Рыбаков К.А., Черепанов Е.В.

Восстановление информации в таблицах эмпирических данных с применением ранговых статистик

Missing data recovery in the empirical data tables using rank values

Аннотация: В статье описана методика выявления недостающей или недостоверной информации в таблицах эмпирических данных с применением ранговых статистик. Подобная методика может оказаться полезной, а иногда и необходимой при первичной статистической обработке различных данных в информационных системах.

Abstract: A technique to identify missing or incorrect information in the empirical data tables by using rank values is described in this paper. A similar technique may be useful and necessary for a primary analysis of the various data in information systems.

Ключевые слова: эмпирические данные, восстановление информации, заполнение пропусков, порядковые статистики, ранги наблюдений.

Keywords: empirical data, missing data recovery, order statistics, rank values.

Введение. В различных эконометрических, социально-экономических и технико-экономических исследованиях, как правило, базой для анализа служат таблицы эмпирических данных (таблицы «объект – свойство»). Они часто оказываются неполными (содержат пропуски значений показателей для некоторых наблюдений) и обладают заметной недостоверностью (часть данных неточна – случайные ошибки, ложные сведения). В этой связи проблема выявления недостающей и недостоверной информации в таблицах эмпирических данных может считаться неотъемлемой частью первичной статистической обработки практически во всех прикладных статистических работах.

Используемый подход выявления недостающей и ложной информации основан на том, что, во-первых, числовые показатели, как правило, коррелированны, и, во-вторых, наблюдения в таблице обладают мерами подобия. С подробным обзором методов восстановления пропущенных значений и проверки данных на достоверность можно ознакомиться в публикациях [1–4].

Целью работы является формирование методики выявления недостающей и недостоверной информации в таблицах эмпирических данных с использованием ранговых статистик и апробация этой методики на реальных данных. Предлагаемый переход к ранговым переменным является одним из наиболее радикальных методов непараметрической статистики, дающих стабильность и относительно высокую эффективность. В статье приведены основные соотношения для решения рассматриваемой задачи (переход к рангам, получение вспомогательных оценок, восстановление информации в таблице) и приведены некоторые примеры их применения (для расчета использовалось специализированное программное обеспечение «Анализ таблиц эмпирических данных» [5]).

1. Исходные данные и их начальные преобразования. Рассмотрим эмпирическую таблицу размеров $m \times n$, где m строк определены числовыми показателями изучаемой области, а n столбцов – наблюдениями этой области. Таким образом, мы располагаем исходной матрицей числовых данных $\{x_j^k, \quad j = \overline{1,m}, \quad k = \overline{1,n}\}$.

Таблица содержит неточные значения (в том числе грубые ошибки), и пропуски значений. Введем индикатор наличия информации по элементам матрицы:

$$\xi_j^k = \begin{cases} 1, & \text{если значение } x_j^k \text{ известно,} \\ 0, & \text{если значение } x_j^k \text{ отсутствует,} \end{cases} \qquad j = \overline{1,m}, \quad k = \overline{1,n}.$$

Число наблюдений, у которых известны значения j-го показателя, равно $n_j = \sum_{k=1}^{n} \xi_j^k$, $j = \overline{1,m}$. По каждому из m показателей (из известных значений) образуем вариационный ряд:

$$x_j^{(1)} < x_j^{(2)} < \ldots < x_j^{(k)} < \ldots < x_j^{(n_j)}, \quad j = \overline{1,m}. \tag{1}$$

Члены $x_j^{(k)}$ вариационного ряда (1) называются порядковыми статистиками, а номер наблюдения в вариационном ряду – рангом этого наблюдения [6, 7]. Перейдем по всем показателям к рангам r_j^k:

$$r_j^k = \begin{cases} l \quad (1 \le l \le n_j), \quad \text{если} \quad \xi_j^k = 1 \quad \text{и} \quad x_j^k \mapsto x_j^{(l)}, \\ 0, \quad \text{если} \quad \xi_j^k = 0, \end{cases} \qquad j = \overline{1,m}, \quad k = \overline{1,n},$$

