МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

МЕТОДИЧНІ ВКАЗІВКИ
ДО ПРАКТИЧНИХ ЗАНЯТЬ
„ Регресійні моделі у гідрологічних розрахунках ”

З ДИСЦИПЛІНИ
„МЕТОДИ ПРОСТОРОВОГО УЗАГАЛЬНЕННЯ ГІДРОЛОГІЧНОЇ ІНФОРМАЦІЇ”

Одеса-2007

Методичні вказівки „ Регресійні моделі у гідрологічних розрахунках ” до практичних занять з дисципліни "Методи просторового узагальнення гідрологічної інформації"./ проф. Лобода Н.С. – Одеса, ОДЕКУ, 2007. – 40 с.

Методичні вказівки призначені для магістрів IV-V курсів навчання за спеціальністю “Гідрологія та гідрохімія”.

ЗМІСТ

ЗАГАЛЬНІ ПОЛОЖЕННЯ 4
1 РЕГРЕСІЙНІ МОДЕЛІ У ГІДРОЛОГІЧНИХ РОЗРАХУНКАХ 6
1.1 Основні положення регресійного аналізу 6
1.2 Оцінка адекватності регресійної моделі за складовими дисперсії випадкової величини 11
1.3 Визначення частинних коефіцієнтів кореляції 16
2 ПРИКЛАД РОЗРАХУНКІВ 21
3 ФАКТОРНИЙ АНАЛІЗ ПРИ ВИРІШЕННІ ЗАДАЧ ГІДРОЛОГІЧНИХ РОЗРАХУНКІВ 28
3.1 Теоретичні основи факторного аналізу 28
3.2 Застосування факторного аналізу до районування за синхронністю коливань стоку 31
4 ПРИКЛАД РОЗРАХУНКІВ 33
ДОДАТОК А 38
ДОДАТОК Б 39
СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ 40

ЗАГАЛЬНІ ПОЛОЖЕННЯ

Мета та задачі

При вивченні розділу дисципліни „Методи просторового узагальнення гідрологічної інформації”, який має назву „Регресійні моделі у гідрологічних розрахунках”, студенти повинні отримати вміння та знання щодо застосування моделей множинної лінійної регресії при узагальненнях закономірностей розподілу характеристик стоку у просторі.

При вивченні розділу „Застосування факторного аналізу до районування за синхронністю коливань стоку” студенти повинні отримати вміння та знання щодо застосування Q - модифікації факторного аналізу до районування території за синхронністю коливань стоку.

Метою методичних вказівок є закріплення студентами знань при вивченні теоретичних розділів „Застосування методів множинної регресії при розрахунках та узагальненнях характеристик стоку”, „Застосування факторного аналізу до районування за синхронністю коливань стоку”.

Задача методичних вказівок – вироблення практичних навичок при застосуванні моделі множинної лінійної регресії до розрахунків стоку невивчених у гідрологічному відношенні річок.

У результаті вивчення розділу „Регресійні моделі у гідрологічних розрахунках” студенти повинні:

знати

- основні принципи побудови моделі множинної лінійної регресії;

вміти

- аналізувати результати розрахунків за моделлю лінійної множинної регресії з покроковим вибором предікторів.

У результаті вивчення розділу „Застосування факторного аналізу до районування за синхронністю коливань стоку” студенти повинні:

знати

- основні принципи факторного аналізу

вміти

- аналізувати результати розрахунків за методом факторного аналізу;
- виконувати на основі графічних побудувань районування території.

Згідно з програмою дисципліни „Методи просторового узагальнення гідрологічної інформації” на вивчення теми „Регресійні моделі у гідрологічних розрахунках” відведено 16 годин лекційного курсу та 16 годин практичних занять, а на вивчення теми „Метод факторного аналізу” – 24 години теоретичних та 20 годин практичних занять.

Теми утворюють окремі модулі, після завершення яких студенти пишуть контрольну роботу, до якої входять питання з розглянутої тематики. До модульного контролю виконання практичних робіт входять відповіді на такі питання.

Контрольні запитання до теми „Регресійні моделі у гідрологічних розрахунках”

1. Визначення коефіцієнтів рівняння множинної лінійної регресії.
2. Визначення коефіцієнту множинної кореляції.
3. Фізичний зміст поняття „частинний коефіцієнт кореляції”.

Контрольні запитання до теми „Застосування факторного аналізу до районування за синхронністю коливань стоку”

1. Записати основне рівняння, яке виражає матрицю вихідних центрованих даних через фактори.
2. Записати основне рівняння, яке виражає коваріаційну матрицю через факторні навантаження та дисперсії залишків.
3. Описати графічні побудови при районуванні території за синхронними коливаннями стоку на основі даних про перші два фактори.

За основу викладених у збірнику методичних вказівок питань взято наукові труди таких авторів, як Є.С. Вентцель „Теория вероятностей”; Є.П. Школьний, І.Д. Лоєва, Л.Д. Гончарова „Обробка та аналіз гідрометеорологічної інформації”; А.В. Рождественський, А.І. Чеботарьов «Статистические методы в гидрологии”; „Исследование и расчеты речного стока” (под ред. Быкова).

Автор виражає вдячність інженеру кафедри гідрології суші Арестовій О.В. за допомогу в оформленні методичних вказівок.

1 РЕГРЕСІЙНІ МОДЕЛІ У ГІДРОЛОГІЧНИХ РОЗРАХУНКАХ

1.1 Основні положення регресійного аналізу

Модель лінійної парної регресії описує зв'язок генеральних сукупностей залежних випадкових величин X і Y. Задача користувача полягає в тому, щоб за обмеженими даними спостережень (вибірками), зробити висновки про характер зв'язку в цілому. У загальному випадку рівняння лінійної парної регресії є рівнянням умовного математичного сподівання випадкової величини Y, залежної від випадкової величини X:

$$m_{y/x} = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x), \tag{1.1}$$

або рівняння умовного математичного сподівання випадкової величини X, залежної від Y:

$$m_{x/y} = m_x + r_{xy}\frac{\sigma_x}{\sigma_y}(y - m_y), \tag{1.2}$$

де $m_{y/x}, m_{x/y}$ - умовні математичні сподівання Y по X та X по Y, відповідно;

r_{xy} - коефіцієнт кореляції;

σ_y, σ_x - середні квадратичні відхилення випадкових величин Y та X, відповідно;

m_y, m_x - безумовні математичні сподівання випадкових величин Y та X, відповідно.

Для вибірок (рядів спостережень) рівняння (1.1) представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = ax_i + b, \tag{1.3}$$

де x_i - дискретні значення випадкової величини X;

y_i - дискретні значення випадкової величини Y;

$\tilde{y}_i$ - значення випадкової величини Y, розраховані за рівнянням регресії;

a,b - параметри рівняння.

Оцінки параметрів, які входять в рівняння лінійної парної регресії, розраховуються на основі методу найменших квадратів. Це метод обробки емпіричного матеріалу, основна вимога якого полягає в тому, щоб сума квадратів відхилень даних спостережень від лінії регресії була найменшою, тобто

$$\Delta = \sum_{i=1}^{n} \left[y_i - \tilde{y}(x_i) \right]^2 = min, \tag{1.4}$$

де n - довжина вибірки.

Відповідно до методу найменших квадратів a та b повинні бути такими, щоб сума Δ досягала свого мінімуму. Вимога екстремуму означає, що частинні похідні від Δ, узяті по a та b, дорівнюють нулю

$$\frac{\partial \Delta(a,b)}{\partial a} = \frac{\partial \left[\left(\sum_{i=1}^{n} y_i - a x_i - b \right)^2 \right]}{\partial a} = 0 \ ; \tag{1.5}$$

$$\frac{\partial \Delta(a,b)}{\partial b} = \frac{\partial \left[\left(\sum_{i=1}^{n} y_i - a x_i - b \right)^2 \right]}{\partial b} = 0 \tag{1.6}$$

Вирішуючи рівняння (1.5) та (1.6) відносно a та b , одержуємо

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}; \tag{1.7}$$

$$b = \bar{y} - a\bar{x}, \tag{1.8}$$

де $\bar{y}, \bar{x}$ - середні арифметичні значення.

Чисельник дробу, який знаходиться в правій частині рівняння (1.7) є оцінкою коваріації (коваріаційного моменту) $\hat{K}_{xy}$, розрахованого за дискретною вибіркою завдовжки n

$$\hat{K}_{xy} = \hat{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \tag{1.9}$$

а знаменник – оцінкою дисперсії випадкової величини X

$$\hat{\sigma}_x^2 = S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \tag{1.10}$$

де S_x - оцінка середнього квадратичного відхилення σ_x випадкової величини X.

