V. КАРТОГРАФІЯ ТА ГЕОІНФОРМАТИКА

УДК УДК 681.518

Берсим І.О.

ІНТЕРПРЕТАЦІЯ ТА СТАТИСТИЧНА ОБРОБКА ГЕОЛОГО-ГЕОФІЗИЧНИХ ДАНИХ ЙМОВІРНІСНО-СТАТИСТИЧНИМИ МЕТОДАМИ З ВИКОРИСТАННЯМ СИСТЕМ АВТОМАТИЗОВАНОГО ПРОЕКТУВАННЯ НА ПРИКЛАДІ ФАКТОРНОГО АНАЛІЗУ У СЕРЕДОВИЩІ MATHCAD

У статті розглядаються основні статистичні функції, ймовірнісно-статистичні методи аналізу при первинній обробці геологічних даних та наводиться приклад використання середовища Mathcad в інтерпретації даних на прикладі множинної кореляції, факторного аналізу.

Актуальність теми. Зараз у процесі обробки та інтерпретації даних, великого значення набувають ймовірнісно-статистичні методи аналізу. Дані, які ми отримуємо в окремих точках спостережень, варто розглядати як випадкові події. Теорія ймовірностей та математична статистика вивчає закономірності випадкових подій у часі та просторі і прийоми їх кількісного опису. Застосовуючи методи математичної статистики проводиться аналіз даних, що дає можливість вивести оцінки характеристик випадкової величини серед яких використовуються: числові характеристики, характеристики розподілу та характеристики взаємозв'язку.

Аналіз попередніх досліджень. До основних числових характеристик відносяться: математичне сподівання, мода, медіана, дисперсія, середньоквадратичне відхилення, коефіцієнт варіації, асиметрія, ексцес [1]. Їх розрахунок можна проводити за допомогою формул, так і за допомогою функцій Excel. Ми віддаємо перевагу і рекомендуємо використовувати середовище систем комп'ютерної алгебри з класу систем автоматизованого проектування – Mathcad. Можна використовувати й інші середовища (інтегровані) швидкої розробки програмного забезпечення - Turbo Pascal, Embarcadero Delphi. Перевага Mathcad: досить зручний інтерфейс, простота виконання операцій та велика кількість вбудованих функцій, доволі прості інструменти програмування.

Якщо виміри виконані в однакових умовах, рівноточні, то центр групування результатів таких вимірів визначається *середнім арифметичним* ($\bar{x}$)

$$\bar{x} = \sum_{i=1}^{n} x_i \cdot \frac{1}{n}$$

Мода (модальне або найімовірніше значення) дискретної випадкової величини – це значення, якого вона набуває з найбільшою ймовірністю. Мода неперервної випадкової величини – це значення, при якому щільність розподілу набуває максимуму.

Медіана (серединне значення) випадкової величини ξ – квантиль її розподілу порядку 0,5.

Дисперсія – характеристика варіації (розсіювання) значень випадкової величини, яка є середнім квадратом відхилення ξ від свого математичного сподівання:

$$\mathrm{D}\xi = \mathrm{M}(\xi - \mathrm{M}\xi)^2 = \mathrm{M}\xi^2 - (\mathrm{M}\xi)^2$$

Середнє квадратичне відхилення (с.к.в.) випадкової величини ξ – це характеристика її варіації, що дорівнює квадратному кореню з дисперсії:

$$\sigma_\xi = \sqrt{\mathrm{D}\xi}$$

Коефіцієнт варіації випадкової величини ξ становить $V\xi = \sigma\xi \,/\, M\xi$ позначається Vξ, Vξ у відсотках – $\mathrm{V}\xi\,\% = \mathrm{V}\xi \cdot 100\,\%$

Асиметрія (коефіцієнт асиметрії) випадкової величини ξ:

$$A = M(\xi - M\xi)^3 / {\sigma_\xi}^3$$

Асиметрія служить характеристикою асиметричності розподілу. Додатна асиметрія вказує на праву асиметрію розподілу. Тоді значні додатні відхилення від моди більш імовірні, ніж такі самі від'ємні відхилення [2].

Ексцес (коефіцієнт ексцесу): $E = \mathrm{M}(\xi - \mathrm{M}\xi)^4 / {\sigma_\xi}^4 - 3$

Для всіх перерахованих вище функцій кількість аргументів обмежена 30.

Якщо об'єм вибірки невеликий($n \leq 30$), то за розрахованими значеннями коефіцієнтів можна судити про нормальність розподілу, в інших випадках значення можуть бути несправедливими. Параметри вибірки (будь-які), як і асиметрія та ексцес, є випадковими величинами, тож при нормальному розподілі можуть відрізнятися від нуля [1]. За дисперсіями можна судити, чи суттєво вибіркові асиметрія та ексцес відхиляються від своїх математичних очікувань, тобто від нуля.

