6. Kurs sotsial'No.-ekonomicheskoy statistiki: Uchebnik / Pod red. M.G. Nazarova. 5-ye izd., pererab. i dop. - M.: Izd-vo Omega-L, 2006. - 984 s. [Course of social-and-economic statistics: Textbook / Ed. by M.G. Nazarov. 5th ed., Rev. and ext. - M.: Omega-L, 2006. - 984 p. (In Russ.)].

7. Lopatin A.A., Nabiyev A.M., Silintsev V.S. Sovershenstvovaniye sistemy pokazateley dolgosrochnogo prognoza sotsial'No.-ekonomicheskogo razvitiya regiona // Ekonomika. Finansy. Rynok. 2005. № 1. URL: http://www.publications.csu.ru/form5.asp?ID=32 [**Lopatin A.A.**, **Nabiyev A.M.**, **Silintsev V.S.** Improving the system of indicators of long-term social-and-economic development of the region // Economy. Finance. Market. 2005. No. 1. URL: http://www.publications.csu.ru/form5.asp?ID=32 (In Russ.)].

8. Pulyayevskaya V.L. Valovoy munitsipal'nyy produkt kak pokazatel' otsenki ekonomicheskogo potentsiala rayonov i gorodov // Vestnik NGUEU. 2012. № 3. S. 159-168 [**Pulyaevskaya V.L.** Gross municipal product as an assessment indicator for the economic potential of regions and cities // Herald NSUEM. 2012. No. 3. P. 159-168 (In Russ.)].

9. Pulyayevskaya V.L. Otsenka munitsipal'nykh obrazovaniy Respubliki Sakha (Yakutiya) // Vestnik NGUEU. 2014. № 1. S. 185-189 [**Pulyaevskaya V.L.** Assessment of municipalities of the Republic of Sakha (Yakutia) // Herald NSUEM. 2014. No. 1. P. 185-189 (In Russ.)].

10. Tret'yakova O.V. O podkhodakh k otsenke effektivnosti zdravookhraneniya // Vestnik NGUEU. 2012. № 2. S. 183-191 [**Tretyakova O.V.** On the approaches to evaluating the effectiveness of health care // Herald NSUEM. 2012. No. 2. P. 183-191 (In Russ.)].

11. Ekonomika rayonov i gorodov Respubliki Sakha (Yakutiya): Stat. sb. / Sakha(Yakutiya)stat: Yakutsk, 2013. - 186 s. [Economics of districts and towns of the Republic of Sakha (Yakutia): Stat. handbook / Sakha(Yakutia)Stat: Yakutsk, 2013. - 186 p. (In Russ.)].

СПОСОБ ВЫЯВЛЕНИЯ ОШИБОК В БОЛЬШИХ МАССИВАХ ЧИСЛОВОЙ ИНФОРМАЦИИ

Г.Н. Хубаев

Предложены методы выявления ошибок в числовой информации. Отмечено, что в больших массивах числовых данных невозможно визуально выделить источники данных, содержащих недостоверную информацию, определить, достоверность каких показателей сомнительна и требует перепроверки. Показано, что «сжатие» исходной информации путем вычисления корреляционных матриц, средних значений коэффициентов корреляции, коэффициентов асимметрии и вариации упрощает выявление ошибок в больших массивах числовых данных. При наличии статистически значимого уравнения регрессии проверку качества исходного и нового массивов числовой информации можно проводить путем сравнения прогнозируемых и фактических значений отклика в очередном массиве числовых данных с последующей проверкой достоверности наблюдений с максимальными значениями остатков.

Ключевые слова: выявление ошибок, числовая информация, «сжатие» информации, анализ остатков, корреляционные матрицы, коэффициенты вариации и асимметрии.

JEL: C1, C4.

Постановка задачи

Пусть n - количество строк в таблице числовых исходных данных (количество объектов любой природы), а m - количество столбцов (число показателей, характеризующих каждый объект, или число моментов фиксирования значений конкретного показателя). Тогда уже при n, $m \geqslant 100$ количество представленных в таблице чисел превысит 10 тыс. Но зачастую в реальности приходится сталкиваться со случаями, когда n, $m >> 100$. И разве можно при таких объемах числовой информации *визуально* обнаружить ошибки в исходных данных, выявить наблюдения, достоверность которых сомнительна?