Оцінка коефіцієнта кореляції, який відображає тісноту лінійного зв'язку між рядами спостережень, які представляють собою спостережені сукупності випадкових величин Y та X, записується у вигляді

$$\hat{r}_{xy} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \tag{1.11}$$

Оцінка параметра a рівняння лінійної парної регресії виражається через коефіцієнт кореляції і середнє квадратичне відхилення випадкових величин Y та X, розрахованих за даними спостережень і позначеними як і S_y та S_x

$$a = r\frac{S_y}{S_x}. \tag{1.12}$$

Математична модель множинної лінійної регресії представляється рівнянням виду

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2) + b_3(x_{3i} - \bar{x}_3)\ldots + b_k(x_{ki} - \bar{x}_k) \tag{1.13}$$

де $\tilde{y}_i - \bar{y}$ - центровані значення залежної величини (предіктанта);

$x_{ji} - \bar{x}_j$ - центровані значення j - того аргументу (предіктора);

$b_1, b_2, b_3, ... b_k$ - коефіцієнти рівняння множинної лінійної регресії;

k - число предікторов.

Ідентифікація структури і параметрів рівняння множинної лінійної регресії виконується, виходячи з принципу найменших квадратів за (1.4), як і для випадку парної лінійної регресії. Результуючі формули для розрахунку коефіцієнтів рівняння множинної лінійної регресії за даними спостережень мають вид

$$b_j = \frac{\sigma}{\sigma_j} \frac{D_{0j}}{D_{00}}, \tag{1.14}$$

де σ - оцінка середнього квадратичного відхилення досліджуваної характеристики y;

σ_j - оцінка середнього квадратичного відхилення j-того предіктора;

D_{0j} - мінор визначника розширеної матриці коефіцієнтів кореляції, у якого викреслений перший рядок і стовпець, який відповідає змінній j, вказаній в мінорі;

D_{00} - мінор визначника розширеної матриці коефіцієнтів кореляції, у якого викреслений перший рядок і перший стовпець.

Елементами початкового визначника є коефіцієнти парної кореляції між предікторами r_{ij} і коефіцієнти парної кореляції r_{oj} між предіктантом і предікторами. При цьому визначник розширеної матриці кореляцій другого порядку записується у вигляді

$$D = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{vmatrix}, \tag{1.15}$$

а мінори цього визначника записуються таким чином

$$D_{00} = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}, \tag{1.16}$$

$$D_{01} = \begin{vmatrix} r_{10} & r_{12} \\ r_{20} & r1 \end{vmatrix}, \tag{1.17}$$

$$D_{02} = \begin{vmatrix} r_{10} & 1 \\ r_{20} & r_{21} \end{vmatrix} \tag{1.18}$$

У записах виду (1.15-1.18) r_{0j} - коефіцієнт кореляції між предіктантом (0) та предіктором (j).

Рівняння лінійної множинної регресії для двох предікторів буде мати вигляд

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2), \tag{1.19}$$

де

$$b_1 = \frac{\sigma}{\sigma_1}\frac{D_{01}}{D_{00}} \tag{1.20}$$

$$b_2 = \frac{\sigma}{\sigma_2}\frac{D_{02}}{D_{00}} \tag{1.21}$$

Рівняння (1.19) може бути записане і таким чином

$$\tilde{y}_i = b_1 x_{1i} + b_2 x_{2i} + b_0, \tag{1.22}$$

де

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2, \tag{1.23}$$

причому $\bar{y}$ - середнє арифметичне значення предіктанта;

$\bar{x}_1$ та $\bar{x}_2$ - середні арифметичні значення предікторів X_1 та X_2.

Середнє квадратичне відхилення $\sigma_{\tilde{y}}$ спостережених даних від обчислених за рівнянням множинної лінійної регресії може бути визначене за наступною залежностю

$$\sigma_{\tilde{y}} = \sigma\sqrt{1-R^2}, \tag{1.24}$$

де R- коефіцієнт множинної лінійної кореляції, який обчислюється за рівнянням

$$R = \sqrt{1-\frac{D}{D_{00}}}, \tag{1.25}$$

причому D - визначник розширеної матриці коефіцієнтів кореляції.

Якщо парні коефіцієнти кореляції, які характеризують лінійний зв'язок між двома залежними випадковими величинами, змінюються від -1 до 1, то повний коефіцієнт кореляції рівняння множинної регресії змінюється від 0 до 1.

Лінійна залежність відсутня при $r = 0$ і $R = 0$. У разі функціональної залежності $R = 1{,}0$. Чим більше коефіцієнт множинної кореляції, тим більшою мірою адекватності характеризується модель множинної регресії. Оцінити міру адекватності можна і іншим шляхом, наприклад, шляхом перевірки статистичної гіпотези про те, що залишкова дисперсія (дисперсія вхідних даних, яка не описується рівнянням регресії) незначущо відрізняється від дисперсії предіктанта. Якщо така гіпотеза приймається, то прогноз (розрахунок) по моделі не відрізняється від випадкового.

Оцінка адекватності регресійної моделі за складовими дисперсії випадкової величини

За вибірковими даними повна або загальна дисперсія змінної Y може бути розрахована за формулою

$$\sigma_y^2 = \frac{\sum_{}^{n}(y_i - \bar{y})^2}{n-1}. \qquad (1.26)$$

В основі формули (1.26) лежить відхилення спостереженої величини y_i від середнього арифметичного значення (рис.1.1).

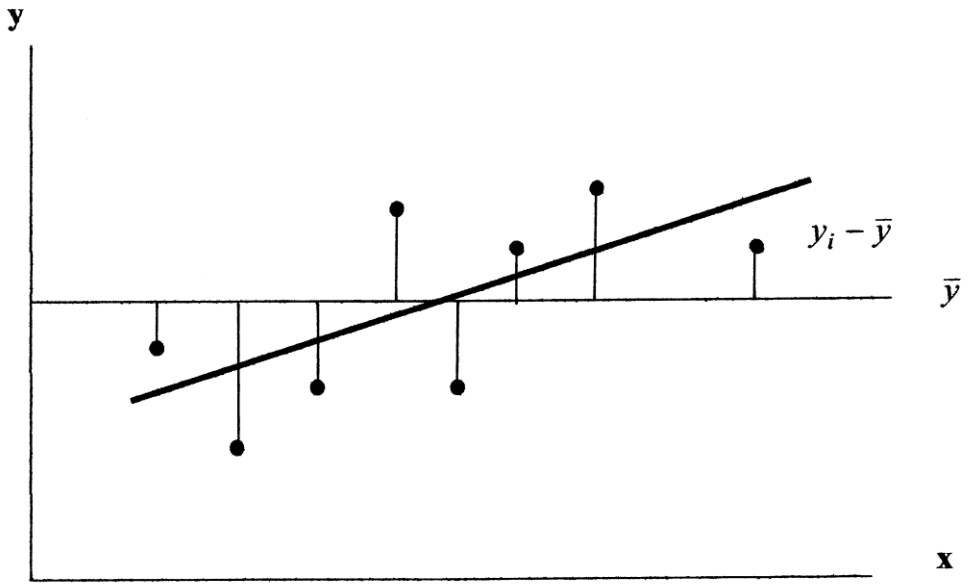

Рисунок 1.1 – Ілюстрація розсіювання спостережених значень y_i від $\bar{y}$

Запишемо $y - \bar{y}$ у вигляді складових

$$y_i - \bar{y} = (y_i - \tilde{y}_i) + (\tilde{y}_i - \bar{y}), \text{ або } y_i - \bar{y} = (\tilde{y}_i - \bar{y}) + (y_i - \tilde{y}) \qquad (1.27)$$

Тобто відхилення $(y_i - \bar{y})$ складається з відхилення значень $\tilde{y}_i$, обчислених за регресій ним рівнянням, від середнього $\bar{y}$ та з відхилення розрахованих значень $\tilde{y}_i$ від спостережених y_i.

Обидві частини рівняння возведемо у квадрат

$$(y_i - \bar{y})^2 = [(y_i - \bar{y}) + (y_i - \tilde{y}_i)]^2. \quad (1.28)$$

Після підсумовування відхилень $(y_i - \bar{y})$ отримаємо

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i) + \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (1.29)$$

Складова $2\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i)$ дорівнює нулю у випадку, коли $(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i)$ некорельовані, що справедливо для нормально розподілених величин.

Отже, можна записати, що

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-1} \quad (1.30)$$

Рівняння (1.30) представляє собою суму дисперсій

$$\sigma_y^2 = \sigma_p^2 + \sigma_{зал}^2 \quad (1.31)$$

Величина σ_p^2 має назву поясненої дисперсії, оскільки вона показує, яка частина загальної дисперсії обумовлена залежністю Y від X.

