необходимы построение и исследование соответствующих моделей.

Примем следующие обозначения и допущения:

M – количество составляющих слагаемых агрегированного показателя (агрегата);

m – количество ошибочных значений в агрегате, искаженных на случайную величину с математическим ожиданием $\pm u^{(1)}$ и дисперсией $d^{(1)}$ (единичное смещение и дисперсия);

n – количество разрядов суммируемых первичных показателей (будем считать его постоянным);

q – основание системы счисления в представлении показателя;

$U^{(m)}$ и $D^{(m)}$ – математическое ожидание и дисперсия модуля смещения суммы m искаженных показателей (смещение и дисперсия агрегата).

С учетом принятых обозначений расчетное значение агрегата $S(M)$ представляет собой случайную величину с математическим ожиданием $S(M) = S_0(M) \pm U^{(m)}$ и среднеквадратичным отклонением $\sigma^{(m)} = \sqrt{D^{(m)}}$, где $S_0(M)$ – истинное значение агрегата.

Задача заключается в построении модели и определении значений $U^{(m)}$, $D^{(m)}$, $\sigma^{(m)}$.

2. Оценка ожидаемого смещения значения агрегата

Рассмотрим возможные результаты ожидаемых смещений суммы двух показателей $(m = 2)$, каждый из которых искажен в среднем на величину $\pm u^{(1)}$. Поскольку математическое ожидание суммы равно сумме математических ожиданий, ожидаемое суммарное смещение U_2 равно

$$U_2 = \pm u^{(1)} \pm u^{(1)} .$$

В зависимости от сочетания знаков единичных смещений возможны 4 варианта значений U_2:

$U_{21} = \left| - u^{(1)} - u^{(1)} \right| = 2u^{(1)}$; $U_{23} = \left| + u^{(1)} - u^{(1)} \right| = 0$;

$U_{22} = \left| - u^{(1)} + u^{(1)} \right| = 0$; $U_{24} = \left| + u^{(1)} + u^{(1)} \right| = 2u^{(1)}$.

Поскольку все варианты равновероятны, то

$$u^2 = \frac{4u^{(1)}}{4} = u^{(1)} .$$

В свою очередь, дисперсия в силу известных положений теории вероятности равна сумме единичных дисперсий, т.е. $D^{(2)} = 2d^{(1)}$.

Обобщая приведенный пример на случай суммы m искаженных показателей, поставим в соответствие m показателям m- разрядное двоичное число. Припишем k-му разряду следующий смысл: если значение разряда равно 0, единичное смещение соответствующего показателя имеет знак "+", в противном случае единичное смещение имеет знак "-". В такой интерпретации, например, для $m = 8$ двоичному числу 01110111 соответствует случайное смещение

$$U_8 = \left| + 2u^{(1)} - 6u^{(1)} \right| = \left| - 4u^{(1)} \right| = 4u^{(1)} .$$

Дальнейшее рассуждение иллюстрирует табл.1, в которой через $m(1)$ обозначено количество единиц в некотором m - разрядном двоичном числе, а через $m(0)$ – количество нулей.

Таблица 1. Характеристики модуля суммарного смещения

$C(1)$	$C(0)$	Количество чисел, соответствующих данной комбинации $m(1)$, $m(0)$	Модуль суммарного смещения
0	m	1	$mu^{(1)}$
1	$m-1$	C_m^1	$(m-2)u^{(1)}$
2	$m-2$	C_m^2	$(m-4)u^{(1)}$
…	…	…	…
$m-1$	1	C_m^1	$(m-2)u^{(1)}$
m	0	1	$mu^{(1)}$

С учетом того, что общее количество всевозможных m – разрядных двоичных чисел равно 2^m, из табл. 1 следует, что

$$U^{(m)} = k^{(m)} u^{(1)}, \tag{1}$$

где

$$k^{(m)} = \begin{cases} \frac{2}{2^m} \sum_{i=1}^{\frac{m}{2}} C_m^i (m-2i) \quad для \quad четных\ m\ ; \\ \frac{2}{2^m} \sum_{i=1}^{\frac{m-1}{2}} C_m^i (m-2i) \quad для \quad нечетных\ m\ . \end{cases} \tag{2}$$

