МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Одеський державний екологічний університет

Н. С. ЛОБОДА

МЕТОДИ ПРОСТОРОВОГО УЗАГАЛЬНЕННЯ ГІДРОЛОГІЧНОЇ ІНФОРМАЦІЇ

Конспект лекцій

Одеса
"Екологія"
2008

ББК 26.22
Л 68
УДК 556.16

Конспект лекцій призначений для магістрів, які навчаються за спеціальністю "Гідрологія та гідрохімія", напрям підготовки "Гідрометеорологія". В конспекті розглянуто методи багатовимірного статистичного аналізу стосовно до їх застосування при географічних узагальненнях стоку річок.

Конспект лекцій може бути використаний при виконанні дипломних, магістерських та аспірантських робіт.

Друкується за рішенням Вченої ради Одеського екологічного університету.
Протокол № 9 від 26.10.2006 р.

ЗМІСТ

ВСТУП

При використанні даних спостережень за річковим стоком у гідрологічних розрахунках надійна інформація може бути отримана у випадку, коли тривалість рядів спостережень становить 50-100 років. Однак добре вивчені у гідрологічному відношенні водозбори – це рідкість, більшість рядів гідрологічних спостережень значно коротша. Ситуація ускладнюється тим, що водний режим багатьох річок трансформується в результаті водогосподарської діяльності та змін глобального клімату, що переводить такі річки до розряду недостатньо вивчених. Компенсація недостатності спостережень відбувається за рахунок просторових узагальнень стоку.

Об'єкт вивчення – характеристики водних ресурсів.

Предмет вивчення – методи просторового узагальнення гідрологічних характеристик.

Головна задача дисципліни *"Методи просторового узагальнення гідрологічної інформації"* – обгрунтування вибору методу просторово-часового узагальнення тої чи іншої гідрологічної характеристики в залежності від якості вихідної інформації з метою отримання максимальної надійності та достовірності кінцевого результату.

Основні методи – методи багатовимірного статистичного аналізу.

Мета дисципліни – набування досвіду при аналізі й інтерпретації результатів розрахунків за методами багатовимірного статистичного аналізу.

До самостійних розділів дисципліни належать:

- проблеми та задачі методів географо-гідрологічного узагальнення гідрологічної інформації;
- метод факторного аналізу;
- метод головних компонентів;
- метод сумісного аналізу даних (аналіз складових просторових дисперсій);
- регресійні моделі у гідрологічних розрахунках;
- дискримінантний аналіз.

У ході вивчення дисципліни студент повинен отримати такі вміння і знання.

Знання:

- задачі просторового узагальнення гідрологічних характеристик;
- основні положення методу сумісного аналізу даних за С.М.Крицьким та М.Ф.Менкелем;

- основні положення методу регресійного аналізу: коефіцієнт кореляції, кореляційне відношення, коефіцієнт множинної кореляції, частинний коефіцієнт;
- основні принципи методу головних компонент;
- головні принципи методу факторного аналізу;
- основні положення дискримінантного аналізу;

Вміння:

- обґрунтовувати спосіб узагальнення тої чи іншої характеристики стоку за методом сумісного аналізу;
- вміти обґрунтовувати оптимальний добір предикторів при аналізі розрахункових рівнянь множинної регресії;
- виконувати аналіз інформації, що міститься в кореляційних матрицях, на основі методу головних компонент;
- виконувати районування за синхронністю коливань стоку на базі методу головних компонент;
- відновлювати ряди стоку на базі методу головних компонент;
- використовувати результати Q- модифікації факторного аналізу до районування стоку;
- використовувати дискримінантну функцію до обґрунтування меж районування.

Аналіз просторової мінливості характеристик водного режиму необхідний для наукового обґрунтування роботи водогосподарських систем та вирішення великомасштабних водних проблем, пов'язаних із змінами водних ресурсів в результаті антропогенної діяльності (перекидання стоку).

Дисципліна забезпечена підручниками, посібниками, методичними вказівками в достатній кількості.

Для успішного засвоєння дисципліни студентам необхідні знання та вміння з таких дисциплін як "Теорія імовірностей та математична статистика", "Методи аналізу та обробки гідрометеорологічної інформації", "Гідрологічні розрахунки".

Значний внесок у розвиток статистичних методів при вирішенні спеціалізованих практичних задач гідрології внесли А.В. Рождественський (Російський гідрологічний інститут, Санкт-Петербург, Росія); С.М. Крицький та М.Ф. Менкель, Раткович Д.Я., Болгов М.В. (Інститут водних проблем, Москва, Росія).

Автор виражає подяку професору, д.техн.н, ***Євгену Павловичу Школьному*** (Одеський Державний Екологічний Університет, Україна) за його успішну та багатоплідну працю в області розвитку та застосування методів математичної статистики і теорії випадкових процесів при вирішенні наукових та практичних задач у гідрометеорології.