Викладення основного матеріалу. У наш час же все частіше постає питання дослідження й оптимізації складних, неорганізованих даних. Це можливо, навіть, лише за допомогою статистичних та імовірнісних методів. Вихідною точкою таких досліджень є аналог фізичної формули – математичної моделі системи(моделі експерименту), тобто рівняння регресії. Хоча й не завжди можна знайти зручний і точний вид моделі. В загальних випадках створюють математичну модель на основі статистичного методу – регресійного аналізу.

Рівняння регресії представляє математичну форму залежності фізичної величини, що досліджується, від факторів що впливають на неї. Вибір рівняння визначається дослідником, який пропонує модель. А вибір рівняння і визначає точність співпадіння моделі з реальними даними. Регресійний аналіз зводиться до визначення на підставі експериментальних даних коефіцієнтів регресії (коефіцієнтів моделі).

Процедура оцінки значень коефіцієнтів регресії і подібності, при якій квадрат відхилення є мінімальним, зветься *методом найменших квадратів(МНК).*

Метод найменших квадратів (МНК), завдяки широкій сфері застосування, посідає виняткове місце серед методів математичної статистики. Задачею МНК є оцінка закономірностей, які спостерігаються на тлі випадкових коливань, та її використання для подальших розрахунків, зокрема, для прогнозів. Особливу роль відіграють МНК у геофізиці, визначаючи концепцію й методологію розв'язання оберненої задачі геофізики [2].

$$
\begin{aligned}
& na_o + a_1 \sum x_1 + a_2 \sum x_2 + \ldots + a_k \sum x_k = \sum y \\
& a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_2 x_1 + \ldots + a_k \sum x_k x_1 = \sum y x_1 \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
& a_0 \sum x_k + a_1 \sum x_1 x_k + a_2 \sum x_2 x_k + \ldots + a_k \sum x_k^2 = \sum y x_k
\end{aligned}
$$

{ { {

При оцінці лінійного множинного зв'язку розраховують коефіцієнт множинної кореляції. Він відображає щільність зв'язку між залежною змінною та варіаціями всіх незалежних змінних, що включені до аналізу [3]:

$$R = \sqrt{\frac{\sigma^2_{Y\,теор}}{\sigma^2_Y}},$$

де $\sigma^2_{Y\,теор} = \dfrac{\sum \left(y_{i\,теор} - \bar{y}\right)^2}{n}$ – факторна дисперсія;

$\sigma^2_Y = \dfrac{\sum \left(y_i - \bar{y}\right)^2}{n} = \overline{y^2} - \bar{y}^2$ – загальна дисперсія.

Коли оцінюється тіснота зв'язку між результативною Y та двома факторними ознаками X1, X2, то використовується множинний коефіцієнт кореляції, який змінюється в межах від 0 до 1 і є позитивною величиною – 0<R<1: R ≤ 0,3 – зв'язок практично відсутній (або не всі важливі фактори взаємозв'язку враховані, або вибрано невірну форму рівняння регресії. Необхідно переглянути змінні, що ввійшли в модель, та можливо її вид); 0,3 < R ≤ 0,5 – слабкий зв'язок; 0,5 < R ≤ 0,7 – помірний зв'язок; R > 0,7 – сильний зв'язок.

Отже, змоделюємо певну ситуацію в середовищі Mathcad: факторний аналіз (багатовимірний випадок), множинна кореляція.

Значення (певну модель) згенеруємо в Excel (рис. 1). Першу головну компоненту визначають як лінійну комбінацію компонент величини $\vec{\eta}$, що має найбільшу дисперсію, за умови, що сума квадратів коефіцієнтів дорівнює одиниці. Другу – як не корельовану з першою лінійну комбінацію, що має найбільшу дисперсію, причому сума квадратів коефіцієнтів теж дорівнює 1, і т.д.

2,107042	6,032147	16,14535	36,75014	51,17945
0,51789	6,038814	13,03759	28,494	7,91911
0,600438	5,385674	5,256963	10,40227	0,494506
2,678603	6,090687	22,68045	53,63139	20,50734
0,500047	1,713332	13,30839	30,14691	9,446989
0,041118	3,651493			

Рис. 1. Зразок даних, генерованих в Excel

Відображення спостережень в координатах показників дає можливість побачити особливості багатовимірної вибірки у зручних проекціях (рис. 2).