а затем к центрированным и масштабированным переменным y_j^k:

$$y_j^k = \begin{cases} \dfrac{r_j^k - \overline{r}_j}{\sqrt{Dr_j}}, \quad \text{если} \quad \xi_j^k = 1, \\ 0, \quad \text{если} \quad \xi_j^k = 0, \end{cases} \qquad j = \overline{1,m}, \quad k = \overline{1,n}, \tag{2}$$

т.е. $\overline{y}_j = 0$, $Dy_j = 1$, $j = \overline{1,m}$. Для совпадающих значений в вариационном ряду (при наличии связок) $x_j^{(1)} < x_j^{(2)} < \ldots < x_j^{(k)} = x_j^{(k)} = \ldots = x_j^{(k)} < \ldots < x_j^{(n_j)}$ ранги одинаковы, значение k вычисляется как среднее между рангами для случая, если бы эти значения в вариационном ряду не совпадали, но занимали бы соседние позиции. Например, если совпадающие значения занимают позиции $l, \ldots, l + p - 1$ (p одинаковых значений), то ранг будет равен $\dfrac{2l + p - 1}{2}$.

В формуле (2) и далее оценка математического ожидания некоторой величины χ обозначается $\overline{\chi}$, а оценка ее дисперсии – $D\chi$. Эти оценки вычисляются с учетом известных элементов матрицы данных, причем если указан нижний индекс, то наблюдения – это элементы строки матрицы с соответствующим номером, например,

$$\overline{r}_j = \frac{1}{n_j} \sum_{k=1}^{n} r_j^k, \quad Dr_j = \frac{1}{n_j - 1} \sum_{k=1}^{n} \xi_j^k (r_j^k - \overline{r}_j)^2, \quad j = \overline{1,m}.$$

Если указан верхний индекс, то наблюдения – это элементы столбца матрицы. Отметим, что могут применяться и другие методы [6, 8, 9] оценивания математического ожидания и дисперсии (параметрические, непараметрические, робастные).

2. Вспомогательные оценки. Рассмотрим, предположительно малую, случайную величину

$$\varepsilon_i = y_i - \sum_{j=1,\ j \ne i}^{m} \alpha_{ij} y_j, \quad i = \overline{1,m}.$$

Ее дисперсия имеет вид

$$D\varepsilon_i = 1 + \sum_{j=1,\ j \ne i}^{m} \alpha_{ij}^2 - 2 \sum_{j=1,\ j \ne i}^{m} \alpha_{ij} C_{ij} + \sum_{j=1,\ j \ne i}^{m} \alpha_{ij} \sum_{p=1,\ p \ne i,\ j}^{m} \alpha_{ip} C_{jp}, \quad i = \overline{1,m}. \tag{3}$$

Учитывая, что число наблюдений, в которых одновременно известны значения показателей y_i и y_j, равно $n_{ij} = \sum_{k=1}^{n} \xi_i^k \xi_j^k$, $i, j = \overline{1,m}$, значения ковариаций $C_{ij} = \operatorname{cov}(y_i, y_j)$ определим следующим образом:

$$C_{ij} = \frac{1}{n_{ij} - 1} \sum_{k=1}^{n} y_i^k y_j^k, \quad i, j = \overline{1, m}, \quad i \neq j.$$

Коэффициенты α_{ij} ($i \neq j$) найдем из условия минимизации дисперсии (3):

$$D\varepsilon_i \to \min_{\alpha_{ij}}, \quad i = \overline{1, m},$$

которое приводит к системе линейных уравнений

$$\alpha_{ij} + \frac{1}{2} \sum_{p=1,\ p \neq i,\ j}^{m} \alpha_{ip} C_{jp} = C_{ij}, \quad i, j = \overline{1, m}, \quad i \neq j. \tag{4}$$