Соответственно

$$D^{(m)} = md^{(1)}. \tag{3}$$

В табл. 2 приведены значения $k^{(m)}$, рассчитанные для $m = 1-101$. Как видно из табл. 2, с ростом m значение $U^{(m)}$ тоже растет, но значительно медленнее, чем m, примерно в логарифмическом масштабе $\log_2 m$ (в очень грубом приближении). Так, например, увеличение m в 100 раз (с 1 до 100) вызывает увеличение $k^{(m)}$ только примерно в 8 раз.

Таблица 2. Зависимость $k^{(m)}$ от m

m	1	2	3	4	7	8	10	50	100	101
$k^{(m)}$	1,00	1,00	1,50	1,50	2,19	2,19	2,46	5,61	7,96	8,04

3. Оценка единичного смещения и дисперсии

Определяя значение $u^{(1)}$, ограничимся предположением, что ошибочное значение n – разрядного показателя возникло в результате искажения истинного значения однократной ошибкой. Это допущение хорошо согласуется с результатами ввода информации сканером и близко к реальности при вводе с клавиатуры [2].

Рассмотрим вначале случай $n=1$. Определим возможные смещения значения одного разряда, представленного в системе счисления с основанием q.

В квадратной матрице табл. 3 приведены значения модуля смещения, соответствующего переходу каждого из истинных значений a_x из множества $\{0,1,\ldots,\mathrm{q}-1\}$ любое ложное a_y из того же множества.

Таблица 3. Значения модуля смещения

a_y \ a_x	0	1	2	...	$q-3$	$q-2$	$q-1$
0	0	1	2	...	$q-3$	$q-2$	$q-1$
1	1	0	1		$q-4$	$q-3$	$q-2$
2	2	1	0		$q-5$	$q-4$	$q-3$
.	.	.	.		.	.	.
.	.	.	.		.	.	.
.	.	.	.		.	.	.
$q-3$	$q-3$	$q-4$	$q-5$		0	1	2
$q-2$	$q-2$	$q-3$	$q-4$		1	0	1
$q-1$	$q-1$	$q-2$	$q-3$		2	1	0

Из табл. 3 вытекает, что возможное случайное отклонение составляет $(q-i)$ с вероятностью $\frac{2\mathrm{i}}{\mathrm{q}(\mathrm{q}-1)}$. Здесь $i=1,\ldots\ldots,q-1$. В n- разрядном показателе может быть искажен любой из n разрядов с вероятностью $\frac{1}{n}$. С учетом того, что вклад разряда с позицией j $(\mathrm{j}=1,\ldots\ldots,\mathrm{n})$ в значение n –разрядного числа составляет q^{j-1}, значение $u^{(1)}$ определяется следующим образом:

$$u^{(1)}=\frac{2\sum_{i=1}^{q-1}\sum_{j=1}^{n} i(q-i)q^{j-1}}{nq(q-1)}. \qquad (4)$$

Соответственно

$$d^{(1)} = \frac{2\sum_{i=1}^{q-1}\sum_{j=1}^{n}\left[(q-i)q^{j-1} - u^{(1)}\right]^2 \cdot i}{nq(q-1)}. \qquad (5)$$

В табл. 4 приведены расчетные данные для $u^{(1)}$, $d^{(1)}$ и среднеквадратичного отклонения $\sigma^{(1)} = \sqrt{d^{(1)}}$ для $q = 10$.