1 ПРОБЛЕМИ Й ЗАДАЧІ МЕТОДІВ ГЕОГРАФО-ГІДРОЛОГІЧНИХ УЗАГАЛЬНЕНЬ

Методи географо-гідрологічних узагальнень використовують при недостатній гідрологічній вивченості розрахункового водозбору або за відсутності даних спостережень. Географо-гідрологічні дослідження спрямовані на характеристику усереднених (на різному просторово-часовому рівні) умов формування стоку. При цьому залучаються дані з інших водозборів, на які розповсюджується умова схожості особливостей формування стоку й підстильної поверхні. Для оцінки статистичних параметрів стоку невивчених, з погляду вимірювань стоку, водозборів розробляються карти ізоліній досліджуваних характеристик, виконується їх районування. Для урахування впливу чинників підстильної поверхні, як правило, за регресійними моделями будуються залежності, які зв'язують характеристики річкового стоку з кількісними показниками цих чинників. Чинники або фактори стоку традиційно діляться на зональні, азональні і інтразональні (місцеві). Пов'язані з кліматом зональні чинники обумовлюють плавну й безперервну зміну характеристик стоку, просторове узагальнення яких представляється,зазвичай, у вигляді карт ізоліній. Інтразональні і азональні чинники спричиняють дискретність просторового розподілу характеристик стоку, яка при просторовому узагальненні виражається в районуванні території, тобто виділенні ділянок, в межах яких зональні відмінності невеликі, що і дозволяє прийняти єдине значення розрахункового параметра. Азональні чинники пов'язані з розміром, формою і структурою конкретних водозборів і не залежать від географічного положення водозбору. Вплив азональних чинників усувається за допомогою поправкових коефіцієнтів до порайонних або знятих з карти ізоліній значень, в окремих випадках ці коефіцієнти можуть бути представлені у вигляді регіональних залежностей від морфометричних характеристик водозборів.

Основи методу географо-гідрологічних узагальнень закладені В.Г. Глушковим і одержали свій подальший розвиток в роботах А.М. Бефані, П.С. Кузіна і В.І. Бабкина, С.М. Крицкого і М.Ф.Менкеля, А.В. Рождественского , Г.П. Калініна , А.В. Хрістофорова та інш.

Перші схеми районування території України наводяться в роботах В. Докучаєва, Л. Берга, які були опубліковані в кінці XIX – на початку XX сторіч. На сучасному етапі слід зазначити роботи А. Маринича, А. Ланько, П. Шищенко, Р. Міллера.

Основні задачі методу географічного узагальнення можна сформулювати таким чином.

1. З'ясування доцільності географічних узагальнень.
2. З'ясування ступеня географічного узагальнення.
3. Вибір способу географічних узагальнень.
4. Оцінка меж географічного узагальнення.
5. Пошук доцільних і обгрунтованих форм опису просторового розподілу узагальнювальних параметрів річкового стоку (проблема картографування або районування).

Вибір методу картографування характеристик стоку пов'язаний з проблемою співвідношення між впливом зональних і азональних чинників. У тих випадках, коли роль зональних чинників переважає, застосовується принцип географічної інтерполяції і будується карта ізоліній. При використанні районування передбачається, що в межах виділеного району спостерігається відносна однорідність стокоформуючих чинників. При гідрологічному районуванні слід ураховувати масштаб районування, оскільки такі таксономічні одиниці як фізико-географічний пояс, країна, зона, провінція, підзона не завжди знаходять своє відображення в масштабах гідрологічних районів. Річ у тому, що гідрологічне районування найчастіше спирається на переважаючий вплив одного чинника формування стоку, тоді як у физико-географічному районуванні переважає принцип комплексності або ландшафтно – генетичний принцип.

У теперішній час широко використовуються нові ймовірнісно-статистичні показники, ефективні способи аналізу і синтезу явищ, оцінки надійності і стійкості географічних узагальнень.

Першим кроком є аналіз вихідних даних та вибір репрезантативних рядів стоку. На другому етапі виконується аналіз закономірностей коливань стоку та виділяються райони з синхронними коливаннями стоку. Надалі на основі виділених районів виконується приведення коротких рядів стоку до довгого періоду спостережень за методом аналогії та аналізуються закономірності коливань стоку у кожному з районів. Виконується обгрунтування способу узагальнень характеристик стоку. При переважанні впливу географічних чинників на просторовий розподіл характеристик виконується картування або розробляються розрахункові залежності за регресійними моделями. Якщо характеристика стоку визначається за даними спостережень з малою достовірністю, то виконується її районування. При проведенні меж гідрологічних районів використовуються класифікаційні методи, які дозволяють вирішити проблему належності приграничних водозборів до тієї чи іншої сукупності.

2 ФАКТОРНИЙ АНАЛІЗ ПРИ ВИРІШЕННІ ЗАДАЧ ГІДРОЛОГІЧНИХ РОЗРАХУНКІВ

2.1 Теоретичні основи методу факторного аналізу

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:

$$x_j = \sum_{p=1}^{k} l_{jp} f_p + v_j, \ (j=1,m) \qquad (2.1)$$

де X_j - центрована початкова змінна;

m - кількість змінних;

k - число факторів ($k << m$);

p - номер фактора;

l_{jp} - навантаження j - тої змінної на p - тий фактор або факторна вага;

f_p - некорельовані між собою фактори;

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

Якщо у рівності (2.1) розгорнути суму, то прийдемо до системи рівнянь [28]:

$$\begin{cases} x_1 = l_{11} f_1 + l_{12} f_2 + \ldots + l_{1k} f_k + v_1 \\ x_2 = l_{21} f_1 + l_{22} f_2 + \ldots + l_{2k} f_k + v_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_m = l_{m1} f_1 + l_{m2} f_2 + \ldots + l_{mk} f_k + v_m \end{cases} \qquad (2.2)$$

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

Матрична форма рівняння (2.1) має вигляд

$$X = LF + V, \tag{2.3}$$

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

$$K = M\left[X \cdot X'\right] \tag{2.4}$$

Ураховуючи (2.3), можна прийти до такого матричного виразу

$$K = LL' + D, \tag{2.5}$$

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

$$\frac{\partial L_П}{\partial l_{jp}} = 0; \quad \frac{\partial L_П}{\partial d_j} = 0, \tag{2.6}$$

де $L_П$ - функція правдоподібності.

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

$$K_{jj} = \sum_{p=1}^{k} l_{jp}^2 + d_j, \text{ при } \quad j = i; \tag{2.7}$$

$$K_{ji} = \sum_{p=1}^{k} l_{jp} l_{ip} \quad \text{при} \quad j \neq i. \tag{2.8}$$

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 [28].