Оценку I-го рода (по коррелированности показателей) построим в виде

$$\hat{y}_i^k = \sum_{j=1,\ j \neq i}^{m} \alpha_{ij} y_j^k, \quad i = \overline{1, m}, \quad k = \overline{1, n}; \tag{5}$$

ее дисперсия определяется соотношением

$$D\hat{y}_i = \sum_{j=1,\ j \neq i}^{m} \left(\alpha_{ij}^2 + \alpha_{ij} \sum_{p=1,\ p \neq j, i}^{m} \alpha_{ip} C_{jp} \right), \quad i = \overline{1, m}. \tag{6}$$

Как правило, коррелированность наблюдается не только между строками матрицы данных, но и между ее столбцами [1–4]. Поэтому есть смысл использовать, помимо корреляций показателей, меры линейной схожести наблюдений.

Сделаем еще одно преобразование данных:

$$z_j^k = \begin{cases} \dfrac{y_j^k - \overline{y}^k}{\sqrt{Dy^k}}, & \text{если} \quad \xi_j^k = 1, \\ 0, & \text{если} \quad \xi_j^k = 0, \end{cases} \qquad j = \overline{1, m}, \quad k = \overline{1, n}, \tag{7}$$

т.е. $\overline{z}^k = 0$, $Dz^k = 1$, $k = \overline{1, n}$.

Значения C^{kl} – ковариации $\operatorname{cov}(z^k, z^l)$ – представляются в форме

$$C^{kl} = \frac{1}{m^{kl} - 1} \sum_{j=1}^{m} z_j^k z_j^l, \quad k, l = \overline{1, n}, \quad k \neq l,$$

где $m^{kl} = \sum_{j=1}^{m} \xi_j^k \xi_j^l$ – число признаков, которые известны одновременно в двух фиксированных наблюдениях x^k и x^l (столбцах исходной матрицы числовых данных).

По аналогии с выражением (5) оценки z_j^k построим в виде

$$\hat{z}_j^k = \sum_{l=1,\ l \neq k}^{n} \beta_{kl} z_j^l, \quad j = \overline{1, m}, \quad k = \overline{1, n}, \tag{8}$$

где коэффициенты β_{kl} ($k \neq l$) определяются подобно (4) из системы уравнений

$$\beta_{kl} + \frac{1}{2} \sum_{p=1,\ p \neq k, l}^{n} \beta_{kp} C^{pl} = C^{kl}, \quad k, l = \overline{1, n}, \quad k \neq l. \tag{9}$$

Дисперсия оценки (8) определяется соотношением

$$D\hat{z}^k = \sum_{l=1,\ l \neq k}^{n} \left(\beta_{kl}^2 + \beta_{kl} \sum_{p=1,\ p \neq k, l}^{n} \beta_{kp} C^{pl} \right), \quad k = \overline{1, n}.$$

Запишем оценку II-го рода величины $\hat{y}_i^k$ (см. (7)):

$$\tilde{y}_i^k = \hat{y}^k + \sqrt{Dy^k}\, \hat{z}_i^k, \quad i = \overline{1, m}, \quad k = \overline{1, n}; \tag{10}$$

ее дисперсия выражается формулой

$$D\tilde{y}^k = Dy^k D\hat{z}^k, \quad k=\overline{1,n}. \tag{11}$$

Остается свести оценки I-го и II-го рода к одной, наиболее точной, оценке. Используя то, что оценки (5) и (10) являются неравноточными измерениями [10, 11] одной и той же величины y_j^k, представим итоговую оценку:

$$\hat{y}_j^k = \gamma_j^k \hat{y}_j^k + (1-\gamma_j^k)\tilde{y}_j^k, \quad j=\overline{1,m}, \quad k=\overline{1,n}.$$

Из условия минимизации дисперсии этой оценки

$$D\hat{y}_j^k \to \min_{\gamma_j^k}, \quad j=\overline{1,m}, \quad k=\overline{1,n},$$