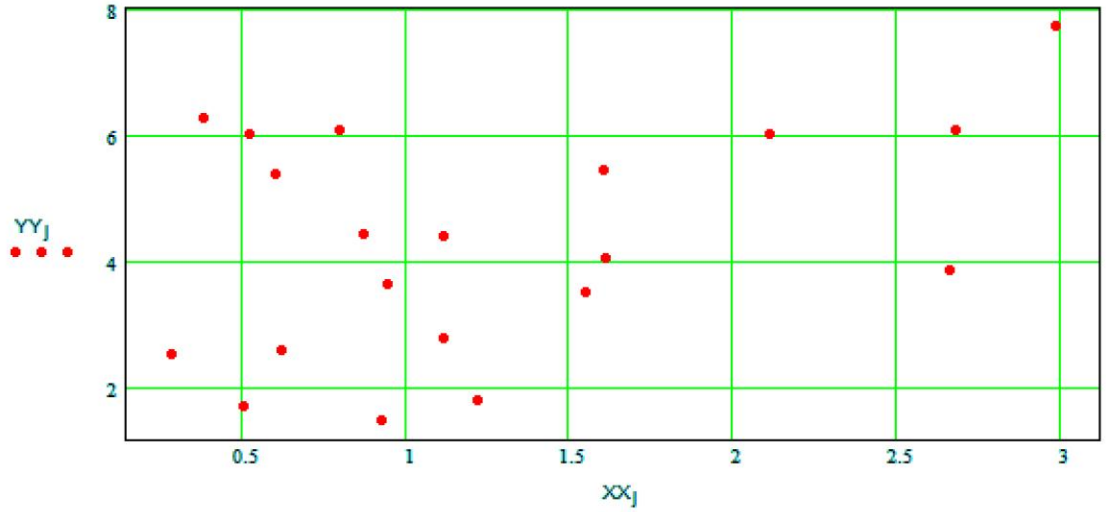

Рис. 2. Відображення даних в координатах показників

Провівши обчислення кореляційної матриці, можна судити чи буде в нашому випадку дієвим весь аналіз, чи ні. Так, ми маємо обчислити визначник матриці, значення якого близькі до нуля. В нашому випадку визначник дорівнював $1{,}6 * 10^{-3}$.

Далі слід провести процедури обчислення власних чисел кореляційної матриці, власних векторів кореляційної матриці, упорядкування власних векторів та власних чисел за спаданням власних чисел (рис. 3а). Наступним етапом є обчислення значень jj-ї головної компоненти для усіх рядків матриці X та нормування матриці спостережень X (рис. 3б).

Останнім етапом є відображення у координатах 1-ї та 2-ї головних компонент спостережень матриці X (рис. 4).

```
Upor(Be, La, M) := | for i ∈ 0.. M – 1
                   |   for j ∈ 0.. M – 2
                   |     Lt0 ← La_j
                   |     Lt1 ← La_{j+1}
                   |     Be0 ← Be^⟨j⟩
                   |     Be1 ← Be^⟨j+1⟩
                   |     La_j ← Lt1 if Lt0 < Lt1
                   |     La_{j+1} ← Lt0 if Lt0 < Lt1
                   |     Be^⟨j⟩ ← Be1 if Lt0 < Lt1
                   |     Be^⟨j+1⟩ ← Be0 if Lt0 < Lt1
                   | Be^⟨M⟩ ← La
                   | return Be
```

а)

```
ii := 0.. N – 1        jj := 0.. M – 1

Xnor_{ii, jj} := (X_{ii, jj} – Xss_{jj}) / Xsk_{jj}

Factorr(jj, X, N) := | Be ← Beta^⟨jj⟩
                     | XT ← X^T
                     | for ii ∈ 0.. N – 1
                     |   F_{ii} ← Be · XT^⟨ii⟩
                     | return F
```

б)

Рис. 3. Процедури: а) – упорядкування власних векторів та чисел за спаданням, б) – нормування матриці спостережень

Відображення спостережень у головних компонентах – це відображення $\vec{\xi}$ у проекціях на вісі $\varsigma_1, \varsigma_2, ..., \varsigma_m$. Для прикладу розглянемо випадок двовимірної величини. Нехай коефіцієнт кореляції між величинами ξ_1 і ξ_2 дорівнює r. Знайдемо факторні навантаження та головні компоненти.

З рівняння $\begin{vmatrix} 1-\lambda & r \\ r & 1-\lambda \end{vmatrix} = (1-\lambda)^2 - r^2 = 0$ маємо $\lambda_1 = 1+|r|,\ \lambda_2 = 1-|r|$.

Запишемо систему рівнянь та умову нормування: $\begin{cases} (1-\lambda)\beta_1 + r\beta_2 = 0, \\ r\beta_1 + (1-\lambda)\beta_2 = 0, \\ \beta_1^2 + \beta_2^2 = 1. \end{cases}$

При $\lambda_1 = 1+|r|$ отримаємо $\beta_1 = \beta_2 = 1/\sqrt{2}$; при $\lambda_2 = 1-|r|$ $\beta_1 = -\beta_2 = 1/\sqrt{2}$. Головні компоненти: $\zeta_1 = \eta_1/\sqrt{2} + \eta_2/\sqrt{2}$ $(\mathbf{D}\zeta_1 = \lambda_1)$; $\zeta_1 = \eta_1/\sqrt{2} - \eta_2/\sqrt{2}$ $(\mathbf{D}\zeta_1 = \lambda_2)$, де $\eta_j = \xi_j / \sqrt{\mathbf{D}\xi_j}$.