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

$$h_j^2 = \sum_{p=1}^{k} l_{jp}^2 \tag{2.9}$$

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

$$S_p = \frac{\sum_{j=1}^{m} l_{pj}^2}{m} 100\%, \tag{2.10}$$

де m - кількість розглядуваних змінних.

Загальний внесок всіх виділених факторів в сумарну дисперсію досліджуваних змінних дорівнює

$$S = \sum_{p=1}^{k} S_p . \qquad (2.11)$$

2.2 Застосування методу факторного аналізу до районування за синхронністю коливань стоку

Синхронними називають коливання стоку річок, на яких спостерігається однаковий хід водності протягом 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 [5]. 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

$$d_j = \sqrt{l_{j1}^2 + l_{j2}^2}, \qquad (2.12)$$

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 i-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

$$\theta_j = \arcsin \frac{l_{j3}}{d_j} \qquad (2.13)$$

$$\lambda_j = \arcsin \frac{l_{j2}}{\sqrt{l_{j1}^2 + l_{j2}^2}}, \qquad (2.14)$$

де

$$d_j = \sqrt{l_{1j}^2 + l_{2j}^2 + l_{3j}^2} \qquad (2.15)$$

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.

Розглянемо приклад районування за закономірностями коливань стоку Північно-Західної частини України. Використано 29 рядів річного стоку з періодом сумісних спостережень з 1955 по 1986 роки, який складає 32 роки. До аналізу залучені дані по водозборах р.Західний Буг, правобережних приток Прип'яті, правобережних приток р. Дніпро - рр. Уж, Ірша, Тетерев) та прилеглих територій (лівобережні притоки р. Дністер, верхів'я р.Південний Буг). Установлено, що перші два фактори пояснюють 57,4% сумарної дисперсії вихідних даних, а перші три - 81,9%. Районування за Q- модифікацією факторного аналізу здійснено на основі графічних побудувань, в яких використовуються результати представлення кореляційної матриці у вагових навантаженнях кожного ряду на виявлений гіпотетичний фактор. За результатами факторного аналізу виділені два угрупування, що утворюють два територіальних райони (західний 1а та східний 1б) з синхронними коливаннями стоку (рис. 2.1).

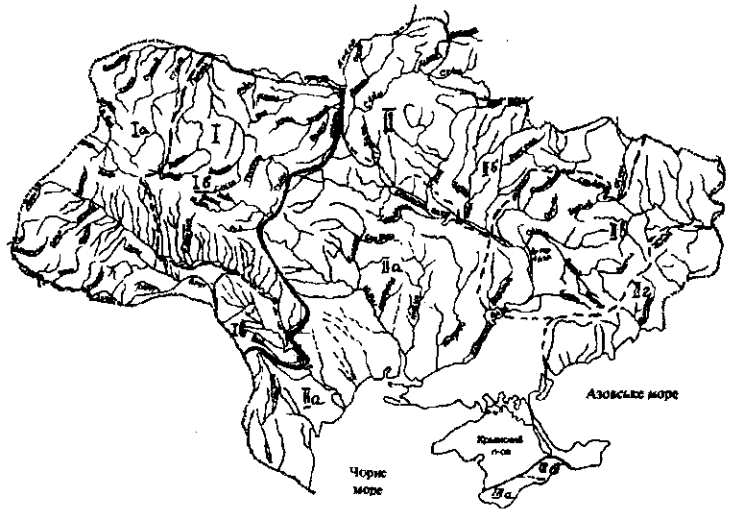

Рис. 2.1 – Карта-схема районів з синхронними коливаннями річного стоку річок України

В перший район входять водозбори річок бассейну р.Західний Буг, верхів'я р. Дністер, а також такі притоки Прип'яті, як Вижівка, Турія, Стир. Друге угруповання утворюють притоки р.Дністер від р. Серет до р. Ушиця включно, верхів'я р. Південний Буг, правобережні притоки р.Прип'ять, починаючи від р.Горинь (рис.2.2). Лінія розмежування проходить через вододіл річок Стир – Горинь.

Доцільність прийняття до розрахунків таких угруповань підтверджується наступним: середній коефіцієнт кореляції $r_{сер}$ між річним стоком усіх розглянутих водозборів дорівнює 0,53, що вказує на синфазність коливань стоку, для району **1а** - $r_{сер} = 0{,}77$, для району **1б** - $r_{сер} = 0{,}64$. Тобто, у межах виділенних районів коливання стоку можуть розглядатися не як синфазні, а як синхронні.

Аналогічним чином були виділені райони IIIа (Західний) та IIIб (Східний) для водозборів Гірського Криму (рис.2.3). Внесок перших двох факторів становить 81%, а кут між двома угрупуваннями близький до $90^{\circ}C$.

Це вказує на суттєву різницю у коливаннях стоку заходу і сходу, оскільки ($r = cos\,90^{\circ} = 0$). Окреме положення займає водозбір р.Салгір-м. Сімферополь, особливість коливань стоку на цьому водозборі пов'язана з інтенсивним використанням стоку для водогосподарських потреб, яке порушує загальну закономірність коливань стоку, обумовлену коливаннями кліматичних факторів.

Як правило, такі водозбори мають низький рівень інформативності, *який описується величиною d. Наприклад, для водозбору р.Салгір* – м.Сімферополь ця величина становить 0,45. Різниця між західною й східною частинами обумовлена кліматичними факторами, насамперед, опадами.

Західні схили знаходяться під значним впливом середземноморських циклонів, які забезпечують приплив зволожених повітряних мас. Захищенність Кримськими горами південно-західного узбережжя забезпечує формування субтропічного клімату, що приводить до значного розбігу між умовами формування стоку заходу і сходу.