получаем

$$\gamma_j^k = \frac{1/D\hat{y}_j}{1/D\hat{y}_j + 1/D\tilde{y}^k}, \quad j=\overline{1,m}, \quad k=\overline{1,n},$$

поэтому окончательное выражение для оценки y_j^k записывается следующим образом:

$$\hat{y}_j^k = \frac{\hat{y}_j^k/D\hat{y}_j + \tilde{y}_j^k/D\tilde{y}^k}{1/D\hat{y}_j + 1/D\tilde{y}^k}, \quad j=\overline{1,m}, \quad k=\overline{1,n},$$

а ее дисперсия имеет вид среднего гармонического дисперсий оценок I-го и II-го рода [8, 10]. Это хорошо согласуется с теорией обработки неравноточных физических измерений:

$$D\hat{y}_j^k = \frac{1}{1/D\hat{y}_j + 1/D\tilde{y}^k}, \quad j=\overline{1,m}, \quad k=\overline{1,n}.$$

Легко заметить, что эта дисперсия меньше минимальной из дисперсий (6) и (11).

3. Выявление недостоверных и оценка неизвестных значений. На основании (2) запишем выражение для оценки ранга r_j^k:

$$\hat{r}_j^k = \overline{r}_j + \sqrt{Dr_j}\,\hat{y}_j^k, \quad j=\overline{1,m}, \quad k=\overline{1,n},$$

тогда гарантированная погрешность выражается в виде $\Delta_j^k = 3\sqrt{D\hat{r}_j^k} = 3\sqrt{Dr_j D\hat{y}_j^k}$. Следовательно, при условии $r_j^k \notin [\hat{r}_j^k - \Delta_j^k, \hat{r}_j^k + \Delta_j^k]$ значение r_j^k заменяется оценкой $\hat{r}_j^k$. Значение $\hat{r}_j^k$ принимается в качестве «ранга» отсутствующего наблюдения x_j^k ($\xi_j^k = 0$). Поскольку значение $\hat{r}_j^k$, как правило, не является целым, то соответствующая ему оценка x_j^k восстанавливается на основе соседних наблюдений с рангами $[\hat{r}_j^k]$ и $[\hat{r}_j^k]+1$, например, с помощью методов интерполяции [11] (далее, в примерах, применялась линейная интерполяция). Если $[\hat{r}_j^k]$ или $[\hat{r}_j^k]+1$ лежат за пределами отрезка $[1, n_j]$, то можно использовать методику экстраполирования (например, [12]) «недостающих» элементов вариационного ряда (1):

$$x_j^{([\hat{r}_j^k])} < \ldots < x_j^{(1)} < x_j^{(2)} < \ldots < x_j^{(k)} < \ldots < x_j^{(n_j)} \quad \text{или} \quad x_j^{(1)} < x_j^{(2)} < \ldots < x_j^{(k)} < \ldots < x_j^{(n_j)} < \ldots < x_j^{([\hat{r}_j^k]+1)}.$$

4. Применение методики выявления недостающей или недостоверной информации. Рассмотрим примеры анализа экономических данных [13, 14] для апробации предложенной методики.

Пример 1. Проведем восстановление недостающей информации в таблице ряда показателей социально-экономического положения федеральных округов РФ (данные 2008 г.). В исходной таблице (табл. 1) пять наблюдений были помечены как отсутствующие, в результате анализа получены следующие оценки (в скобках указано исходное значение и относительная погрешность оценивания): инвестиции в основной капитал, Центральный фед. округ – 93.929 (57.932, 62.14%); иностранные инвестиции, Сибирский фед. округ – 363.271

(364.972, 0.47%); стоимость минимального набора продуктов питания в декабре, Северо-западный фед. округ – 2095.452 (2319.8, 9.67%), Сибирский фед. округ – 2062.149 (2128.8, 3.13%); общая численность безработных, Приволжский фед. округ – 0.029 (0.033, 12.12%).