Величина $\sigma_{зал}^2$ показує ту частину дисперсії величини Y, яка не описується залежністю Y від X і має назву залишкової.

Відхилення лінії регресії $(\tilde{y}_i)$ від $\bar{y}$ графічно представлене на рисунке 1.2.

Величина $(y_i - \tilde{y}_i)$ характеризує розсіювання точок, які відповідають даним спостережень від значень, розрахованих за рівнянням лінійної регресії (рис. 1.3).

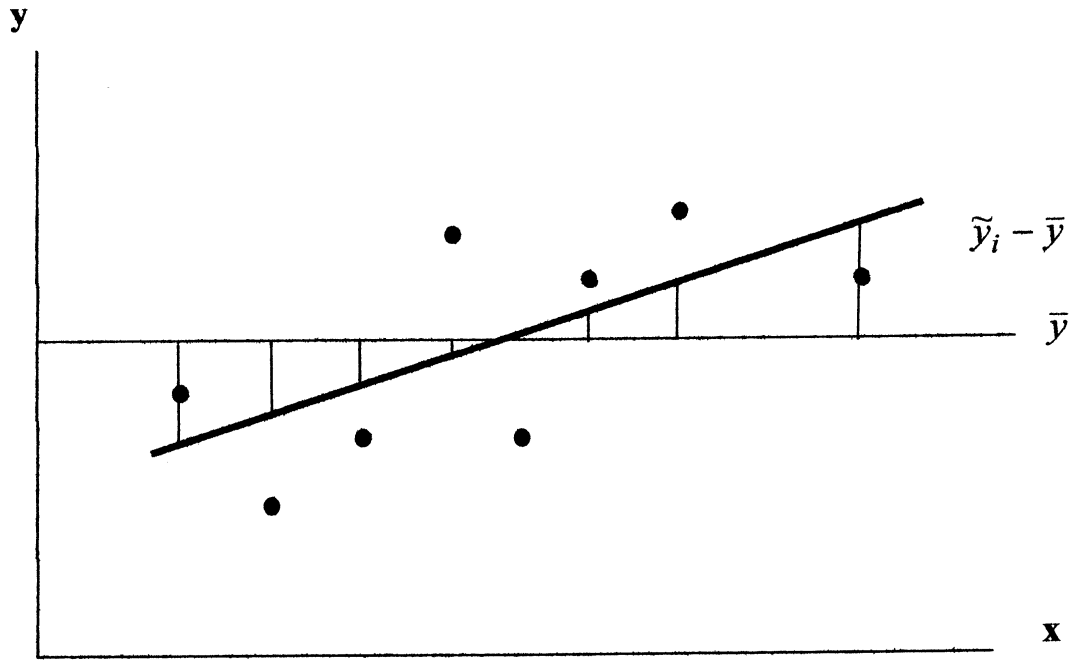

Рисунок 1.2 – Ілюстрація відхилення лінії регресії від $\bar{y}$

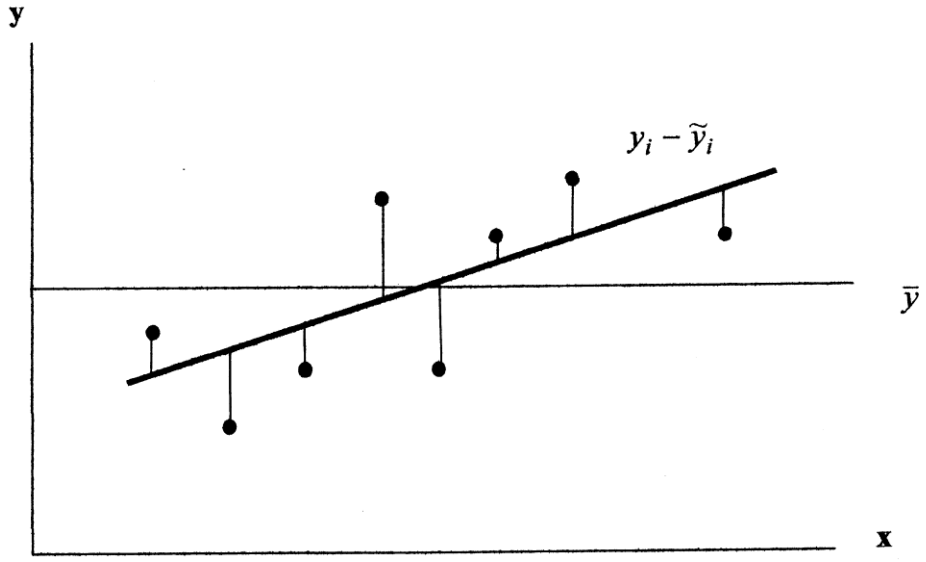

Рисунок 1.3 – Ілюстрація відхилення спостережених даних від лінії регресії.

Якщо у якості розрахункової моделі розглядається рівняння лінійної парної регресії, то гіпотеза про адекватність обраної моделі перевіряється за критерієм Фішера, який формується таким чином

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 / n-1}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n-1} = \frac{\sigma_y^2}{\sigma_{зал}^2}. \qquad (1.32)$$

Гіпотеза H_0 про те, що залишкова дисперсія незначуще відрізняється від загальної дисперсії, не відхиляється, коли

$$F < F_{кр}(\alpha, \nu_1, \nu_2), \qquad (1.33)$$

де $\nu_1 = n-1$; $\nu_2 = n-2$; α - заданий рівень значущості.

Коли ж мова йде про множинну регресію, то залишкова дисперсія - це та частина дисперсії вихідної величини Y, яка не описується залежністю предіктанта Y від предикторів $X_1, X_2, ..., X_k$. За аналогією з (1.32) перевірка гіпотези про адекватність моделі множиннї лінійної регресії даним спостережень здійснюється за критерієм Фішера, який визначається за формулою

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 /(n-1)}{\sum_{i=1}^{n-k}(y_i - \tilde{y}_i)^2 /(n-k-1)}, \qquad (1.34)$$

де n - об'єм вибірок;
k - кількість предікторів;
$y_i, \tilde{y}_i$ - фактичні і розрахункові величини.

Нульова гіпотеза не відкидається, коли $F < F_{KP}(\alpha,\upsilon_1,\upsilon_2)$, де $\upsilon_1 = n-1$, а $\upsilon_2 = n-k-1$, α - заданий рівень значущості.

Певну проблему при побудові моделі множинної лінійної регресії складає вибір оптимальних предікторів, які відображають вплив основних стокоформуючих чинників.

Збільшення числа предікторів далеко не завжди приводить до кращих результатів, оскільки при збільшенні числа предікторів збільшується порядок матриці кореляції. Більш того, серед потенційних предікторів існує багато таких, які тісно зв'язані між собою, у зв'язку з чим матриця кореляції може бути погано обумовленою.

Звичайно це приводить до значних помилок при оцінках коефіцієнтів регресії і, отже, до погіршення якості розрахункової моделі. Щоб уникнути проблем такого роду з числа потенційних предікторів вибирають ті, які є статистично значущими.

Вибір оптимальних предікторів називають операцією “просіювання”.

Просіювання може відбуватися за допомогою частинних коефіцієнтів кореляції.

1.3 Визначення частинних коефіцієнтів кореляції

Приведемо визначення частинного коефіцієнта кореляції й розглянемо алгоритм його оцінки. Припустимо, що на випадкову величину Y впливають дві випадкові величини X_1 и X_2. Частинним коефіцієнтом кореляції між випадковими величинами Y и X_1 ($r_{yx_1 \bullet x_2}$) називають коефіцієнт кореляції між ними при умові, що вплив другої випадкової величини X_2 на Y вже є врахованим. Таким же чином визначається частинний коефіцієнт кореляції $r_{yx_2 \bullet x_1}$.

