

DOI: 10.58168/OpEq2025_50-55

УДК 549.22 (075.8)

**ОБРАБОТКА МАССИВОВ СТАТИСТИЧЕСКИХ ДАННЫХ БОЛЬШОЙ
РАЗМЕРНОСТИ С ОБОСНОВАНИЕМ КЛАСТЕРИЗАЦИОННОГО
ПРАВИЛА**

**PROCESSING OF LARGE-DIMENSIONAL STATISTICAL DATA ARRAYS
WITH JUSTIFICATION OF THE CLUSTERING RULE**

Калач Андрей Владимирович

Доктор химических наук, профессор, начальник кафедры безопасности и
защиты сведений, составляющих государственную тайну ВИ ФСИН России, г.

Воронеж, Россия

Ковальчук Андрей Александрович

Соискатель, ВГТУ, г. Воронеж, Россия

Ряжских Виктор Иванович

доктор технических наук, профессор, заведующий кафедрой
прикладной математики и механики ВГТУ, г. Воронеж, Россия

Kalach Andrey Vladimirovich

Doctor of Chemical Sciences, Professor, head of the Department
Voronezh Institute of the Federal Penitentiary Service of Russia

Voronezh, Russia

Kovalchuk Andrew Alexsandrovich

the applicant, Voronezh State Technical University, Voronezh, Russia

Ryazhskikh Viktor Ivanovich

D.S. of Technical, Professor, Head of the Department of Applied Mathematics and
Mechanics, Voronezh State Technical University, Voronezh, Russia

Аннотация. Проведен анализ снижения размерности исходных статистических данных с применением кластерного правила с группировкой независимых параметров для обеспечения расположения их среднего значения в группах по возрастанию. Это дало возможность сформулировать задачу отыскания коэффициентов линейного уравнения пространственной регрессии, которая решена путем проецирования значений зависимой величины на координатные плоскости с применением метода наименьших квадратов и целевой функции в виде суперпозиций квадратов отклонений. Проецирование одного из регрессионных уравнений формата 2-D на плоскость, параллельную оси значений зависимой переменной, позволило идентифицировать через параметрическое регрессионное уравнение прямой в пространстве. Приведен пример расчета.

Abstract. The analysis of the reduction of the dimensionality of the initial statistical data was carried out using the cluster rule with grouping of independent parameters to ensure the arrangement of their average value in groups in ascending order. This made it possible to formulate the problem of finding the coefficients of the linear equation of spatial regression, which was solved by projecting the values of the dependent variable onto coordinate planes using the least squares method and the objective function in the form of superpositions of squared deviations. Projecting one of the regression equations of the 2-D format onto a plane parallel to the axis of the values of the dependent variable made it possible to identify a straight line in space through a parametric regression equation. An example of calculation is given.

Ключевые слова: кластеризация данных, статистические массивы, метод наименьших квадратов, пространственная регрессия.

Keywords: data clustering, statistical arrays, least squares method, spatial regression.

Необходимость построения аналитических соотношений на поле неупорядоченных данных различной предметной направленности часто

возникает в практической реализации вспомогательных формализаций при генерации феноменологических моделей [1]. Особенно эта проблема трудно формализуется в многомерном случае, когда требуемая величина должна быть идентифицирована на различных метриках. В данном случае синтезируется построение системы разнородных параметров, определяющих некую наблюдаемую величину. Выбор реперной шкалы строится по принципу возрастания, имитирующей числовую дискретную ось. Другая шкала формируется за счет упорядоченных кластерных массивов с неубывающей мощностью и инвариантной параметрической характеристикой с применением метода пассивной стратегии [2].

При таком подходе создается упорядоченное поле, на котором однозначным образом определяется значение рассматриваемого зависимого параметра с построением для него, например, регрессивного соотношения технологией минимизации интегрального квадратичного отклонения [3].

Пусть после кластеризации исходные статистические данные представлены в табличном виде:

z^j	z_1	...	z_i	...	z_n
x^j	x_1	...	x_i	...	x_n
y^j	y_1	...	y_i	...	y_n

Табл.1. Исходные данные.

Будем рассматривать декомпозицию данных на координатные (оху), (охz), (оуz), считая систему координат охуz ортогональной. Допустим линейную структуру регрессионных уравнений на соответствующих координатных плоскостях

$$y^T = a_0 + a_1x, \quad z_1^T = b_0 + b_1x, \quad z_2^T = c_0 + c_1y, \quad (1)$$

где $a_0, a_1, b_0, b_1, c_0, c_1$ - неизвестные параметры, подлежащие определению.