Таблица 1. Показатели социально-экономического положения федеральных округов РФ

	Центральный	Северо-западный	Южный	Приволжский	Уральский	Сибирский	Дальневосточный
Инвестиции в осн. капитал, млн. р./чел.	–	74.028	39.559	48.121	114.463	42.202	74.537
Иностранные инвестиции, тыс. дол. США/чел.	1490.933	1073.968	140.100	255.046	512.778	–	1145.549
Строит. жилых домов, тыс. кв. м./чел.	0.513	0.482	0.427	0.448	0.482	0.333	0.181
Стоим. фикс. набора потреб. товар. и усл. в дек., р.	7502.100	7455.900	6415.300	6254.500	7170.600	6563.000	9043.800
Стоим. мин. набора продуктов питания в дек., р.	2099.500	–	1989.500	1904.700	2278.000	–	2942.700
Среднемес. начисленная з/п одного работника, р.	20459.200	19113.400	11783.500	13181.600	21707.900	15395.400	21147.500
Просроч. задолж. по з/п на 1.01, млн. р./чел.	0.030	0.017	0.029	0.024	0.017	0.058	0.068
Общая числ. безработных, %/100	0.020	0.029	0.050	–	0.029	0.040	0.037

Если использовать аналогичные алгоритмы оценивания за исключением этапа перехода к рангам (фактически вместо r_j^k используются значения x_j^k [3]), то результаты будут соответственно следующими: 86.103 (48.63%), 489.474 (34.11%), 2153.747 (7.16%), 2278.045 (7.01%), 0.033 (0.00%).

Пример 2. Проанализируем недельные курсы валют по отношению к рублю в 2007 г. Табл. 2 содержит две ошибки: одно значение в десять раз больше истинного, другое – в десять раз меньше (одна из характерных ошибок в таблицах данных при вводе информации – неправильное положение десятичного разделителя); ошибочные значения выделены.

В результате применения изложенной выше методики получены следующие результаты (форма представления такая же, как и в примере 1): курс евро (8-й столбец) – 35.031 (34.837, 0.56%); курс австралийского доллара (3-й столбец) – 21.814 (22.308, 2.21%). Наряду с этим курс евро (7-й столбец) был признан недостоверным – 35.032 (34.651, 1.1%). При использовании методики [3] значение курса евро (8-й столбец) признано достоверным и получена оценка курса австралийского доллара (3-й столбец) – 13.544 (39.29%).

Таблица 2. Недельные курсы валют

	1	2	3	4	5	6	7	8	9
Доллар США	25.691	25.581	25.441	25.438	25.579	25.451	25.631	25.779	25.674
Евро	34.933	35.057	35.103	35.046	34.978	35.034	34.651	**348.37**	35.023
Австралийский доллар	22.007	22.032	**2.2308**	22.406	21.834	21.783	21.026	20.832	21.070
Японская иена	20.926	20.860	20.854	21.174	21.565	21.468	22.028	22.372	22.224

При аналогичном тесте (табл. 3) десятикратная ошибка в курсе евро (7-ой столбец вместо 8-го) не была обнаружена, но была получена оценка курса австралийского доллара (3-й столбец) – 21.833 (2.13%). Без перехода к рангам [3] обе ошибки выявлены, но погрешность оценивания слишком велика: курс евро (7-й столбец) – 161.373 (365.71%); курс австралийского доллара (3-й столбец) – 13.42 (39.84%). Если же из таблицы вычеркнуть эти данные, т.е. курс евро (7-й и 8-й столбцы) и курс австралийского доллара (3-й столбец), оба алгоритма достаточно хорошо восстанавливают недостающую информацию (что является

следствием «однородности» данных). Для методики с применением ранговых статистик соответственно получаем: 35.026 (1.08%), 34.96 (0.35%), 22.187 (0.54%); без перехода к рангам: 35.026 (1.08%), 34.984 (0.42%), 22.095 (0.95%).

