ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 30 МАТЕМАТИКИ Выпуск 1
2023

Т. В. Жгун (Великий Новгород, НовГУ)**. Сравнительный анализ качества статистических данных России и Японии.**

УДК 519.25 DOI https://doi.org/10.52513/08698325_2023_30_1_1

Резюме: В работе предлагается методика оценки качества статистических данных с использованием аппарата конечных разностей. Предложенная методика была применена для анализа статистических данных, характеризующих качество жизни населения субъектов Российской Федерации и префектур Японии за 2010–2019 годы.

Ключевые слова: количественный математико-статистический анализ, измерения качества данных, ошибки данных, статистические данные, метод конечных разностей.

В мировой статистической практике принята концепция качества, основанная на принципе максимального удовлетворения потребностей пользователей. Исходя из этого принципа и в соответствии с международными рекомендациями и стандартами Приказ Росстата от 07.12.2018 N 732 в качестве критериев качества статистической информации называет: востребованность; достоверность точность оценок показателей; своевременность; доступность; интерпретируемость; сопоставимость; согласованность. Из обозначенных восьми позиций только одна имеет числовые характеристики — точность оценок. Остальные характеристики в значительной степени являются субъективными и зависят от знаний экспертов, производящих оценки компонентов модели. Поэтому разработка показателей качества, позволяющих однозначно характеризовать рассматриваемую совокупность данных является актуальной проблемой.

Во всех случаях оценка «качества данных» представляет собой сравнение фактического состояния конкретного набора данных с желаемым состоянием данных (данных без дефектов). Качественны те данные, которые точно представляют конкретную систему [1-3] и, следовательно, не имеют ошибок регистрации.

Предложим для оценки ошибок регистрации в качестве характеристик качества точность и достоверность данных. Определим точность данных как совпадение характеристики набора данных с неискаженными характеристиками реального объекта (явления), достоверность данных как несовпадение характеристики набора данных с характеристиками объекта с полным отсутствием определенности, т. е. для объекта, все регистрируемые параметры абсолютно случайны.

Меры точности и достоверности определяют вычисленные по ряду наблюдений приближенные конечные разности максимального порядка. Если вместо точных значений параметра j для объекта iy_{ij} известны приближенные значения y^*_{ij}, то, соответственно, вместо точных значений конечных разностей Δ^k_{ij} — рассматриваются значения приближенных конечных разностей Δ^{*k}_{ij}. Ошибка измерений ε_{ij}. Мерой точности параметра j будет максимальная из наблюдаемых ошибок [4].

$$\varepsilon^*_j = \max_i |\varepsilon^*_{ij}|, \quad \text{где} \quad \varepsilon^*_{ij} = |\Delta^{*k}_{ij}|/2^k. \tag{1}$$

Для определения меры достоверности нужно рассмотреть поведение конечных разностей случайной величины, равномерно распределённой на заданной интервале. Мож-

но показать, что если случайные величины независимы и равномерно распределены на интервале [0, a], то математическое ожидание модуля k-й конечной разности $M(\Delta^k) \leqslant a2^k/6$. Величина вычисленного отношения математического ожидания модуля последней приближенной разности к аналогичной характеристике случайного процесса

$$\mu_j = \frac{6M(|\Delta_{ij}^{*k}|)}{a2^k} \cdot 100\% \tag{2}$$

характеризует достоверность j-го параметра входных данных. Значение введенных характеристик более 5% в наборе данных будет свидетельствовать о значительном уровне искажений и случайной компоненты в сигнале и о необходимости применять методы устранения случайных искажений — методы шумоподавления — для анализа сигнала.

Также вычисленные оценки позволяют оценить среднее качество выборки одной величиной:

$$Q = \frac{1}{n}\left(\sum_{j=1}^{n}(\varepsilon_j^2 + \mu_j^2)\right)^{1/2} \tag{3}$$

С помощью введенных метрик качества оценим статистические выборки, предоставленные государственными системами статистики Японии и России. Будем сравнивать 2 выборки: доступную японскую статистику, описывающую качество жизни населения (43 показателя) и российскую выборку статистических данных, используемую автором многократно для вычисления интегральных показателей качества жизни [5] (37 показателей) за 2010–2019 годы. Данные для корректного сравнения приведены к единому масштабу [0,100] [6]. Характеристики всей совокупности оценок, вычисленные по (1), (2) для данных двух выборок, представлены в табл. 1.