Составим суперпозиционную целевую функцию

$$\Phi(a_0, a_1, b_0, b_1, c_0, c_1) = \sum_{i=1}^n \left[\left(y_i^{\mathfrak{z}} - y_i^T \right)^2 + \left(z_i^{\mathfrak{z}} - z_{1i}^T \right)^2 + \left(z_i^{\mathfrak{z}} - z_{2i}^T \right)^2 \right], \quad (2)$$

которую минимизируем с помощью совокупности необходимых условий существования экстремума

$$\frac{\partial \Phi}{\partial a_0} = \frac{\partial \Phi}{\partial a_1} = \frac{\partial \Phi}{\partial b_0} = \frac{\partial \Phi}{\partial b_1} = \frac{\partial \Phi}{\partial c_0} = \frac{\partial \Phi}{\partial c_1} = 0. \quad (3)$$

Из (1) – (3) следует система линейных уравнений для определения $a_0, a_1, b_0, b_1, c_0, c_1$:

$$na_0 + \left(\sum_{i=1}^n x_i \right) a_1 = \sum_{i=1}^n y_i^{\mathfrak{z}}; \quad (4)$$

$$nb_0 + \left(\sum_{i=1}^n x_i \right) b_1 = \sum_{i=1}^n z_i^{\mathfrak{z}}; \quad (5)$$

$$nc_0 + \left(\sum_{i=1}^n y_i \right) c_1 = \sum_{i=1}^n z_i^{\mathfrak{z}}; \quad (6)$$

$$\left(\sum_{i=1}^n x_i \right) a_0 + \left(\sum_{i=1}^n x_i^2 \right) a_1 = \sum_{i=1}^n y_i^{\mathfrak{z}} x_i; \quad (7)$$

$$\left(\sum_{i=1}^n x_i \right) b_0 + \left(\sum_{i=1}^n x_i^2 \right) b_1 = \sum_{i=1}^n z_i^{\mathfrak{z}} x_i; \quad (8)$$

$$\left(\sum_{i=1}^n y_i \right) c_0 + \left(\sum_{i=1}^n y_i^2 \right) c_1 = \sum_{i=1}^n z_i^{\mathfrak{z}} y_i. \quad (9)$$

Уравнение плоскости

$$-a_1 x + y + 0 \cdot z - a_0 = 0, \quad (10)$$

которое параллельно оси oz пересекается с плоскостью, например,

$$-b_1 x + 0 \cdot y + z - b_0 = 0, \quad (11)$$

по прямой, у которой направляющий вектор определяется из соотношения (см. (10) и (11))

$$\begin{vmatrix} \bar{i} & \bar{j} & \bar{k} \\ -a_1 & 1 & 0 \\ -b_1 & 0 & 1 \end{vmatrix} = p_1 \bar{i} + p_2 \bar{j} + p_3 \bar{k}, \quad (12)$$

где $\bar{i}, \bar{j}, \bar{k}$ орты ортогональной системы координат $oxyz$. Если $(\bullet)M(x_0, y_0, z_0)$ принадлежит этой прямой, то регрессионное уравнение будет иметь следующий вид:

$$\frac{x - x_0}{p_1} = \frac{y - y_0}{p_2} = \frac{z - z_0}{p_3}. \quad (13)$$

Введем параметр t следующим образом

$$\frac{x - x_0}{p_1} = t, \quad \frac{y - y_0}{p_2} = t, \quad \frac{z - z_0}{p_3} = t. \quad (14)$$

Из первых двух соотношений (14) имеем

$$x = x_0 + p_1 t, \quad y = y_0 + p_2 t,$$

откуда

$$\frac{x}{y} = \frac{x_0 + p_1 t}{y_0 + p_2 t}$$

и поэтому

$$t = \frac{yx_0 - xy_0}{xp_2 - yp_1},$$

что в итоге дает структуру регрессионного пространственного уравнения

$$z = z_0 + \left(\frac{yx_0 - xy_0}{xp_2 - yp_1} \right) p_3. \quad (15)$$

Список литературы

1. Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К. Алгоритмы: построения и анализ. – М.: ООО «И.Д. Вильямс», 2013. – 1328с.
2. Соловьев С.В. Методы оптимизации. Примеры и задачи. – Хабаровск: Изв-во Тихоокеан. гос. ун-та, 2017. – 164с.
3. Семушин И.В. Наименьшие квадраты от теории к алгоритмам. Часть I. Теория. – Ульяновск: УлГУ, 2005. – 90с.

References

1. Cormen T., Leiserson C., Rivest R., Stein K. Algorithms: constructions and analysis. – M.: ООО I.D. Williams, 2013. – 1328 p.
2. Soloviev S.V. Optimization methods. Examples and tasks. – Khabarovsk: Izv-vo Pacific. state. university, 2017. – 164s.
3. Semushin I.V. Least squares from theory to algorithms. Part I. Theory. – Ulyanovsk: UlGU, 2005. – 90 p.