Таблица 3. Недельные курсы валют

	1	2	3	4	5	6	7	8	9
Доллар США	25.691	25.581	25.441	25.438	25.579	25.451	25.631	25.779	25.674
Евро	34.933	35.057	35.103	35.046	34.978	35.034	**346.51**	34.837	35.023
Австралийский доллар	22.007	22.032	**2.2308**	22.406	21.834	21.783	21.026	20.832	21.070
Японская иена	20.926	20.860	20.854	21.174	21.565	21.468	22.028	22.372	22.224

Пример 3. Восстановим пропущенные значения в табл. 4 (данные о валовом внутреннем продукте и валовой добавленной стоимости в РФ по видам экономической деятельности в ценах 2011 г., млрд. рублей).

Таблица 4. Валовой внутренний продукт и валовая добавленная стоимость

	2002	2003	2004	2005	2006	2007	2008	2009	2010
1	10819.2	13208.2	17027.2	21609.8	26917.2	33247.5	41276.8	38786.4	44939.2
2	9570.0	11619.8	14858.8	18517.7	22977.3	28484.5	35182.7	33804.1	38682.3
3	573.8	667.4	773.4	864.2	981.3	1194.8	1486.6	1502.4	1482.2
4	29.0	59.4	61.7	55.5	58.1	61.6	62.7	80.6	81.5
5	638.4	769.8	1411.6	2064.3	2509.4	2865.5	3284.6	3007.9	4020.3
6	1634.3	1897.7	2590.9	3388.5	4116.0	5025.2	6163.9	–	6353.9
7	349.4	414.1	548.3	608.4	727.0	855.9	1034.0	1368.5	1625.5
8	513.5	703.0	847.1	989.9	1202.0	1633.9	2225.3	2104.5	2188.0
9	2192.6	2572.2	3012.2	3610.5	4673.6	5745.0	7137.7	6129.6	7064.2
10	88.0	93.9	139.9	167.8	206.7	286.3	358.0	343.9	369.1
11	978.7	1244.2	1642.4	1897.0	2247.6	2750.9	3258.3	3246.3	3760.8
12	280.3	388.0	474.1	701.2	977.2	1253.8	1537.8	1723.4	1703.3
13	1019.7	1246.7	1408.0	1828.8	2287.6	3102.8	3959.4	4098.2	4406.8
14	488.7	651.3	802.5	959.1	1189.2	1466.4	1884.4	2201.7	2392.0
15	280.0	317.9	400.1	493.2	619.3	–	970.7	1134.2	1187.9
16	321.5	375.9	472.6	564.7	765.5	950.5	1197.8	1359.9	1439.7
17	182.0	218.2	273.8	324.7	417.1	522.1	621.5	589.2	607.3
18	1415.2	1775.1	2352.1	3248.2	4090.1	4977.6	6323.8	5202.1	6491.8
19	165.9	186.6	183.7	156.1	150.2	214.5	–	219.9	–
20	1249.2	1588.5	2168.4	3092.1	3939.9	4763.0	6094.2	4982.3	6256.9

Примечание: 1 – Валовой внутр. продукт в рыночных ценах; 2 – Валовая добавл. стоимость в основных ценах; 3 – Сельское хозяйство, охота и лесное хозяйство; 4 – Рыболовство, рыбоводство; 5 – Добыча полезных ископаемых; 6 – Обрабат. производства; 7 – Производство и распределение электроэнергии, газа и воды; 8 – Строительство; 9 – Опт. и розн. торг., рем. автотранс. средств, быт. изд. и предм. личн. польз.; 10 – Гостиницы и рестораны; 11 – Транспорт и связь; 12 – Финансовая деят.; 13 – Операции с недвиж. имуществом, аренда и предоставл. услуг; 14 – Государственное управление и обеспеч. военной безопасности, соц. страх.; 15 – Образование; 16 – Здравоохр. и предоставл. соц. услуг; 17 – Предоставл. прочих коммунальных, соц. и персон. услуг; 18 – Налоги на продукты; 19 – Субсидии на продукты; 20 – Чистые налоги на продукты

Результаты применения методики выявления недостающей информации с применением ранговых статистик (форма представления аналогична примерам 1 и 2): 6, 2009 г. – 5797.149 (4913.9, 17.97%); 15, 2007 г. – 766.413 (769.9, 0.45%); 19, 2008 г. – 212.241 (229.7, 7.6%); 19, 2010 г. – 219.163 (234.9, 6.7%). Без применения ранговых статистик [3] соответственно получаем: 5192.778 (5.68%), 820.741 (6.6%), 206.924 (9.92%), 217.559 (7.38%).