Известно, что при составлении рейтинга субъектов РФ за 2012 г. РИА Новости, например, использовало «данные публикуемой официальной статистики. Рейтинг строился на основе комплексного учета различных показателей, фиксирующих фактическое состояние тех или иных аспектов условий жизни, а также оценок удовлетворенности населения складывающейся в регионах ситуацией в различных социальных сферах. Источники информации для составления рейтинга: данные Росстата, Минздрава России, Минрегиона России, Минфина России, Минприроды России, Банка России, сайты региональных органов власти, другие открытые источники. При составлении рейтинга были отобраны 64 показа-

Хубаев Георгий Николаевич (gkhubaev@mail.ru) - канд. техн. наук, профессор кафедры «Информационные системы и технологии» Ростовского государственного экономического университета (РИНХ).

теля, которые объединены в 11 групп, характеризующих *основные аспекты, влияющие на качество жизни в регионе* [1]. Таким образом, получается, что, во-первых, при таком количестве исходных данных невозможно визуально выделить *источники данных*, *представившие сомнительную, недостоверную информацию*. Во-вторых, нельзя определить, *достоверность каких показателей качества жизни населения сомнительна и требует перепроверки*.

Как отмечается в Предисловии Ю.П. Адлера и В.Г. Горского в [2], к исходной информации «часто примешиваются чужеродные элементы, даже в малых количествах существенно ухудшающие ситуацию. Опыт показывает, что в больших массивах данных появление засорений практически неизбежно (подчеркнуто нами - *Г.Х.*). Долгое время разрабатывались методы выявления подозрительных наблюдений, которые называют “дикими”, или сорными. Однако чтобы их выявить, надо знать закон распределения».

Очевидно, что разработка методов выявления недостоверных данных в исходной статистической информации является исключительно актуальной задачей для всех уровней статистического обеспечения социально-экономических расчетов. В то же время в большинстве информационных бюллетеней, издаваемых учреждениями государственной статистики, не указано, какова достоверность результатов статистических расчетов, как проводилась оценка качества входной информации, как осуществлялось выявление ошибок в исходных данных или выделение наблюдений, достоверность которых вызывает сомнение. Ведь при наличии недостоверных наблюдений результаты выполняемых расчетов также окажутся недостоверными.

Ниже предлагается один из возможных способов выявления подозрительных наблюдений, ориентированный на использование статистических процедур ***«сжатия»*** больших массивов числовой информации, содержащей десятки тысяч значений показателей.

Методы выявления ошибок в исходной статистической информации

1. ***Анализ корреляционных матриц и средних по каждому из j (jЄm) столбцов значений коэффициентов корреляции.*** Ранее (см. [3, 4]) показано, что весьма эффективным способом выявления недостоверных наблюдений (или наблюдений, достоверность которых вызывает сомнение) в больших объемах числовых данных является «сжатие» исходной информации путем вычисления корреляционных матриц, средних значений коэффициентов корреляции. Анализ корреляционных матриц, корреляционных функций и значений коэффициентов корреляции облегчает выявление ошибок в больших объемах числовых статистических данных, представленных, например, в виде таблицы 1. Причем вычисление корреляционных матриц и средних значений коэффициентов корреляции целесообразно проводить и для *транспонированной* таблицы 1, когда строки и столбцы поменяются местами.

Таблица 1

Варианты представления исходной статистической информации

Объект наблюдения	Значения k-го показателя у n объектов в моменты времени j ($jЄm$) (Значения m показателей у n объектов в момент времени t_k)				
	1	. . .	j	. . .	m
1	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
i	. . .	. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .	. . .	. . .
n	. . .	. . .	. . .	. . .	. . .

Предположим, что нас интересует оценка наличия недостоверных наблюдений в исходной статистической информации. Для этого обратимся к примеру, описанному в [4]. Как оказалось, *только* анализ корреляционных матриц и сравнительный анализ средних значений коэффициентов корреляции позволили обнаружить ошибки, допущенные экспериментатором при проведении измерений значений показателей. Например, в таблице 2 представлен фрагмент матрицы корреляций, рассчитанной на основе реальных статистических данных. При рассмотрении значений коэффициентов корреляции обращает на себя внимание наличие значений, *существенно отличающихся от остальных*: пятый столбец и пятая строка в таблице 2. И такие отклонения значений коэффициентов были обнаружены в дни, когда измерения значений показателей проводил один и тот же исследователь. Но ведь *глядя на исходную (не «сжатую») таблицу данных эксперимента, содержащую тысячи близких по значениям чисел* (чисел одного порядка), такие отклонения *обнаружить невозможно*!