Будемо вважати, що нам відома матриця кореляцій

$$R_x = \begin{vmatrix} 1 & r_{x_2} \\ r_{x_2 x_1} & 1 \end{vmatrix} \qquad (1.35)$$

і вектор парних кореляцій між Y і X_1 та Y і X_2

$$R_{yx} = \begin{vmatrix} r_{yx_1} \\ r_{yx_2} \end{vmatrix} \tag{1.36}$$

На їх основі сформуємо розширену матрицю кореляцій

$$\tilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1x_2} \\ r_{yx_2} & r_{x_1x_2} & 1 \end{vmatrix} \tag{1.37}$$

Як очевидно, вона утворюється з матриці R_x шляхом додавання до неї рядка та стовпця, що складаються з координат вектора R_{yx}. На основі матриці (1.37) розрахуємо мінори $|R_x|, D_{yx_i}, D_{\overline{yx}_i}$ $(i = 1,2)$. Мінори D_{yx_i} складаються таким чином: стовпець, на першому місці котрого розташовується парна кореляція r_{yx_i}, переставляється на перше місце, а на його місці становиться перший стовпець і, після цього, викреслюють перші рядок і стовпець. Очевидно:

$$D_{yx_1} = r_{yx_1} - r_{yx_2} r_{x_1x_2}, \tag{1.38}$$

$$D_{yx_2} = r_{yx_2} - r_{yx_1} r_{x_1x_2} \tag{1.39}$$

Означення мінора $D_{\overline{yx}_1}$ має такий сенс: це мінор визначника $|\tilde{R}|$, який не утримує парної кореляції r_{yx_1}. Очевидно ми його будемо мати, якщо викреслимо із мінора $|\tilde{R}|$ рядок і стовпець, що утримують парну кореляцію r_{yx_1}. Отже

$$D_{\overline{yx}_1} = \begin{vmatrix} 1 & r_{yx_2} \\ r_{yx_2} & 1 \end{vmatrix} = 1 - r_{yx_2}^2, \tag{1.40}$$

$$D_{\overline{yx}_2} = \begin{vmatrix} 1 & r_{yx_1} \\ r_{yx_1} & 1 \end{vmatrix} = 1 - r_{yx_1}^2 \qquad (1.41)$$

Частинні коефіцієнти кореляції визначаються таким чином:

$$r_{yx_1 \bullet x_2} = \frac{D_{yx_1}}{\sqrt{|R_x| D_{\overline{yx}_1}}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_2}^2\right)}}, \qquad (1.42)$$

$$r_{yx_2 \bullet x_1} = \frac{D_{yx_2}}{\sqrt{|R_x| D_{\overline{yx}_2}}} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} \qquad (1.43)$$

Виникає питання, яка суттєва інформація утримується в частинних коефіцієнтах кореляції? Щоб відповісти на нього, розглянемо такий приклад.

Нехай коефіцієнти парної кореляції між випадковими величинами мають значення: $r_{yx_1} = 0.72; r_{yx_2} = 0.91; r_{x_1 x_2} = 0.84$. Що можна сказати про ці випадкові величини? Звісно те, що випадкова величина Y характеризується дуже тісними кореляційними зв'язками і з величиною X_1, і з величиною X_2. Але треба звернути увагу на те, що дві останні випадкові величини теж зв'язані дуже тісним кореляційним зв'язком між собою.

Отже, щоб визначити, яка з величин X дійсно чинить вплив на величину Y, треба розрахувати частинні коефіцієнти кореляції за допомогою формул (1.42) і (1.43).

Розрахунки дають такі їх значення: $r_{yx_1 \bullet x_2} = -0.05; r_{yx_2 \bullet x_1} = 0.75$. Таким чином ясно, що в дійсності на випадкову величину Y чинить вплив випадкова величина X_2.

Кореляційний зв'язок Y з величиною X_1, якщо урахувати її зв'язок з величиною X_2, є не тільки незначним, але навіть має обернений характер.

Поширюючи отриманий алгоритм розрахунків частинних коефіцієнтів кореляції на *n* змінних, треба побудувати розширену матрицю

$$\tilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \dots & r_{yx_k} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} & \dots & r_{x_2x_n} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} & \dots & r_{x_3x_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 & \dots & r_{x_kx_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & r_{x_nx_3} & \dots & r_{x_nx_k} & \dots & 1 \end{vmatrix} \quad (1.44)$$

і на її основі визначити мінори $|R_x|, D_{yx_k}, D_{\overline{yx_k}}$ $(k = 1,2,\dots,n)$

$$r_{yx_k \bullet x_1, x_2, \dots, x_{k-1}, x_{k+1}, \dots x_m} = \frac{D_{yx_k}}{\sqrt{|R_x| D_{\overline{yx_k}}}}, \quad (1.45)$$

де D_{yx_k} - мінор розширеної матриці R_{yx}, який складається таким чином: стовпець, на першому місці в якому розташовується кореляційний коефіцієнт r_{yx_k}, переставляється на перше місце, а на його місце ставиться перший стовпець, після чого викреслюється перший рядок і перший стовпець з матриці кореляцій;

$D_{\overline{yx_k}}$ - мінор розширеної матриці R_{yx}, з якої викреслюється рядок і стовпець, що містять парний коефіцієнт кореляції r_{yx_k}, іншими словами, виключається парна кореляція r_{yx_k};

$|R_x|$ - визначник матриці, що містить тільки коефіцієнти кореляцій між предікторами.

Вибір оптимальних предікторов виконується за наступною схемою.

1. З числа предикторов вибирається той, який має найтісніший зв'язок з предіктантом і йому привласнюється номер 1.

2. Розраховується матриця частинних коефіцієнтів кореляції, за умови, що вплив першого предіктора вже враховано.

3.З числа частинних коефіцієнтів кореляції, які характеризують зв'язок між предіктантом і предикторами за умови, що вплив першого предіктора вже враховано, вибирається найбільший за абсолютною величиною (номер 2) і знов розраховується матриця частинних коефіцієнтів кореляції, але вже з урахуванням впливу перших двох предикторів.

Процедура повторюється до тих пір, поки на деякому $k+1$ етапі всі приватні коефіцієнти кореляції не втрачають статистичну значущість.

Гіпотеза про статистичну незначущість параметра перевіряється за допомогою критерію Стьюдента. Вона не спростовується, якщо $t < t_{kp}(\alpha,\upsilon)$, де $\upsilon = m - k$.

2 ПРИКЛАД РОЗРАХУНКІВ

Розглянуто 40 водозборів у басейні р. Уссурі (табл. 2.1). До розрахунків залучений пакет статистичних програм „Microstat".

Завдання: добрати оптимальні предіктори та отримати розрахункове рівняння множинної лінійної регресії для визначення середньобагаторічної величини річного стоку невивчених у гідрологічному відношенні водозборів басейну р. Уссурі.

Як потенційні предіктори розглядаються логарифм площі водозборів $lg(F+1)$; норма річних опадів $\overline{X}$; середня висота водозборів $H_{сер}$; заболоченість $f_б$; залісеність $f_л$; умовна довгота λ; умовна широта φ.

На першому етапі розраховуються коефіцієнти лінійної парної кореляції між предіктантом та усіма предікторами. Обирається предіктор, який має найбільш тісний зв'язок з предіктантом $(\overline{q})$.

У розглянутому випадку таким предіктором визнається умовна довгота $\lambda: r_{\overline{q}\lambda} = 0{,}692$.

Таблиця 2.1 – Вихідні дані застосовані для річок водозбору р. Уссурі.

N пп	$\overline{q}$, л/с·км^2	$lg(F+1)$	$\overline{X}$, мм	$H_{сер}$, м	$f_б$, %	$f_л$, %	λ, см	φ, см
88	17.2	3.44	813	679	-	100	9.3	10.1
45	8.64	3.03	823	260	19	81	4.0	5.9
46	11.4	1.93	848	340	8	92	3.2	6.3
48	8.72	2.69	764	176	12	78	3.9	8.3
81	10.6	3.83	886	670	0.5	100	8.2	6.0
83	12.3	4.27	881	685	0.5	100	8.8	7.8
85	12.1	4.36	871	601	5	95	8.1	8.7
92	10.3	3.28	823	205	17	83	6.6	9.9
93	10.3	3.67	831	373	4	96	5.5	7.1
96	11.1	3.25	977	445	-	94	6.1	7.4
97	13.8	2.63	1005	591	-	100	6.2	6.9
98	12.5	3.07	1180	568	-	100	6.1	6.2
100	9.74	3.41	914	403	2	98	5.4	6.6
105	12.4	4.12	956	790	0.5	100	12.4	11.4
107	11.4	4.33	919	560	4	93	9.9	11.8
1	9.34	3.71	924	879	-	91	3.6	3.8