Таблица 1. Характеристики совокупности оценок для данных двух выборок.

	Оценка погрешности		Доля случайного компонента	
	Япония	Россия	Япония	Россия
Среднее значение показателя	7.8	2.9	11.6	4.1
Выборочное среднеквадратичное отклонение	8.4	2.3	13.6	5.0
Максимум	32.6	7.8	59.4	24.9
Минимум	0.2	0.1	0.1	0.2

Сравнение числовых значений позволяет говорить о более высоком качестве выборки данных, предоставленных Росстатом. Также вычисленные оценки позволяют оценить качество выборки одной величиной (3). Чем меньше значение показателя качества выборки Q, тем качественнее рассматриваемая выборка. Для рассматриваемых выборок для Японии: $Q_J = 3,280$, для России: $Q_R = 1,244$. Вычисленные суммарные оценки качества также убедительно свидетельствуют о более высоком качестве российских статистических данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Bian J., Lyu T., Loiacono A., Viramontes T. M., Lipori G. Y., Guo Y., Wu Y., Prosperi M., George T. J., Harle C. A., Shenkman E. A.* Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. — J Am Med Inform Assoc. 2020 Dec; 1999–2010. Published online 2020 Nov 9.
2. *Furber C.* Data Quality. Data Quality Management with Semantic Technologies. Springer, 2015, p. 20–55.
3. *Batini C., Scannapieca M.* Data quality. Springer-Verlag, Berlin, Germany. 2006, p. 19–31.

4. *Жгун Т. В.* Оценка качества статистических данных в задаче вычисления интегральной характеристики системы по ряду наблюдений. — Современные информационные технологии и ИТ-образование, 2020, т. 16, № 2, с. 295–303. // *Zhgun T. V.* Evaluation of Statistical Data Quality in the Problem of Calculating the Integral Characteristic of a System for a Number of Observations. Sovremennye informacionnye tehnologii i IT-obrazovanie. Modern Information Technologies and IT-Education. 2020, is. 16, № 2, p. 295–303. DOI: https://doi.org/10.25559/SITITO.16.202002. 295–303
5. *Жгун Т. В.* Построение интегральной характеристики качества жизни субъектов Российской Федерации с помощью метода главных компонент. — Экономические и социальные перемены: факты, тенденции, прогноз, 2017, т. 10, № 2, с. 214–235. // *Zhgun T. V.* Building an integral measure of the quality of life of constituent entities of the Russian Federation using the principal component analysis. Economic and Social Changes: Facts, Trends, Forecast, 2017, is. 10, № 2, p. 214–235. DOI: 10.15838/esc/2017.2.50.12
6. *Zhgun T. V.* Data transformations when constructing a composite system quality index: 2021 J. Phys.: Conf. Ser. 2052 012058 Doi: 10.1088/1742-6596/2052/1/012058

UDC 519.25 DOI https://doi.org/10.52513/08698325_2023_30_1_1

Zhgun T. V. (Veliky Novgorod, Yaroslav-the-Wise Novgorod State University)**.**
Comparative analysis of the quality of statistical data from Russia and Japan.

Abstract: The proposed technique provides a formalized and computationally simple algorithm for assessing the quality of a set of input parameters of a complex dynamic system using the finite difference apparatus. The proposed methodology was applied to analyze a set of statistical data characterizing the quality of life of the population of the constituent entities of the Russian Federation and the prefectures of Japan for 2010-2019. The calculated quality scores provide convincing evidence of the higher quality of Russian statistical data.

Keywords: quantitative mathematical-statistical analysis, data quality, data errors, data quality dimensions, specific data quality metrics, finite difference method.