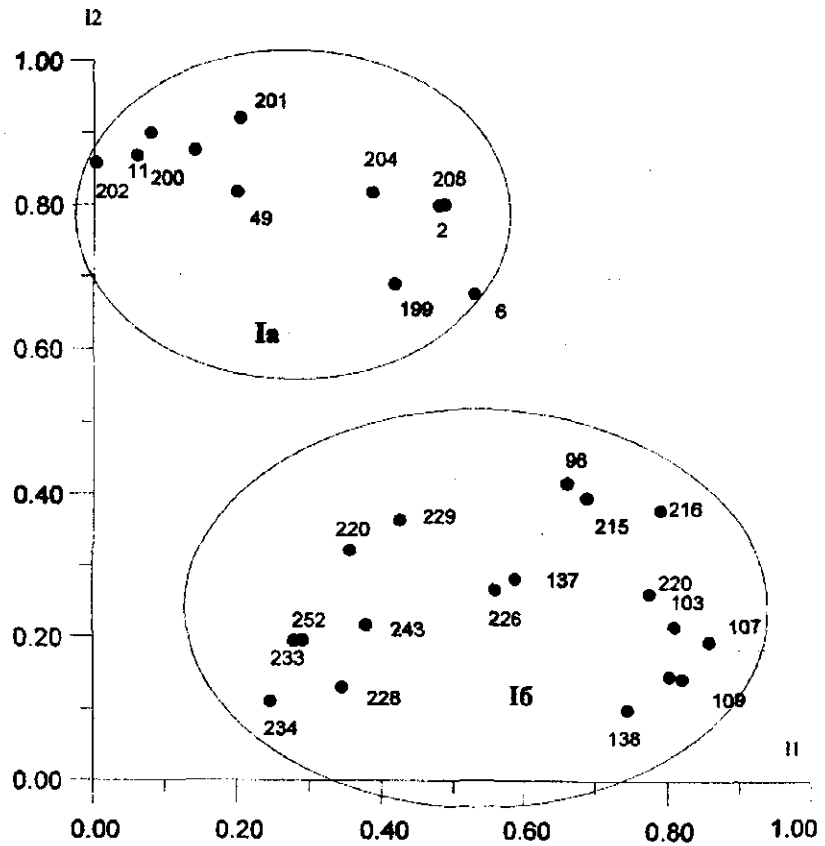

Рис. 2.2 Виділення угруповань з синхронними коливаннями річного стоку за двома факторами для річок Північно-Західної України (біля точок – номери водозборів)

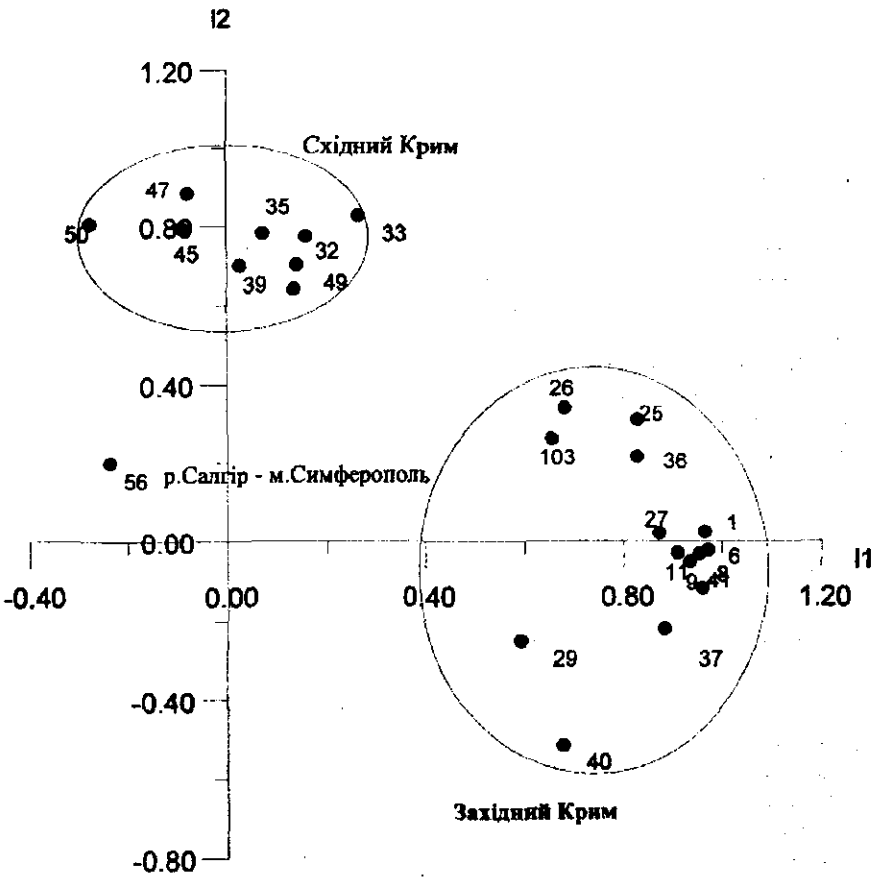

Рис. 2.3 Виділення угруповань з синхронними коливаннями річного стоку Гірського Криму за двома першими факторами (біля точок - номери водозборів)

3 МЕТОД ГОЛОВНИХ КОМПОНЕНТІВ У ГІДРОЛОГІЧНИХ РОЗРАХУНКАХ

3.1 Теоретичні основи методу головних компонентів

Метод головних компонентів (природних ортогональних функцій або ПОФ) являє собою один з методів лінійного перетворення інформації, який полягає в лінійному ортогональному перетворенні полів вхідних величин у базисі власних векторів матриці кореляцій або коваріацій. Інакше кажучи, на основі матриць кореляції визначається система ортогональних, лінійно незалежних функцій, називаних власними векторами, які відповідають системі незалежних випадкових величин, іменованих власними значеннями або власними числами матриці кореляцій. Пошук власних векторів і власних значень досягається шляхом розв'язання матричного рівняння вигляду

$$R_X - \lambda_i U_i = 0, \tag{3.1}$$

де R_X - матриця коефіцієнтів кореляції розміром $m \times m$ (m відповідає числу розглянутих об'єктів);

U_i - власний вектор матриці кореляцій;

λ_i - відповідне власному вектору власне значення.