Таблица 2

Фрагмент матрицы корреляций
(коэффициенты корреляции между значениями показателя у множества объектов в разные моменты времени)

	Столбец 1	Столбец 2	Столбец 3	Столбец 4	Столбец 5	Столбец 6	Столбец 7	Столбец 8	...
Столбец 1	1,0	0,85	0,84	0,79	0,54	0,69	0,68	0,67	...
Столбец 2		1,0	0,82	0,81	0,48	0,75	0,69	0,68	...
Столбец 3			1,0	0,92	0,59	0,88	0,83	0,85	...
Столбец 4				1,0	0,53	0,89	0,87	0,86	...
Столбец 5					1,0	0,54	0,52	0,43	...
Столбец 6						1,0	0,95	0,95	...
Столбец 7							1,0	0,96	...
Столбец 8								1,0	...
...	...	...	...	...	...	...	...	...	...

Среднее значение коэффициента корреляции в пятом столбце таблицы 3 также заметно отличается от остальных (см. значения $M[r_{j.m}]$ в таблице 3).

Таблица 3

Средние значения коэффициентов корреляции (по столбцам) - $M[r_{j.m}]$

$r_{1.m}$	$r_{2.m}$	$r_{3.m}$	$r_{4.m}$	$r_{5.m}$	$r_{6.m}$	$r_{7.m}$	$r_{8.m}$	...
0,72	0,72	0,81	0,81	0,52	0,81	0,79	0,77	...

2. ***Анализ динамики значений коэффициентов асимметрии и вариации.*** При оценке качества исходной статистической информации *индикаторами наличия аномальных наблюдений* могут быть скачки значений коэффициентов асимметрии и вариации [Коэффициент асимметрии распределения (*coefficient of skewness of a distribution*) - мера асимметрии распределения $\gamma = \mu_3/\sigma^3$, где μ_3 - третий центральный момент, а σ^2 - дисперсия распределения]. Значения коэффициентов асимметрии и вариации также можно представлять в форме таблицы 3. В случаях обнаружения резких отклонений значений этих коэффициентов целесообразно осуществить проверку достоверности исходной информации *по конкретному показателю, наблюдению или по конкретной группе объектов*. Например, разница в численных значениях коэффициентов вариации и асимметрии в первой и третьей группах стран (см. [5] и таблицы 4 и 5) весьма значительна. Это обстоятельство свидетельствует о целесообразности перепроверки исходной статистической информации с целью выявления ошибочных наблюдений.

Таблица 4

Значения коэффициентов вариации показателей качества жизни населения по группам выделенных стран
(в процентах)

Группы стран	Наименование показателя качества жизни населения страны				
	Индекс уровня жизни	Индекс доступности образования в учебные заведения всех ступеней	Индекс прогнозируемой продолжительности жизни	Индекс образования	Индекс ВВП
Страны с высоким уровнем жизни	6,06	12,3	6,08	6,44	9,60
...	...	...	...	...	...
Страны с низким уровнем жизни	13,35	29,05	29,40	28,27	17,44

Таблица 5

Значения асимметрии распределения показателей качества жизни населения по группам выделенных стран

Группы стран	Наименование показателя качества жизни населения страны				
	Индекс уровня жизни	Индекс доступности образования в учебные заведения всех ступеней	Индекс прогнозируемой продолжительности жизни	Индекс образования	Индекс ВВП
Страны с высоким уровнем жизни	-0,07	-0,10	-0,37	-1,24	-0,4
...	...	...	...	...	...
Страны с низким уровнем жизни	-0,31	0,017	0,20	0,07	0,01

Ориентируясь на [6, 7], можно также реализовать проверку достоверности исходных данных, попавших за пределы интервала $\mu \pm (2 \div 3)\sigma$, где μ - математическое ожидание значений анализируемого показателя.

3. ***Анализ остатков в процессе прогнозирования с использованием построенного уравнения регрессии.*** Еще одним направлением выявления подозрительных наблюдений в массивах числовых данных, направлением, ориентированным на использование «*сжатой*» информации, является анализ остатков, представляющих собой разности между экспериментальными (фактическими) и расчетными

значениями откликов. Известно (см., например, [2]), что исходные предпосылки регрессионного анализа выполняются далеко не всегда. Метод наименьших квадратов (МНК) строит оценки регрессии на основе минимизации суммы квадратов остатков. Исследования остатков e_i предполагают проверку наличия следующих предпосылок МНК: случайный характер остатков; нулевая средняя величина остатков, не зависящая от x_i; гомоскедастичность - дисперсия отклонений e_i одинакова для всех значений x_i; отсутствие автокорреляции остатков - значения остатков распределены независимо друг от друга; остатки подчиняются нормальному распределению.