Заключение. Выводы, сделанные по результатам примеров, вряд ли могут претендовать на общность, тем не менее, отметим, что методика, приведенная в пп. 1–3, в ряде случаев позволяет решить задачу выявления недостающей или недостоверной информации, оценивать пропущенные или ошибочно введенные наблюдения более точно. В то же время пе-

реход к рангам снижает вероятность выявления аномальных данных (выбросов), поскольку при переходе к ранговым статистикам теряется информация о масштабе. Если анализировать исходные данные, не переходя к рангам, то точность оценки невысока особенно в случае «однородных» данных из примера 2, поскольку грубые ошибки в таблице сильно сказываются на результате. Возможные пути устранения этих недостатков следующие: предварительный анализ данных на наличие грубых ошибок, удаление их из таблицы, применение для оценки недостоверных данных параллельно как методики с применением ранговых статистик, так и без них (в последнем случае представляется целесообразным использование робастных методов оценивания математического ожидания, дисперсии, ковариации наблюдений, которые участвуют при вычислении оценок).

Предложенная методика может быть использована на этапе начальной обработки статистических данных социально-экономического и технико-экономического характера. Для ее применения разработано программное обеспечение «Анализ таблиц эмпирических данных», позволяющее решать ряд важных задач обработки информации: выявление недостоверных данных, восстановление пропущенных значений в таблицах, прогнозирование и планирование временных рядов, классификация данных и анализ уровня [5].

Литература

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск, Изд-во ИМ СО РАН, 1999.
2. Литтл Р.Дж., Рубин Д.Б. Статистический анализ данных с пропусками. – М., Финансы и статистика, 1991.
3. Черепанов Е.В. Анализ полноты и достоверности информации в таблицах эмпирических данных // Анализ социально-экономических и политических процессов и систем. Вып. 4: Сб. науч. работ. – М., АМИ, 2007. – С. 147–153.
4. Рыбаков К.А., Черепанов Е.В. Анализ данных в эмпирических таблицах с использованием порядковых статистик // Информатика, социология, экономика, менеджмент. Вып. 7, ч. 2: Межвуз. сб. науч. тр. – М., АМИ, 2010. – С. 60–65.
5. Рыбаков К.А., Черепанов Е.В. О программном обеспечении анализа таблиц эмпирических данных // Теоретические вопросы вычислительной техники и программного обеспечения. Т. 1: Межвуз. сб. науч. тр. – М., МИРЭА, 2011. – С. 47–51.
6. Тарасенко Ф.П. Непараметрическая статистика. – Томск, Изд-во ТГУ, 1976.
7. Кендэл М. Ранговые корреляции. – М., Статистика, 1975.
8. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. – М., Финансы и статистика, 1983.
9. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. – М., Финансы и статистика, 1982.
10. Свешников А.А. Основы теории ошибок. – Л., Изд. ЛГУ, 1972.
11. Щиголев Б.М. Математическая обработка наблюдений. – М., ГИФМЛ, 1962.
12. Жеруль А.О., Черепанов Е.В. Экономическое прогнозирование на основе непараметрического экстраполирования коротких временных рядов // Анализ социально-экономических и политических процессов и систем. Вып. 3. Математические вопросы социально-экономических исследований: Сб. науч. работ. – М., АМИ, 2006. – С. 28–35.
13. Федеральная служба государственной статистики [Электронный ресурс]. – http://www.gks.ru.
14. РосБизнесКонсалтинг [Электронный ресурс]. – http://www.rbc.ru.