Матриця R_X має m коренів або m власних чисел λ, які є дійсними, позитивними й простими. Для знаходження m власних векторів, що відповідають m власним числам, необхідно розв'язання m систем лінійних рівнянь. Процедура розрахунку здійснюється, як правило, за допомогою ітераційних методів, серед яких найпоширенішим є метод Якобі [28].

Сукупність власних векторів утворює базис, у якому проводиться розкладання полів вихідних даних

$$U' \cdot \varphi_i = Z_i, \tag{3.2}$$

де U' - транспонована матриця U розміром $m \times m$;

φ_i - i-ий випадковий вектор (поле) центрованих і нормованих вихідних даних, що підлягає розкладанню;

Z_i - вектор головних компонентів, який є результатом лінійного перетворення φ_i -того поля відповідним власним вектором. Оскільки

власні вектори ортонормовані, головні компоненти поля є статистично незалежними. Рівність (3.2) означає, що вхідне поле розкладене на m незалежних компонент.

Складові вектора Z_i для p-тої компоненти розкладання визначаються таким чином:

$$z_{ip} = \sum_{k=1}^{m} U_{pk} \varphi_{ik} \; ; \; p = \overline{1,m} \; , \tag{3.3}$$

де z_{ip} - складові p - тої компоненти розкладання;

U_{pk} - вагові коефіцієнти, що відображують внесок k- того об'єкта в кожну p-ту компоненту (або внесок p-тої компоненти в k-тий об'єкт), які є складовими власних векторів матриці кореляцій;

φ_{ik} - i-ий випадковий вектор (поле) центрованих і нормованих вихідних даних, який підлягає розкладанню.

Значення U_{pk} змінюються в просторі при переході від об'єкта до об'єкта, але не залежать від часу. Система функцій U_{pk} часто представляється як функція координат (x_k, y_k) для k- того об'єкта й має назву "базисної функції".

Отримані компоненти мають наступну властивість: кожне p- е власне значення λ_P матриці кореляцій є дисперсією p-ої головної компоненти σ^2_{Zp}

$$\lambda_P = \sigma^2_{Zp}, \tag{3.4}$$

Тоді сума дисперсій m компонент дорівнює сумі власних чисел матриці й, отже, дорівнює сліду матриці кореляцій

$$\sum_{p=1}^{m} \sigma^2_{Zp} = \sum_{p=1}^{m} \lambda_p = t_R R_x = m, \tag{3.5}$$

де t_R - слід матриці кореляцій (слідом матриці кореляцій називається сума елементів, розташованих на головній діагоналі).

Якщо розкладання за ПОФ виконувати в базисі власних векторів вихідного центрованого поля ΔX_i, то (3.2) буде мати вигляд

$$W'\Delta X_i = Z_i, \tag{3.6}$$

де W - матриця власних векторів матриці коваріації.
У такому випадку

$$\sum_{p=1}^{m} \sigma_{Zp}^2 = \sum_{p=1}^{m} \lambda_p = t_R K_x = \sum_{p=1}^{m} \sigma_{Xp}^2. \tag{3.7}$$

Сума власних значень матриці є, як це було показано раніше, сумою дисперсій головних компонент. Ця сума для матриці коваріацій дорівнює сумарній дисперсії поля. Остання розподіляється таким чином, що найбільша її частина являє собою дисперсію першої головної компоненти, яка у свою чергу, згідно з (3.4) є першим власним значенням. Сума дисперсій головних компонент матриці коваріацій дорівнює сумі дисперсій вихідних рядів, тобто

$$\sum_{p=1}^{m} \sigma_{Zp}^2 = \sum_{j=1}^{m} \sigma_{Xj}^2 \tag{3.8}$$

Таке подання дозволяє більш наочно зрозуміти суть методу головних компонент, тому що ці некорельовані лінійні комбінації вихідних змінних містять у собі всю дисперсію, укладену в m змінних вхідного масиву даних. Декілька перших власних чисел ($\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_m$) вичерпують основну частину сумарної дисперсії поля, тому при аналізі результатів розкладання особлива увага приділяється першим власним значенням і відповідних їм компонентам. А оскільки великомасштабні процеси характеризуються великою дисперсією, то справедливо допущення, що саме вони відображені у перших компонентах.

Якщо розташувати власні числа матриці у спадному порядку, то перше власне число буде являти собою величину дисперсії, що відповідає першій компоненті, друге власне число - величину дисперсії, що відповідає другій компоненті й т.д. Через те, що при використанні кореляційної матриці сума власних чисел дорівнює числу розглянутих змінних m, то розділивши кожне власне число на m або $\sum_{j=1}^{m} \lambda_j$, можна одержати частку від сумарної дисперсії, що описується кожною з розглянутих компонент.

Частка істотної інформації із всієї сумарної інформації про поле оцінюється за допомогою співвідношення

$$S = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{s=1}^{m} \lambda_s}, \tag{2.9}$$

де чисельник дорівнює сумі дисперсій, що припадає на p перших головних компонент, а знаменник дорівнює сумарній дисперсії поля. Задаючись значенням S (наприклад, $S = 0{,}70\text{-}0{,}80$), можна встановити число перших компонент, які варто враховувати, щоб скоротити обсяг вхідної інформації й зберегти при цьому її основний зміст. При використанні методу головних компонент вхідна інформація не тільки заміщається малим числом статистичних функцій, але й зберігає в цих функціях їхнє фізичне навантаження. Одержані в результаті розкладання функції є відображенням реальних фізичних процесів, які обумовлюють просторово-часовий розподіл досліджуваних гідрометеорологічних величин.