Продовження таблиці 2.1

N пп	$\bar{q}$, л/с·км^2	$lg(F+1)$	$\bar{X}$, мм	$H_{сер}$, м	$f_б$, %	$f_л$, %	λ, см	φ, см
2	8.95	4.39	802	435	5	96	4.3	5.2
12	10.8	2.73	904	879	-	91	4.1	1.1
13	9.84	3.24	843	729	-	89	3.3	1.3
16	9.19	3.97	832	558	1	91	4.0	2.5
20	1.6	3.06	907	811	-	98	4.3	1.8
24	9.61	3.06	907	811	-	98	5.4	3.0
25	9.22	3.53	852	578	0.5	94	5.3	3.2
34	9.33	3.39	843	552	-	94	1.3	1.8
36	7.79	3.71	810	402	6	82	1.5	3.0
38	9.78	2.88	892	573	-	97	1.9	2.1
39	9.94	2.97	931	635	-	99	0.7	1.5
41	7.2	2.79	739	235	6	76	1.2	3.4
227	10.7	2.35	907	810	-	100	2.8	1.4
228	8.01	1.26	860	619	-	100	3.7	2.3
230	8.12	1.97	854	599	-	100	0.6	1.6
231	9.08	2.13	853	591	-	98	1.1	2.1
232	6.87	1.54	809	416	0.5	98	1.2	3.4
113	10.1	2.86	889	264	8	88	6.6	13.9
114	8.36	2.16	817	218	7	88	6.3	13.8
119	8.73	3.37	816	189	17	81	9.1	15.0
127	15.9	4.39	1048	629	0.5	94	12.2	15.7
128	13.3	3.49	1022	948	-	97	14.2	13.6
130	13.3	3.05	1040	350	-	99	9.3	13.3
131	11.3	2.70	905	216	10	81	10.2	15.7

Отримане рівняння лінійної парної регресії має наступний вигляд

$$\begin{aligned} &\bar{q} = 0.440\lambda + 8.03; \\ &\sigma_{\tilde{q}} = 1.61 \end{aligned} \qquad (2.1)$$

Підраховуються повна дисперсія предіктанта й регресійна та залишкова складові повної дисперсії

$$\sigma_{\bar{q}}^2 = \frac{\sum_{i=1}^{n}(q_i - \bar{q})^2}{n-1} = \frac{\sum_{i=1}^{40}(q_i - \bar{q})^2}{39} = \frac{189}{39} = 4.85 \qquad (2.2)$$

$$\sigma_p^2 = \frac{\sum_{i=1}^{n}(\tilde{q}_i - \bar{q})^2}{n-1} = \frac{90.4}{39} = 2.32 \qquad (2.3)$$

$$\sigma_{зал}^2 = \frac{\sum_{i=1}^{n}(q_i - \tilde{q}_i)^2}{n-1} = \frac{98.5}{39} = 2.51 \qquad (2.4)$$

Підраховується критерій Фішера за (1.34)

$$F = \frac{189/39}{98.5/38} = \frac{4.85}{2.59} = 1.87 \qquad (2.5)$$

$$\nu = 39; \nu = 38; F_{kp} = 1.7 \text{ (додаток А)} \quad F > F_{kp}$$

Отже нульова гіпотеза про те, що залишкова дисперсія незначуще відрізняється від загальної відкидається.

Надалі розраховуються частинні коефіцієнти кореляції між предіктанотом $\bar{q}$ та предікторами, які не увійшли до рівняння (2.5):

$$\begin{aligned} r_{\bar{q},lg(F+1)\bullet\lambda} &= 0.16; \\ r_{\bar{q},\bar{x}\bullet\lambda} &= 0.44; \\ r_{\bar{q},H\bullet\lambda} &= 0.40; \\ r_{\bar{q},f_\delta\bullet\lambda} &= 0.45; \\ r_{\bar{q},f_л\bullet\lambda} &= 0.46; \\ r_{\bar{q},\varphi\bullet\lambda} &= 0.26 \end{aligned} \qquad (2.6)$$

До складу оптимальних предікторів додається $f_л$ (залісеність). Цей предіктор має найбільш тісний зв'язок з $\bar{q}$, за умови, що вплив довготи урахований ($r_{\bar{q},f_л\bullet\lambda}=0.46$).

Рівняння множинної лінійної кореляції між нормою річного стоку та двома обраними оптимальними предікторами має вигляд

$$\begin{aligned}&\bar{q}=0.105f_л+0.420\lambda-1.66;\\&\sigma_{\tilde{q}}=1.44; \qquad R=0.768\end{aligned} \tag{2.7}$$

Коефіцієнт кореляції зростає від 0.692 до 0.768.

Регресійна складова, що показує, яка частина дисперсії величини $\bar{q}$ описується її залежністю від λ та $f_л$, збільшується й становить 2.94, а залишкова складова зменшується до 2.09.

Критерій Фішера приймає значення

$$F=\frac{189/39}{77.3/37}=\frac{4.85}{2.09}=2.32 \tag{2.8}$$

При $\nu_1=39$ та $\nu_2=37$ $F_{kp}=1.8$.

Оскільки $F>F_{kp}$, розрахунки по моделі будуть відрізнятися від випадкових.

Знов обчислюються частинні коефіцієнти кореляції між предіктантом $\bar{q}$ та предікторами, які не увійшли до рівняння (2.7)

$$\begin{aligned}&r_{\bar{q},lg(F+1)\bullet\lambda,f_л}=0.25;\\&r_{\bar{q},\bar{X}\bullet\lambda,f_л}=0.27;\\&r_{\bar{q},H\bullet\lambda,f_л}=0.13;\\&r_{\bar{q},f_\delta\bullet\lambda,f_л}=0.14;\\&r_{\bar{q},\varphi\bullet\lambda,f_л}=0.00\end{aligned} \tag{2.9}$$

Найбільший частинний коефіцієнт кореляції за умови, що вплив довготи та залісеності урахований, установлений між нормою річного стоку ($\bar{q}$) та нормою опадів ($\bar{X}$).

Оскільки величина частинного коефіцієнта кореляції $r_{\bar{q},\bar{x}\bullet\lambda,f_л}$ має невисоке значення , то необхідно перевірити його значущість за формулами

$$t = \frac{r_{xy}}{\sqrt{\sigma^2_{r_{xy}}}} \ , \tag{2.10}$$

$$\sigma_{r_{xy}} = \frac{1-r^2_{xy}}{\sqrt{n-1}} \ , \tag{2.11}$$

де t - критерій Стьюдента.

$$\sigma_{r_{\bar{q},\bar{x}\bullet\lambda,f_л}} = \frac{1-0.27^2}{\sqrt{40-1}} = \frac{1-0.0729}{6.24} = \frac{0.927}{6.24} = 0.149 \tag{2.12}$$

$$t = \frac{0.27}{0.149} = 0.85 \tag{2.13}$$

Критичне значення t для рівня значущості 0.05 та числа степенів свободи 40-1=39 дорівнює 3 (додаток Б).

Отже при $t < t_{kp}$ нульова гіпотеза про те, що $r_{x,y}$ не відрізняється від нуля ($H_0 : r_{xy} = 0$) не відхиляється, тобто частинний коефіцієнт кореляції $r_{\bar{q},\bar{x}\bullet\lambda,f_л}$=0.25 не розглядається як значущий і подальшого набору оптимальних предікторів не відбувається. Слід зазначити, що у більшості пакетів статистичних програм відбувається перевірка значущості всіх частинних коефіцієнтів кореляції.

Після виконання перевірки частинних коефіцієнтів кореляції на їх значущість. Як розрахункове приймається рівняння (2.7).

На основі отриманого розрахункового рівняння виконуються перевірні розрахунки, які зводяться у таблицю.

На основі результатів розрахунків обчислюється відносна похибка $\delta_i = \frac{(y_i - \tilde{y}_i)}{y_i} \cdot 100\%$ та середнє арифметичне значення її абсолютних величин.