Если построено уравнение регрессии, адекватно описывающее исходную статистическую информацию, то такое уравнение может быть использовано *для оценки качества вновь поступившей*, аналогичной по составу и структуре *информации*. Для выявления *подозрительных наблюдений* в этом случае тоже анализируются графики остатков, но уже графики *разностей откликов*, прогнозируемых по ранее построенному уравнению регрессии *на основе новых значений* независимых переменных и фактических значений *откликов в очередном массиве* данных. Затем выполняется проверка достоверности наблюдений *с максимальными значениями* остатков.

Например, по данным из [8, 9] за 2011 г. (110 наблюдений-стран) нами построена статистически значимая регрессионная модель, позволяющая с минимальной ошибкой прогнозировать значения одного из рассматриваемых показателей. Регрессионное уравнение имеет вид:

$$Y1 = 0{,}6229+0{,}4476X1+0{,}4369X2+0{,}1043X3;$$

$$R^2=0{,}903;\ R^2_{ск}=0{,}9;\ F_{кр}=330{,}97;$$

все коэффициенты при неизвестных статистически значимы:

$$\sigma_{b1\text{-}x1} = 0{,}076;\ \sigma_{b2\text{-}x2} = 0{,}07;\ \sigma_{b3\text{-}x3} = 0{,}056.$$

Как показали графики остатков и график нормального распределения, регрессионное уравнение в целом достаточно хорошо описывает исходную информацию (за 2011 г.).

По построенному уравнению регрессии выполнен прогноз значений анализируемого показателя *на очередной период времени* (на 2012 г.) и проведено *сравнение прогнозируемых и фактических значений, рассчитаны остатки*. Оказалось, что построенное по данным за 2011 г. уравнение в целом позволяет *практически адекватно* прогнозировать вновь полученную за 2012 г. информацию по 142 странам.

Однако при анализе 142 остатков (см. рис. 1-3) обнаружено несколько значений, заметно превышающих остальные (выбросов). Очевидно, что информацию именно по этим наблюдениям необходимо перепроверить в первую очередь, то есть качество поступившей от этих стран информации может вызывать сомнение. В случае если и при построении используемой регрессионной модели аномальные выбросы остатков были у тех же источников информации, то, по-видимому, следует выяснять причины появления таких отклонений.

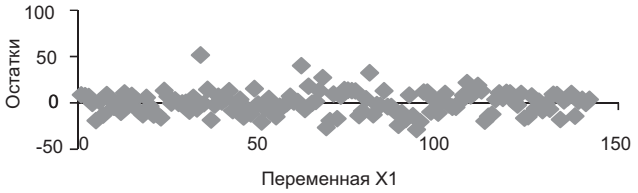

Рис. 1. **График остатков для уравнения *Y*1 в зависимости от очередных значений *X*1**

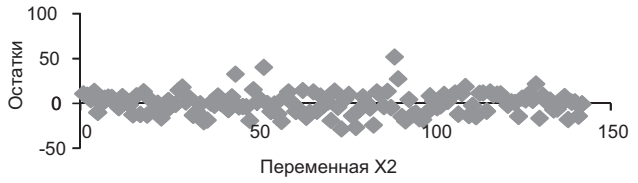

Рис. 2. **График остатков для уравнения *Y*1 в зависимости от очередных значений *X*2**

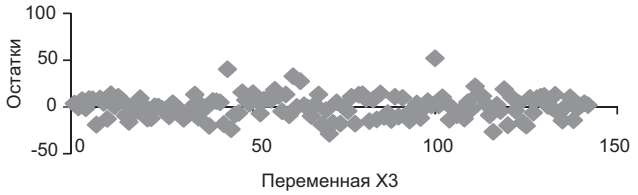

Рис. 3. **График остатков для уравнения *Y*1 в зависимости от очередных значений *X*3**

Выводы

1. Отмечено, что, *во-первых*, при больших объемах числовых данных невозможно *визуально* выделить *источники исходных данных*, представившие сомнительную, недостоверную информацию. *Во-вторых*, невозможно определить, достовер-

ность каких показателей сомнительна и требует перепроверки.

2. При наличии уравнения регрессии, адекватно описывающего исходную статистическую информацию, *оценку качества* аналогичного по составу и структуре массива *вновь поступившей* статистической информации можно проводить *путем сравнения прогнозируемых* (по построенному ранее уравнению) и *фактических* значений отклика в очередном массиве статистических данных с последующей проверкой достоверности наблюдений с максимальными значениями остатков.

3. Предложены апробированные на практике методы выявления в больших массивах числовой информации наблюдений, достоверность которых вызывает сомнение. Показано, что способ «*сжатия*» информации путем вычисления корреляционных матриц, средних значений коэффициентов корреляции, коэффициентов асимметрии и вариации облегчает выявление ошибок в больших массивах числовых данных.