3.2 Застосування методу головних компонентів до аналізу синхронності коливань стоку

Аналізується кореляційна матриця 20 рядів річного стоку басейну р.Уссурі за період сумісних спостережень, починаючи з 1960 року і закінчуючи 1986 роком, тобто 27 років. Внесок перших компонент у загальну дисперсію становить: 63% для першої компоненти, 20% - для другої та 5% - для третьої [17]. Оскільки для виявлення просторових закономірностей зміни базисних функцій перших компонент розкладання за природними ортогональними функціями необхідні дані про координати центрів тяжіння водозборів, досліджуваний водозбір був вкритий координатною "сіткою" й положення центрів тяжіння було представлене у вигляді умовних координат (рис.3.1).Знак вагових коефіцієнтів першої компоненти розкладання не змінюється (табл.3.1), що інтерпретується як однорідний вплив найбільш масштабного фізичного процесу на формування полів річного стоку. Зазвичай, перша компонента розкладання розглядається як статистичне відображення дії атмосферних процесів глобального масштабу. Вагові коефіцієнти

другої компоненти змінюють знак, що у літературних джерелах [25] інтерпретується як існування різниці у закономірностях коливань гідрологічних характеристик на різних ділянках території.

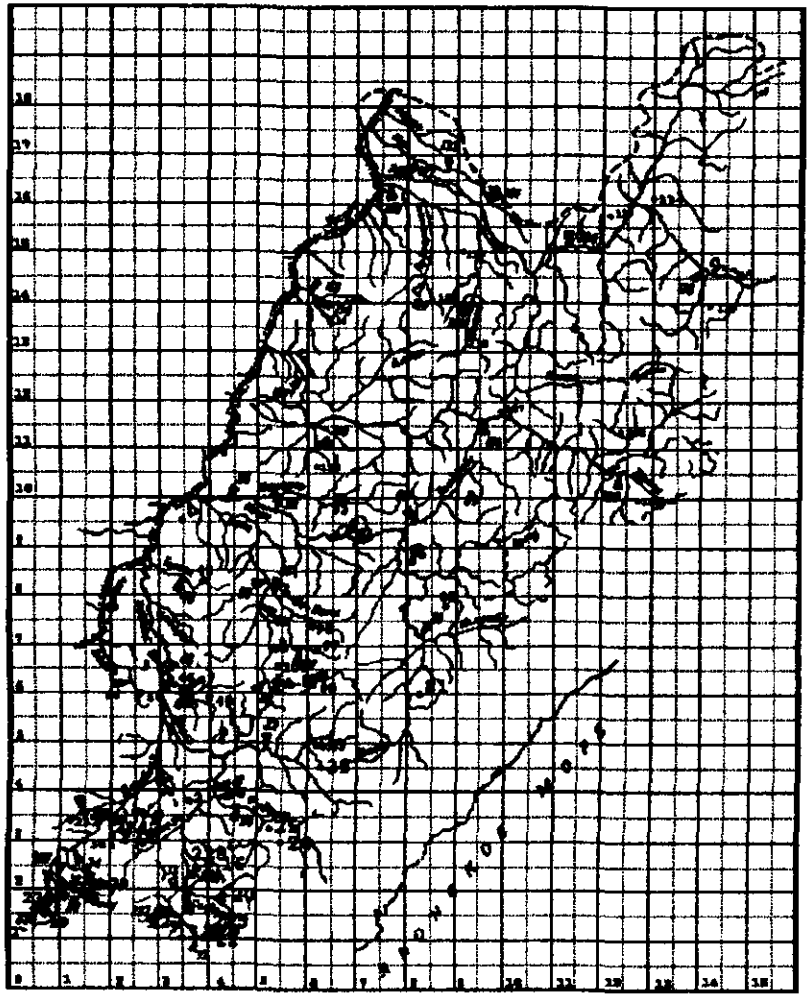

Рисунок 3.1 - Координатна сітка, гидрологічні пости та центри тяжіння водозборів для басейну р.Уссурі

Таблиця 3.1 – Значення перших базисних функцій

Номер за картою	Річка –пост	U_1	U_2	U_3
2	р.Уссурі-сел. Кіровське	0,2582	0.2184	0,0398
12	р.Улахе-с.Березняки	0,2587	0.2592	0,4140
20	р.Сидагоу-с.Извілінка	0,2208	0,2684	0,3047
25	р.Фудзін-с.Уборка	0,2566	0,2193	0,1200
36	р.Даубіхе-с.Яковлівка	0,2266	0,2933	-0,0678
42	р.Хоніхеза-с.Варфоломєєвка	0,2465	0,2414	0,0768
45	р.Шетуха-с.Криловка	0,2359	0,1001	-0,3672
48	р.Тамга-с.Тамга	0,2446	0,0027	-0,3074
83	р.Іман-с.Картун	0,2317	-0,0645	-0,1489
85	р. Іман-сел. Вагутон	0,2406	-0,0745	-0,1916
89	р.Сібічі-с. Сібічі	0,2078	-0,2371	-0,1818
93	р.Вака-с.Ракітне	0,1819	0,0402	-0,2741
98	р.Тудо-Вака-с.Аріадне	0,2067	0,0836	-0,3054
107	р.Бікін-ст.Звєньєва	0,2290	-0,2414	0,0145
114	р.Горбун-с.Пушкіно	0,1943	-0.3282	0,1602
119	р.Подхорьонок-с.Дормидонтівка	0,1914	-0,3282	0,1762
120	р.Правий Подхорьонок – лзу. Медвежий Ключ	0,2159	-0,3268	0,1797
128	р.Сукпай-мет.ст.Сукпай	0,2337	-0,2188	0,2627

продовження табл.3.1

131	р.Кія-с.Марусіно	0,2038	-0,3472	0,1932
204	р.Каменка-сел. Каменський	0,1548	0,2030	0,1436

Інакше кажучи, другий за значущістю фізичний процес обумовлює прояв несинхронності у коливаннях характеристик стоку. Положення нульової ізолінії ($U_2 = 0$) розглядається як межа між районами з *несинхронними коливаннями цих характеристик. Нульова ізолінія поділяє* територію водозбору р. Уссурі навпіл (рис.3.2). Різницю у коливаннях стоку між північчю і півднем можна пояснити тим, що південна частина басейну р.Уссурі знаходиться у області дії субтропічних мусонів, а північна розташована у області впливу мусонів помірних широт. Тобто другий за значущістю фізичний процес може бути інтерпретований як атмосферний процес синоптичного масштабу.