Таблиця 2.2 – Оцінка похибки перевірних розрахунків

N	$\bar{q}_i$	$\tilde{q}_i$	$\Delta\bar{q} = \bar{q}_i - \tilde{q}_i$	$\delta_i = \frac{\Delta\bar{q}_i}{\bar{q}_i} \cdot 100\%$
1	17.20	12.76	4.44	25.80
2	8.64	8.53	0.11	1.23
3	11.40	9.35	2.04	17.80
4	8.72	8.17	0.54	6.20
5	10.60	12.29	-1.69	-15.90
6	12.30	12.55	-0.24	-1.95
7	12.10	11.72	0.37	3.05
8	10.30	9.83	0.46	4.50
9	10.30	10.74	-0.44	-4.27
10	11.10	10.78	0.31	2.88
11	13.80	11.45	2.34	16.90
12	12.50	11.41	1.08	8.64
13	9.74	10.90	-1.16	-11.90
14	12.40	14.06	-1.66	-13.38
15	11.40	12.27	-0.87	-7.63
16	9.34	9.41	-0.07	-0.75
17	8.95	10.23	-1.28	-14.30
18	10.80	9.62	1.17	10.83
19	9.84	9.08	0.76	7.72
20	9.19	9.58	-0.39	-4.24
21	11.60	10.44	1.15	9.90
22	9.61	10.90	-1.29	-13.40
23	9.22	10.44	-1.22	-13.20
24	9.33	8.76	0.56	6.00
25	7.79	7.58	0.20	2.56
26	9.78	9.33	0.44	4.59
27	9.94	9.03	0.90	9.05
28	7.20	6.83	0.36	5.13
29	10.70	10.02	0.67	6.26

Продовження таблиці 2.2

N	$\bar{q}_i$	$\tilde{q}_i$	$\Delta\bar{q} = \bar{q}_i - \tilde{q}_i$	$\delta_i = \frac{\Delta\bar{q}_i}{\bar{q}_i} \cdot 100\%$
30	8.01	10.40	-2.39	-29.80
31	8.12	9.10	-0.98	-12.06
32	9.08	9.10	-0.02	-0.22
33	6.87	9.14	-2.27	-33.04
34	10.10	10.36	-0.26	-2.57
35	8.36	10.23	-1.87	-22.36
36	8.73	10.67	-1.94	-22.22
37	15.90	13.34	2.55	16.03
38	13.30	14.50	-1.20	-9.00
39	13.30	12.65	0.64	4.86
40	11.30	11.14	0.16	1.42

$|\delta|_{сер}$=10.08

3 ФАКТОРНИЙ АНАЛІЗ ПРИ ВИРІШЕННІ ЗАДАЧ ГІДРОЛОГІЧНИХ РОЗРАХУНКІВ

3.1.Теоретичні основи методу факторного аналіза

В факторному аналізі висувається гіпотеза про те, що ***дані спостережень є лише непрямими характеристиками явища, яке вивчається, і це явище можна описати за допомогою невеликого числа деяких параметрів або властивостей. Такі теоретичні параметри або властивості називаються факторами.*** Фактори є однаковими для всіх розглядуваних гідрометеорологічних величин, але входять в кожну з них із своєю вагою. Зазначені властивості не повністю описують вихідні змінні. Залишається частина інформації, яку називають залишками. Основна перевага методу факторного аналізу полягає в тому, що безліч корельованих змінних описується набагато меншим числом факторів.

Задача факторного аналізу - представити дані спостережень у вигляді лінійних комбінацій факторів:

$$x_j = \sum_{p=1}^{k} l_{jp} f_p + \nu_j \,,\ (j = 1, m) \tag{3.1}$$

де X_j - центрована початкова змінна;

m - кількість змінних;

k - число факторів ($k << m$);

p - номер фактора;

l_{jp} - навантаження j - ої змінної на p - ий фактор або факторна вага;

f_p - некорельовані між собою фактори;

ν_j - незалежні залишки (частина даних, яка не описується кінцевим числом факторів).

Якщо у рівності (3.1) розгорнути суму, то прийдемо до системи рівнянь:

$$\begin{cases} x_1 = l_{11} f_1 + l_{12} f_2 + \ldots + l_{1k} f_k + \nu_1 \\ x_2 = l_{21} f_1 + l_{22} f_2 + \ldots + l_{2k} f_k + \nu_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_m = l_{m1} f_1 + l_{m2} f_2 + \ldots + l_{mk} f_k + \nu_m \end{cases} \tag{3.2}$$

Всі вихідні величини x_i виражаються через однакові випадкові величини f_p, але с різними ваговими коефіцієнтами.

Матрична форма рівняння (3.1) має вигляд

$$X = LF + V, \tag{3.3}$$

де X - матриця центрованих вихідних величин;
L - матриця факторних навантажень;
V - матриця незалежних залишків.
Матриця коваріацій знаходиться як

$$K = M\left|X \cdot X^{'}\right| \tag{3.4}$$

Ураховуючи (3.3), можна прийти до такого матричного виразу

$$K = LL' + D, \tag{3.5}$$

де K - матриця коваріацій;
L - матриця факторних навантажень;
L' - транспонована матриця L;
D - діагональна матриця, що складається з дісперсій незалежних залишків.

Таким чином, матрицю системи величин X можна виразити через матрицю вагових навантажень на фактори і діагональну матрицю дисперсій залишків. Фактори ураховують зв’язок між змінними, тобто вони представляють структуру кореляційної або коваріаційної матриці в термінах моделі. Проте самі фактори представляються некорельованими (ортогональними). Залишки є випадковими величинами, не зв'язаними ні між собою, ні з факторами.

Пошук факторних вагів та дисперсій залишків може відбуватися на основі методу найменьших квадратів, методу найбільшої правдоподібності, альфа-факторного аналізу, аналізу образів.

Зупинимося на методі найбільшої правдоподібності. Задача полягає у тому, щоб на основі вибіркової матриці коваріації найти ефективні умотивовані та незсунені оцінки шуканих величин, отже

$$\frac{\partial L_{П}}{\partial l_{jp}} = 0; \quad \frac{\partial L_{П}}{\partial d_j} = 0, \tag{3.6}$$

де $L_П$ - функція правдоподібності.

Результатом пошуку є наступні свідвідношення між елементами коваріаційної матриці, факторними навантаженнями й дисперсіями залишків

$$K_{jj} = \sum_{p=1}^{k} l_{jp}^2 + d_j \text{ , при } \quad j = i; \tag{3.7}$$

$$K_{ji} = \sum_{p=1}^{k} l_{jp} l_{ip} \text{ при } \quad j \neq i. \tag{3.8}$$

Оскільки у матриці коваріацій на діагоналі розташовані дисперсії змінних, то можна зробити висновок, що квадрати факторних навантажень l_{ip}^2 є частками дисперсій змінних, які описуються відповідними факторами.

Оскільки вибіркова коваріаційна матриця може бути розрахованою, то пошук факторних навантажень та дисперсій залишків відбувається шляхом ітераційного процесу.

Сума квадратів навантажень по всіх виділених факторах може бути розрахована таким чином

$$h_j^2 = \sum_{p=1}^{k} l_{jp}^2 \tag{3.9}$$

Отримана величина визначає повноту відображення j-тої змінної в усіх факторах f_p.

Повний внесок S_p (у відсотках) фактора у сумарну дисперсію змінних визначається виразом

$$S_p = \frac{\sum_{j=1}^{m} l_{pj}^2}{m} 100\%, \tag{3.10}$$

де m - кількість розглядуваних змінних.

Загальний внесок всіх виділених факторів в сумарну дисперсію досліджуваних змінних дорівнює

$$S = \sum_{p=1}^{k} S_p . \qquad (3.11)$$

3.2. Застосування методу факторного аналізу до районування за синхронністю коливань стоку

Синхронними називають коливання стоку річок, на яких спостерігається однаковий хід водності на протязі всього інтервалу часу, а асинхронними - коливання стоку, які мають протилежній хід водності. Під синфазністю і асинфазністю стоку розуміють однаковий або протилежний хід коливань не на всьому розглядуваному інтервалі часу, а по періодах водності (група багатоводних та маловодних років).

Кількісною мірою синхронності є коефіцієнт кореляції між двома рядами. Коливання вважаються синхронними, якщо коефіцієнт кореляції перевищує 0,7, й несинхронними, коли коли коефіцієнт кореляції меньший 0,4.

Виділення районів з синхронними коливаннями стоку можливе на основі матриці кореляцій, але при великій кількості рядів для аналізу структури кореляційної матриці застосовуються методи багатовимірного статистичного аналізу (факторного і головних компонент).

При аналізі синхронності коливань стоку розглядаються не зв'язки між ознаками, а зв'язки між рядами, при цьому використовується Q-техника факторного аналізу, яка може розглядаєтися як варіант класифікаційного аналізу. Внесок кожного фактора у дисперсію змінної у даному випадку представляє собою внесок у дисперсію ряду спостережень за стоком. Не зупиняючись на фізичній інтерпретації факторів, при дослідженні синхронності коливань річного стоку в практиці гідрологічних розрахунків застосовують наступні графічні побудови. У випадку, коли перших два фактора описують більше 60% загальної дисперсії вихідних даних, на графіку, осі якого являють собою два фактори, проводять вектори з початку координат у точку з координатами, відповідними факторним навантаженням. Довжина вектора розраховується за виразом

$$d_j = \sqrt{l_{j1}^2 + l_{j2}^2}\,, \tag{3.12}$$

де l_{j1} і l_{j2} - вагові коефіцієнти першого та другого факторів.

Величина d ототожнюється з h. Вона визначає повноту відображення j-го ряду спостережень першими двома факторами, а косинус кута між j-тим та і-тим вектором є коефіцієнт кореляції між ними. Таким чином, про ступінь зв'язку між рядами можна судити по угрупуваннях точок, які утворюються на площині. Як міра схожості в даному випадку використовується міра відстані: чим ближче розташовані точки на графіку і менше косинус кута між ними, тим ближче значення коефіцієнта кореляції до 1.