Литература

1. РИА Рейтинг. URL: http://www.riarating.ru/.
2. **Дрейпер Н., Смит Г.** Прикладной регрессионный анализ: В 2-х кн. Кн. 1 / Пер. с англ. Ю.П. Адлера и В.Г. Горского. - М.: Финансы и статистика, 1986. - 366 с.
3. **Шор Я.Б., Хубаев Г.Н.** Корреляционный анализ надежности тиратронов с холодным катодом // Надежность и контроль качества. 1969. № 8. С. 29-44.
4. **Хубаев Г.Н.** Математическое моделирование на предприятии. - Ростов-на-Дону, 1973. С. 67-77.
5. РБК. Рейтинг. URL: http://rating.rbc.ru/article.shtml?2006/11/29/31275053.
6. **Шор Я.Б.** Статистические методы анализа и контроля качества и надежности. М.: Советское радио, 1962. - 553 с.
7. **Закс Л.** Статистическое оценивание. Пер. с нем. / Под ред. Ю.П. Адлера, В.Г. Горского. - М.: Статистика, 1976. - 508 с.
8. Рейтинг 2012 стран мира. URL: http://7sekretov.ru/world-ranking-2012.html.
9. The Legatum Prosperity Index™. URL: http://www.prosperity.com.

WAYS OF IDENTIFY ERRORS IN LARGE ARRAYS OF NUMERICAL INFORMATION

Georgy Khubaev

Author affiliation: Rostov State University of Economics (Rostov, Russia). E-mail: gkhubaev@mail.ru.

In this article are suggested methods for detecting errors in the numerical information. It is pointed out that in large arrays of numerical data it is impossible to visually highlight data sources with false information and to determine accuracy of which indicators is questionable and requires validation. The author shows that the «compression» of the original information by calculating the correlation matrices, average values of the correlation coefficients, coefficients of skewness and variation facilitates error detection in large arrays of numeric data. If there is a statistically significant regression equation quality check of the source and new array of numerical information can be carried out by comparing the predicted and actual values of the response in a regular array of numerical data with subsequent validation of observations with maximum values of residues.

Keywords: error detection, numerical information, «compression» of information, analysis of residuals, correlation matrix, coefficients of skewness and variation.

JEL: C1, C4.

References

1. RIA Rating. URL: http://www.riarating.ru/.
2. Dreyper N., Smit H. Prikladnoy regressionnyy analiz: V 2-kh kn. Kn. 1 / Per. s angl. Yu.P. Adlera i V.G. Gorskogo. - M.: Finansy i statistika, 1986. - 366 s. [**Draper N**., **Smith H**. Applied Regression Analysis: In 2 books. Proc. 1 / Per. Translated from English by Yu.P. Adler and V. G. Gorsky. - M.: Finance and statistics, 1986. - 366 p. (In Russ.)].
3. Shor Ya.B., Khubaev G.N. Korrelyatsionnyy analiz nadezhnosti tiratronov s kholodnym katodom // Nadezhnost' i kontrol' kachestva. 1969. № 8. S. 29-44 [**Shor Ya.B.**, **Khubaev G.N.** Correlation analysis of the reliability of the cold cathode thyratrons // Reliability and quality control. 1969. No. 8. P. 29-44 (In Russ.)].
4. Khubaev G.N. Matematicheskoye modelirovaniye na predpriyatii. - Rostov-na-Donu, 1973. S. 67-77 [**Khubaev G.N.** Mathematical modeling of the enterprise. - Rostov-on-Don, 1973. P. 67-77 (In Russ.)].
5. RBK. Rating. URL: http://rating.rbc.ru/article.shtml?2006/11/29/31275053.
6. Shor Ya.B. Statisticheskiye metody analiza i kontrolya kachestva i nadezhnosti. M.: Sovetskoye radio, 1962. - 553 s. [**Shor Ya.B.** Statistical methods for analysis and quality control and reliability. M.: Soviet Radio, 1962. - 553 p. (In Russ.)].
7. Zaks L. Statisticheskoye otsenivaniye. Per. s nem. / Pod red. Yu.P. Adlera, V.G. Gorskogo. - M.: Statistika, 1976. - 508 s. [**Sachs L**. Statistical estimation. Translated from German. / Ed. Yu.P. Adler, V.G. Gorsky. - M.: Statistics, 1976. - 508 p. (In Russ.)].
8. Reyting 2012 stran mira. URL: http://7sekretov.ru/world-ranking-2012.html [2012 Rating of countries of the world. URL: http://7sekretov.ru/world-ranking-2012.html].
9. The Legatum Prosperity Index™. URL: http://www.prosperity.com.