Правильність прийнятих рішень щодо районування за синхронністю підтверджується підрахунками осередненого у просторі коефіцієнта кореляції, який для усього басейну р.Уссурі дорівнює 0,52, для району 1 – 0,77 та для району 2 – 0,79.

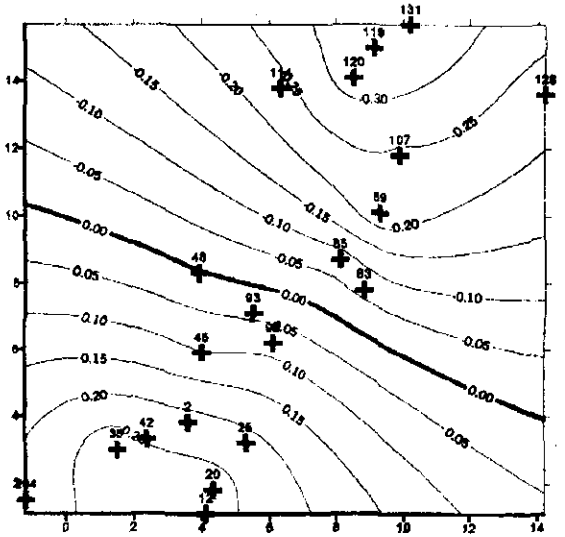

Рисунок 3.2 - Схема розподілу U_2 у басейні р.Уссурі (+ - центри тяжіння водозборів)

3.3 Вирішення задач фільтрації та відновлення гідрологічної інформації

Метеорологічні поля формуються під дією атмосферних процесів різного масштабу: макропроцесів, пов'язаних із планетарними рухами мас повітря; процесів, обумовлених атмосферною циркуляцією (синоптичними процесами); процесів мезо- і мікромасштабу, а також дрібномасштабних флуктуацій, пов'язаних з похибками вимірів і недоліками первинної обробки результатів спостережень. При необхідності виключити вплив малоінформативних фізичних процесів на формування метеорологічних полів використовують математичний апарат, вже викладений в (3.1), роблячи зворотний перехід від компонент до значень спостереженої величини.

Структура гідрологічних полів визначається атмосферними процесами, оскільки стік є "продукт клімату". Але на рівні складових 3-4 порядку розкладання може проявитися вплив факторів підстильної поверхні [14].

Будь-який елемент матриці вихідних спостережених значень φ_{ij} (на i -тім розглянутім об'єкті в j - ий момент часу) може бути розрахований, якщо проблему власних векторів вирішено

$$\varphi_{ij} = \sum_{k=1}^{m} U_{ki} z_{kj}, \text{ при } i = 1,m; j = 1,n, \qquad (3.10)$$

де φ_{ij} - складові j-того випадкового вектора (поля) центрованих і нормованих вихідних даних, що підлягає розкладанню;

U_{ki} - вагові коефіцієнти, які відображають внесок i-того об'єкта в кожну k -ту компоненту або складові власних векторів матриці кореляцій;

z_{kj} - складові k-тої компоненти розкладання;

m - число об'єктів;

n - довжина вихідних рядів.

Значення U_{ki} змінюються в просторі при переході від об'єкта до об'єкта, але не залежать від часу. Система функцій U_{ki} часто представляється як функція координат (x_i, y_i) для i - того об'єкта і має назву базисних функцій

$$U_{ki} = f(x_i, y_i) = U_k(x_i, y_i) \qquad (3.11)$$

Приклад просторового розподілу другої базисної функції розкладання полів річного стоку у басейні р.Уссурі показаний на рис.3.2.

Вагові коєффіцієнти можуть розподілятись у просторі із певною закономірністю. Пошук зв'язків між ваговими навантаженнями на перші компоненти розкладання та показниками кліматичних чинників і чинників підстильної поверхні для Українських Карпат показав, що найбільш оптимальним предіктором при розрахунках річного стоку з водозборів Українських Карпат є висота місцевості. Саме висота місцевості має найбільш тісний кореляційний зв'язок з ваговими навантаженнями w_1 на першу, найбільш інформаційну, складову. Висота водозборів у даному випадку є інтегральним показником впливу кліматичних факторів, насамперед, опадів та випаровування, на формування річного стоку гірських райнів. Друге місце за своєю значущостю при установленні зв'язків w_1 із різними чинниками займає норма інфільтрації річних опадів у водоносні горизонти U_0. Вагові коефіцієнти на другу компоненту розкладання мають тісний зв'язок з логарифмом площі водозборів $\lg(F)$, а на третю компоненту - із залісеністю ($f_Л$). Отримані результати свідчать про те, що фізичні процеси, представлені другою та третьою компонентами, обумовлюють мезомасштабні зміни характеристик стоку у просторі, які можуть бути пов'язані із азональними $\lg(F)$ та інтразональними $f_Л$ чинниками [18]. Для Північно-Західної України установлені такі зв'язки між ваговими коефіцієнтами перших компонент розкладання полів річного стоку та стокоформувальними чинниками (Лобода Н.С., 2005)

$$w_{1i} = -0{,}000884\overline{E}_{mi} + 0{,}000361\overline{X}_i + 0{,}660; \quad R = 0{,}884 \qquad (3.12)$$

$$w_{2i} = -0{,}00150\overline{E}_{mi} - 0{,}00290\overline{X}_i + 3{,}014; \quad R = 0{,}734 \qquad (3.13)$$

$$w_{3i} = 0{,}00180\overline{E}_{mi} + 1{,}450; \quad r = -0{,}788, \qquad (3.14)$$

де $\overline{E}_{mi}$ - річна норма максимально можливого випаровування для i-того об'єкта;

$\overline{X}_i$ - річна норма опадів для i- того объекта.