З метою виконання одночасного аналізу трьох ефективних факторів рекомендується представляти навантаження не в декартових, а в полярних координатах

$$\theta_j = arcsin\frac{l_{j3}}{d_j} \tag{3.13}$$

$$\lambda_j = \arcsin\frac{l_{j2}}{\sqrt{l_{j1}^2 + l_{j2}^2}}, \tag{3.14}$$

де

$$d_j = \sqrt{l_{1j}^2 + l_{2j}^2 + l_{3j}^2} \tag{3.15}$$

а θ та λ - полярні координати (в градусах або радіанах), які визначають положення перетину j-тим вектором одиничної сфери.

Величина d_j оцінює внесок усіх трьох факторів у формування стоку j-того водозбору, включеного в аналіз. Якщо розглядається не матриця кореляцій, а коваріацій, де діагональні елементи дорівнюють 1, то близькість d_j до одиниці указує на те, що дисперсія даної змінної значною мірою пояснюється першими трьома факторами.

Таким чином, застосування Q-модіфікації факторного аналізу дозволяє стиснути інформацію, яка міститься в кореляційній матриці, й інтерпретувати її. Компактне трактування кореляційної матриці досягається при розгляді кожного угрупування водозборів. При наближенні кута між векторами, спрямованими з початку координат до центрів угрупувань, до $90^\circ C$ виділені угрупування розглядаються як

райони з асинхронними коливаннями стоку. Чим меньший кут між угрупуваннями, тим тісніший зв'язок між стоком розглядуваних річок. Усередині угрупувань (районів) можна виділяти окремі групи точок, які по їх територіальному розташуванню і особливостям формування стоку можна інтерпретувати як підрайони.

4 ПРИКЛАД РОЗРАХУНКІВ

Використані дані по 19 гідрологічним постам Гірського Криму. Період сумісних спостережень становить 15 років (з 1965 по 1980 роки).

Гідрологічні пости рівномірно розташовані по території (рис. 4.1).

Перший фактор описує 43% дисперсії вихідних даних, другий – 25% і третій – 8%.

У таблиці 4.1 представлені факторні ваги, полярні координати, розраховані за (3.13) та (3.14), та коефіцієнти повноти відображення d.

Коефіцієнт d знаходиться за виразом

$$d = \sqrt{l_{j1}^2 + l_{j2}^2 + l_{j3}^2} \tag{4.1}$$

Як можна побачити із рисунків 4.2 та 4.3, розглянуті водозбори утворюють 2 угрупування, які за територіальним розташуванням можна поділити на Західний та Східний Гірський Крим. Поява районів з коливаннями стоку різної синхронності обумовлена різною зволоженістю заходу та сходу Кримського півострова.

Таблиця 4.1 – Факторні ваги та полярні координати при використанні трьох факторів

№ за картою	Річка - пост	Факторні ваги			Полярні координати		d
		l_{j_1}	l_{j_2}	l_{j_3}	θ_j	λ_j	
11а	стр. Куйбишевський – смт. Щебетівка	0.651	0.209	0.159	0.227	0.307	0.701
36	р.Су-Индол – с.Топольовка	0.841	0.206	0.069	0.080	0.240	0.867
37	р. Салгір – с. Пионерське	0.906	-0.124	-0.226	-0.243	-0.136	0.941
40	р. Кізіл -Коба – с. Червонопечерне	0.694	-0.396	-0.347	-0.410	-0.519	1.289
41	р. Ангара – с. Перевальне	0.959	-0.103	0.130	0.134	-0.107	0.972
25	р. Демерджи – м. Алушта	0.826	0.221	0.235	0.268	0.261	0.886
26	р. Куру Узень – с. Сонячногірське	0.627	0.042	0.606	0.767	0.067	0.872
27	р. Улу Узень – с. Сонячногірське	0.874	-0.001	0.039	0.044	-0.001	0.874
29	р. Арпат- м. Зеленогір’я	0.563	-0.356	0.115	0.171	-0.564	0.674
32	р.Ай- Серез – м. Межиріччя	0.197	0.811	0.140	0.166	1.333	0.844
33	р. Таракташ – м. Судак	0.255	0.651	0.523	0.643	1.198	0.872
35	руч. Кізілташ – смт. Щебетівка	0.118	0.854	0.082	0.095	1.434	0.749
39	р. Салгір – с. Двуріччя	-0.036	0.348	0.801	1.159	1.469	0.873
38а	р. Салгір – м. Сімферополь	-0.162	0.489	-0.450	-0.719	1.251	0.683
45	р. Бурульча - смт. Межгір’я	-0.103	0.766	0.077	0.097	1.438	0.775
47	р. Біюк –Карасу - м. Білогірськ	-0.111	0.600	0.549	0.732	1.388	0.820
48	р. Біюк-Карасу – с. Зибіны	-0.065	0.824	0.341	0.390	1.400	0.893
49	р.Біюк- Карасу – с. Заріччя	0.051	0.206	0.931	1.374	1.327	0.954
50	р.Кучук - Карасу – с. Богате	-0.229	0.892	0.064	0.070	1.319	0.922

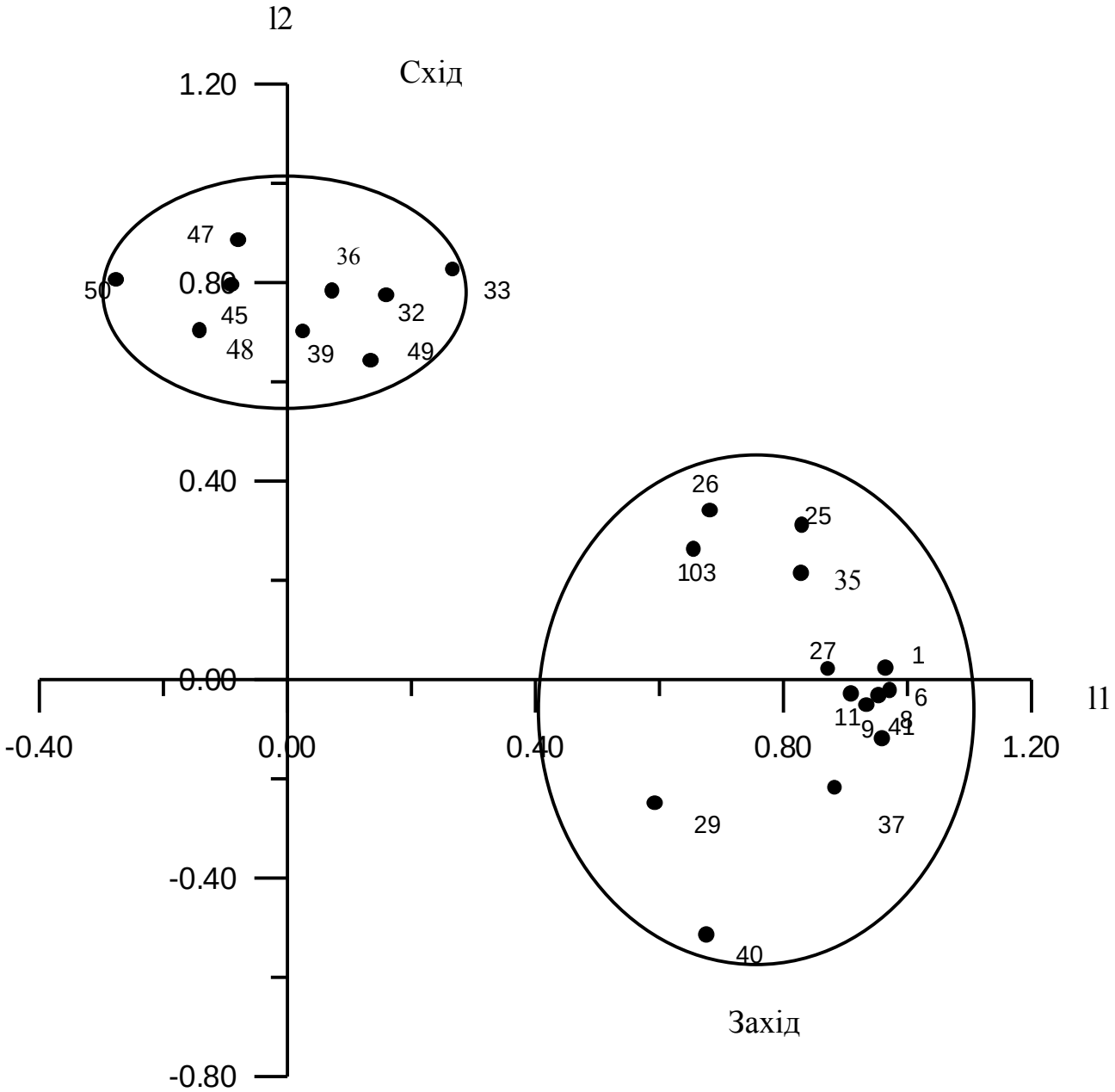

Рисунок 4.2 - Виділення угрупувань з синхронними коливаннями річного стоку на основі двох факторів