Таблица 4. Расчетные данные

n	1	4	6	8	10	12
$u^{(1)}$	3,67	$1{,}02\cdot10^{3}$	$6{,}79\cdot10^{4}$	$5{,}09\cdot10^{6}$	$4{,}07\cdot10^{8}$	$3{,}39\cdot10^{10}$
$d^{(1)}$	4,89	$3{,}59\cdot10^{6}$	$2{,}62\cdot10^{10}$	$2{,}05\cdot10^{14}$	$1{,}68\cdot10^{18}$	$1{,}42\cdot10^{22}$
$\sigma^{(1)}$	2,21	$1{,}89\cdot10^{3}$	$1{,}62\cdot10^{5}$	$1{,}43\cdot10^{7}$	$1{,}29\cdot10^{9}$	$1{,}19\cdot10^{11}$

4. Анализ полученных зависимостей

Для адекватной и наглядной интерпретации выражений (1) – (5) и иллюстрирующих их данных табл. 2, 4 рассмотрим оценки относительных значений $U^{(m)}$ и $\sigma^{(m)}$ по отношению к среднему значению агрегата $\overline{S}(M)$.

Примем среднее значение агрегата равным половине его максимального значения, т.е. $\overline{S}(M) = 0{,}5M\,q^{n}$. Тогда

$$U_{отн}^{(m)} = \frac{U^{(m)}}{0{,}5\cdot M\cdot q^{n}} = \frac{k^{(m)}\cdot u^{1}}{0{,}5\cdot M\cdot 10^{n}}; \qquad (6)$$

$$\sigma_{отн}^{(m)} = \frac{\sqrt{D^{(m)}}}{0{,}5\cdot M\cdot q^{n}} = \frac{\sqrt{m\,d^{(1)}}}{0{,}5\cdot M\cdot 10^{n}}. \qquad (7)$$

Возьмем $m = \gamma M$. Основываясь на экспериментальных данных [2], можно показать, что при определенных предположениях относительно возможностей возникновения и обнаружения первичных ошибок, допущенных на этапе формирования исходных данных, и вторичных ошибок, возникающих на этапе ввода в ЭВМ, ориентировочное значение $\frac{\gamma}{\mathrm{n}}$ находится в диапазоне $10^{-2} \div 10^{-3}$. Для ориентировочных оценок ожидаемых значений $\mathrm{U}_{\mathrm{отн}}^{(\mathrm{m})}$ и $\sigma_{\mathrm{отн}}^{(\mathrm{m})}$ и их трендов положим $\frac{\gamma}{\mathrm{n}} = 5\cdot10^{-3}$. На рис. 2 и 3 приведены группы зависимостей $\mathrm{U}_{\mathrm{отн}}^{(\mathrm{m})}$ и $\sigma_{\mathrm{отн}}^{(\mathrm{m})}$ от M для $q = 10,\ n = 3,4,6,\ \frac{\gamma}{n} = 5\cdot10^{-3}$.