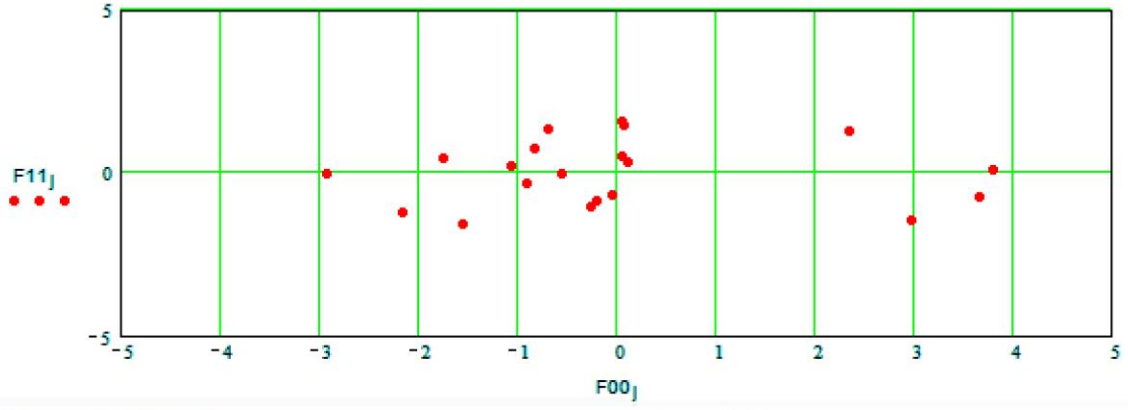

Рис. 4. Відображення у координатах 1-ї та 2-ї головних компонент спостережень матриці X на графіку

Факторний аналіз призначений для інтерпретації кореляційної матриці m-вимірної величини ξ . У результаті дістають: оцінку кількості незалежних факторів, які керують розподілом ξ ; факторні навантаження, які характеризують вплив факторів; головні компоненти, які посередньо відображають фактори через нормовані компоненти $\eta = \xi / D\xi$. *Завдання кореляційного аналізу* полягають в вимірі щільності зв'язку між ознаками, визначенні невідомих причинних зв'язків і в оцінці факторів, що мають найбільший вплив на результативну ознаку. *Регресійний аналіз* має на меті встановлення форми залежності, визначення функції регресії, використання рівняння для оцінки невідомих значень залежної змінної.

Висновки. Отже, ймовірнісно-статистичні методи аналізу є досить універсальними, їх використання спрощує інтерпретацію, аналіз та обробку геологічних даних, наступне прогнозування(моделювання). Слід використовувати ці методи при навчанні студентів, та безпосередньо використовувати їх при обробці геолого-геофізичних даних.

Література

1. Штогрин Л.В., Багрій С.М. Інформатика та обробка геологічних даних: Лабораторний практикум. – Івано-Франківськ: Факел, 2008. – 110 с. 2. Жуков М.Н. Математична статистика та обробка геологічних даних. – К.: Вид-во Київ. нац. ун-ту, 2008. – 450 с. 3. Жуков М.Н. Статистичний аналіз геологічних даних. – К., 551 с. 4. Херхагер М., Партоль М. Mathcad 2000: полное руководство. – К.: «Ирина», BHV, 2000. – 414 с.

Summary

Bersym I.O. **Interpretation and Statistical Processing of Geological and Geophysical Data Probabilistic and Statistical Methods Using Computer-Aided Design for Example Factor Analysis in an Environment Mathcad.**

This article reviews the main statistical functions, probability and statistical analysis methods in the primary processing of geological data and an example of using Mathcad environment in the interpretation of data in case of multiple correlation,factor analysis.

УДК 911.37 (475)

А.С. Соколов

АНАЛИЗ И КАРТОГРАФИРОВАНИЕ ПРОСТРАНСТВЕННОЙ НЕОДНОРОДНОСТИ РАЗМЕЩЕНИЯ НАСЕЛЕНИЯ БЕЛАРУСИ

В статье рассматриваются вопросы, связанные с расчётом показателей, характеризующих степень равномерности размещения населения по территории Беларуси, а также подходы и методы картографического представления полученных данных. Определены такие показатели, как соотношение реальной и социальной плотности, поля плотности населения, потенциал поля расселения, центры тяжести населения и ряд других. Проведён анализ размещения населения по областям и всей стране в целом.