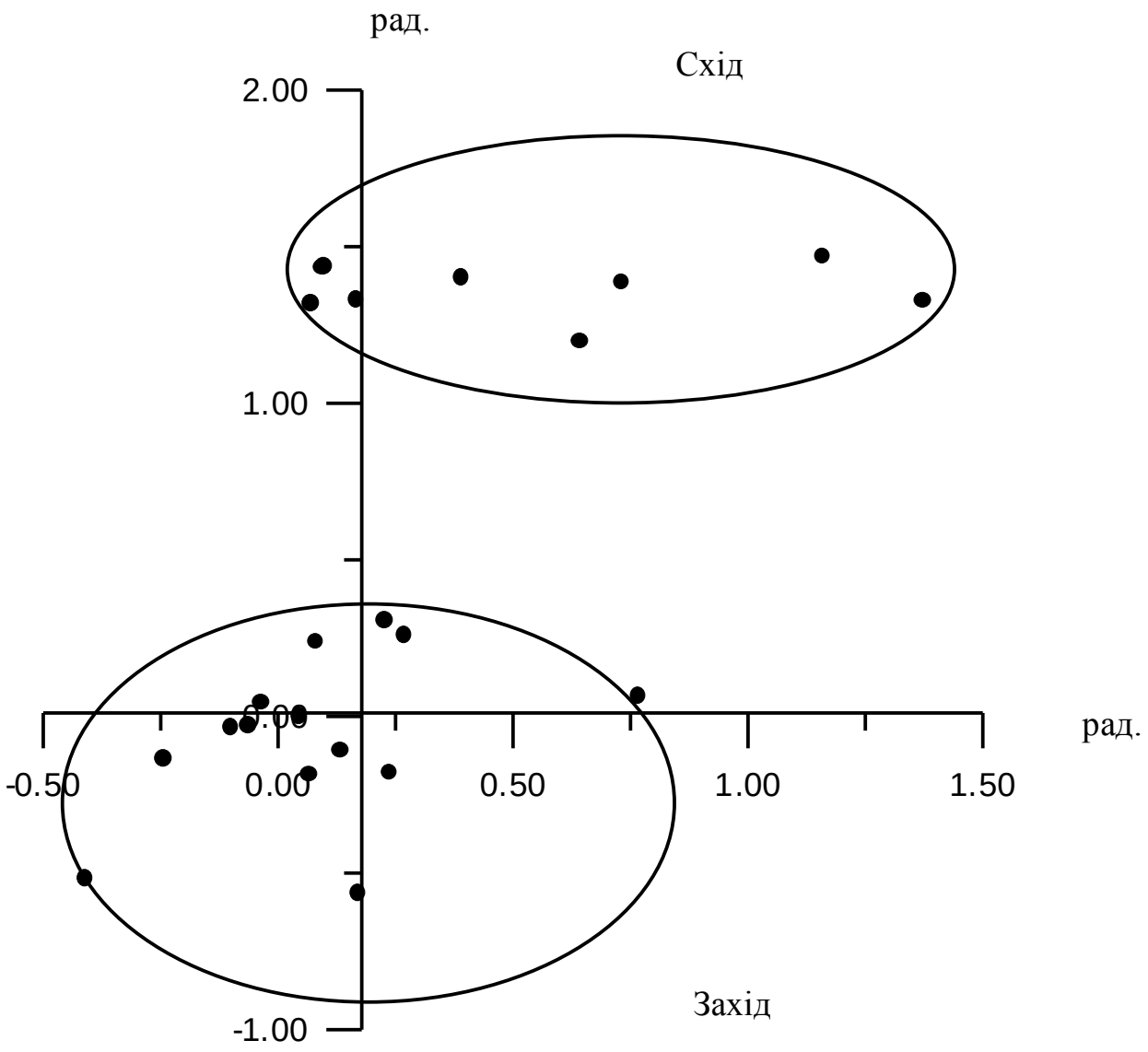

Рисунок 4.3 – Виділення угрупувань з синхронними коливаннями річного стоку Гірського Криму на основі трьох факторів

Додаток А

Значення критерію Фішера F для рівня значущості 0.05
ν_1 - число ступенів свободи для більшої дисперсії
ν_2 - число ступенів свободи для меншої дисперсії

ν_2	ν_1								
	12	14	16	20	24	30	40	50	75
10	2.9	2.9	2.8	2.8	2.7	2.7	2.7	2.6	2.6
11	2.8	2.7	2.7	2.7	2.6	2.6	2.5	2.5	2.5
12	2.7	2.6	2.6	2.5	2.5	2.5	2.4	2.4	2.4
13	2.6	2.6	2.5	2.5	2.4	2.4	2.3	2.3	2.3
14	2.5	2.5	2.4	2.4	2.4	2.3	2.3	2.2	2.2
15	2.5	2.4	2.3	2.3	2.3	2.3	2.2	2.2	2.2
16	2.4	2.4	2.3	2.3	2.2	2.2	2.2	2.1	2.1
17	2.4	2.3	2.3	2.3	2.2	2.2	2.1	2.1	2.0
18	2.3	2.3	2.3	2.3	2.2	2.1	2.1	2.0	2.0
19	2.3	2.3	2.2	2.2	2.1	2.1	2.0	2.0	2.0
20	2.3	2.2	2.2	2.1	2.1	2.0	2.0	2.0	1.9
21	2.3	2.2	2.2	2.1	2.1	2.0	2.0	1.9	1.9
22	2.2	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.9
23	2.2	2.1	2.1	2.0	2.0	2.0	1.9	1.9	1.8
24	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.9	1.8
25	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8
26	2.2	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8
27	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8
28	2.1	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8
29	2.1	2.1	2.0	1.9	1.9	1.9	1.8	1.8	1.7
30	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.8	1.7
32	2.1	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7
34	2.1	2.0	2.0	1.9	1.8	1.8	1.7	1.7	1.7
36	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.7
38	2.0	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6
40	2.0	2.0	1.9	1.8	1.8	1.7	1.7	1.7	1.6
42	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
44	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
46	2.0	1.9	1.9	1.8	1.8	1.7	1.7	1.6	1.6
48	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.6	1.6
50	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.6	1.6
100	1.9	1.8	1.8	1.7	1.6	1.6	1.5	1.5	1.4

Продовження додатку А

ν_2	ν_1								
	12	14	16	20	24	30	40	50	75
200	1.8	1.7	1.7	1.6	1.6	1.5	1.7	1.4	1.4
∞	1.8	1.6	1.6	1.6	1.5	1.5	1.4	1.4	1.3

Додаток Б

Значення критерію Стьюдента при рівні значущості 0.05

Число ступенів свободи	Рівень значущості 0.05	Число ступенів свободи	Рівень значущості 0.05
1	12.70	18	2.10
2	4.30	19	2.09
3	3.18	20	2.09
4	2.78	21	2.08
5	2.57	22	2.07
6	2.45	23	2.07
7	2.36	24	2.06
8	2.31	25	2.06
9	2.26	26	2.06
10	2.23	27	2.05
11	2.20	28	2.05
12	2.18	29	2.05
13	2.16	30	2.04
14	2.14	40	2.02
15	2.13	60	2.00
16	2.12	120	1.98
17	2.11		

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

1. Вентцель Е.С. Теория вероятностей. – М. ФМ, 1962, 264 с.

2. Лобода Н.С., Гопченко Є.Д. Стохастичні моделі у гідрологічних розрахунках. - Навчальний посібник.– Одеса: Екология, 2006. – 200 с.

3. Рождественский А.В.,Чеботарев А.И. Статистические методы в гидрологии. - Л.: Гидрометеоиздат, 1974. - 424с.

4. Шелутко В.А. Численные методы в гидрологии. -Л.: Гидрометеоиздат,1991.- 238 с.

5. Школьний Є.П., Лоєва І.Д., Гончарова Л.Д. Обробка та аналіз гідрометеорологічної інформації: навчальний підручник. - К.: Міносвіти України, 1999. - 600 с.