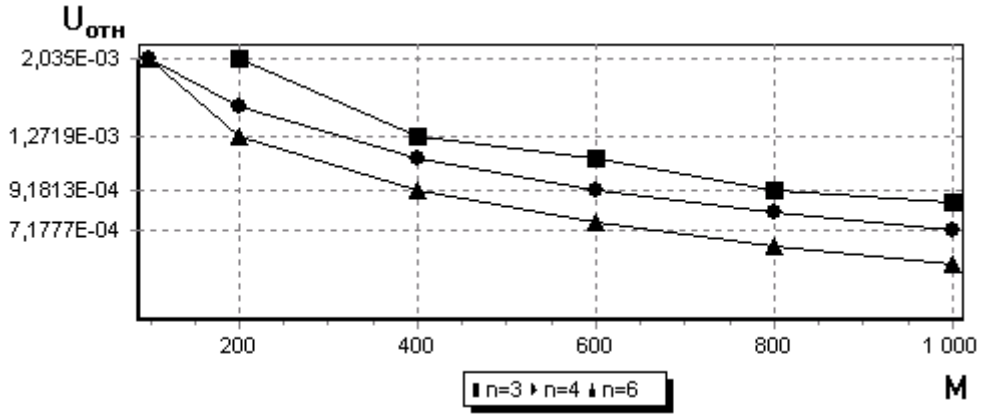

Рис. 2. Зависимости $U_{отн}$ от M

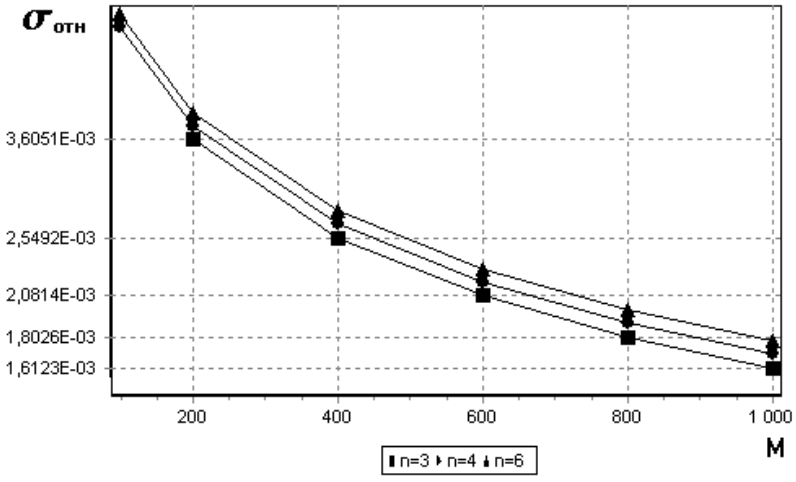

Рис. 3. Зависимости $\sigma_{отн}$ от M

Анализ выражений $(1) \div (7)$, данных таблиц 2, 4 и рисунков 2, 3 позволяет сделать следующие общие выводы:

1. Ожидаемое относительное смещение значения агрегата медленно убывает с ростом n и M, оставаясь при этом весьма небольшим. Так, для $n = 6, M = 100 - 1000$, $m = 3 - 30$ значение $\mathrm{U}_{\mathrm{отн}}^{(\mathrm{m})}$ составляет всего $0{,}2\% \div 0{,}059\%$.

2. Относительное среднеквадратичное отклонение также убывает с ростом M, n (как видно из (3) и (7), пропорционально $\sqrt{m}$) и, как и следовало бы ожидать, сравнительно велико. Оценки предельных величин фактических значений смещения агрегата относительно математического ожидания могут быть получены на основе известного неравенства Чебышева [3]:

$$p(t) \leq \sigma^2 / t^2 , \qquad (8)$$

где $\mathrm{p(t)}$ – вероятность отклонения значения случайной величины (в данном случае, смещения) от математического ожидания более чем на t единиц;

σ – среднеквадратическое отклонение.

Так, например, на основе (8) можно показать, что для $n = 6,\ M = 600$ вероятность смещения в 7,5% относительно $\overline{S}(M)$ меньше 0,09. При этом следует оговориться, что универсальная оценка (8), не учитывающая распределения случайных величин, является весьма грубой и, чаще всего, завышенной.

5. Заключение

1. В работе построена модель процесса искажения грубыми ошибками (на примере однократных транскрипций) агрегируемых показателей.
2. Полученные соотношения и иллюстрирующие их данные позволяют ориентировочно оценить ожидаемое значение отклонения агрегата от истинного значения в зависимости от количества суммируемых показателей и числа ошибок.
3. Полученные в работе оценки ожидаемого отклонения могут быть полезны в системах поддержки принятия решений, ориентированных на различные проблемные области (вооруженные силы, экономика, социология и др.). Уточнение этих оценок возможно за счет учета различных типов ошибок при определении единичного смещения.
4. В перспективе предложенный подход может быть распространен на оценку точности количественных показателей и для других алгоритмов их обработки, выходящих за рамки простого суммирования.

СПИСОК ЛИТЕРАТУРЫ

1. Пивоваров А.Н. Методы обеспечения достоверности информации в АСУ: Обзор методов и фактические данные. – М.: Радио и связь, 1982. – 144 с.
2. Литвинов В.А., Крамаренко В.В. Контроль достоверности и восстановление информации в человеко-машинных системах. – К.: Техника, 1988. – 200 с.
3. Феллер В. Введение в теорию вероятностей и ее приложения. – Москва: Мир, 1967. – 498 с.
4. Горбань І.І. Теорія ймовірностей і математична статистика. – К.: Інститут проблем математичних машин та систем, 2003. – 244 с.