Зміст отриманих рівнянь можна інтерпретувати наступним чином: перші компоненти розкладання полів річного стоку є відображенням кліматичних процесів, які відбуваються на розглядуваній території.

Складові вектора-рядка матриці Z $[z_{k1} \quad z_{k2} \quad \cdots \quad z_{kp} \quad \cdots \quad z_{km}]$ можуть бути представлені як функції часу (так звані амплітудні функції) і є загальними для всіх об'єктів

$$z_{kj} = f(t) = z_k(t) \qquad (3.15)$$

Наприклад, на рис.3.3 наведена перша компонента (перша амплітудна функція) розкладання полів річного стоку річок Західної та Східної Європи, включаючи Україну.

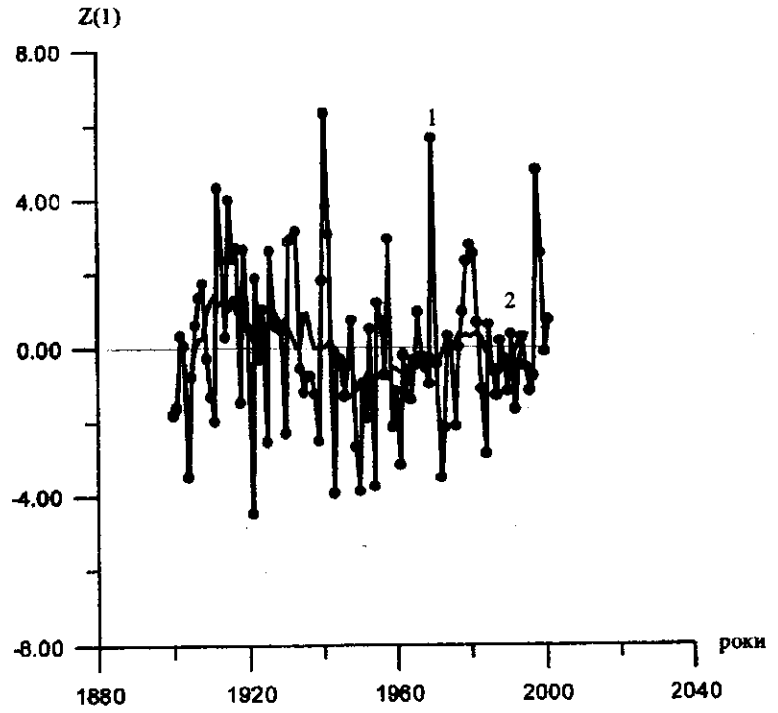

Рисунок 3.3 - Перша амплітудна функція розкладання полів річного стоку Західної та Східної Європи за природними ортогональними функціями (2 – згладжений за 11-тьма річками ряд)

Цей хронологічний графік відображає характер коливання стоку річок, обумовлене найбільш великомасштабним атмосферним процесом.

У зв'язку з вищевикладеним, формула (3.10) може бути представленою у вигляді

$$\varphi(x,y,t)=\sum_{k=1}^{m}U_k(x,y)z_k(t), \qquad (3.16)$$

При розгляді тільки перших компонентів розкладання, у яких утримується основна частина інформації, укладена у вихідних полях, вираз (2.10) перетвориться до вигляду

$$\tilde{\varphi}_{ij}=\sum_{k=1}^{p}U_{ki}z_{kj}\text{, при } i=1,m;\ j=1,n\ , \qquad (3.17)$$

де $\tilde{\varphi}_{ij}$ - відфільтровані або згладжені значення стоку;

p - число перших компонент.

Процедура відновлення вихідного ряду за (3.17) має назву процедури фільтрації.

Перехід від результатів розкладання до відфільтрованих даних $\tilde{x}_{ij}$ здійснюється за виразом

$$\tilde{x}_{ij}=\bar{x}_i+\sigma_i\sum_{k=1}^{p}U_{ki}z_{kj}\text{; при } i=1,m; \qquad (3.18)$$

або

$$\tilde{x}_{ij}=\bar{x}_i+\sum_{k=1}^{p}w_{ki}z_{kj}\text{; при } i=1,m;\ j=1,n\ , \qquad (3.19)$$

де $\bar{x}_i$ - середнє арифметичне значення вихідного ряду;

σ_i - середнє квадратичне відхилення вихідного ряду;

w_{ki} - вагові коефіцієнти амплітудної функції (k-тої компоненти розкладання), які є складовими власного вектора розкладання матриці коваріацій.

Прикладне застосування виразів (3.18-3.19) дозволяє представити процес стоку у вигляді штучного хронологічного ряду, який відображає властивості тільки тих компонентів, а, отже і відповідних їм фізичних процесів різних масштабів, які були

застосовані при фільтрації. При цьому використовуються амплітудні функції (компоненти) і вагові коефіцієнти (значення власних векторів матриці кореляцій) перших складових розкладання, а також середні величини $\bar{x}_i$. Отриманий хронологічний ряд є згладженим, тому що не враховує вплив на формування досліджуваної величини процесів більш дрібного масштабу.

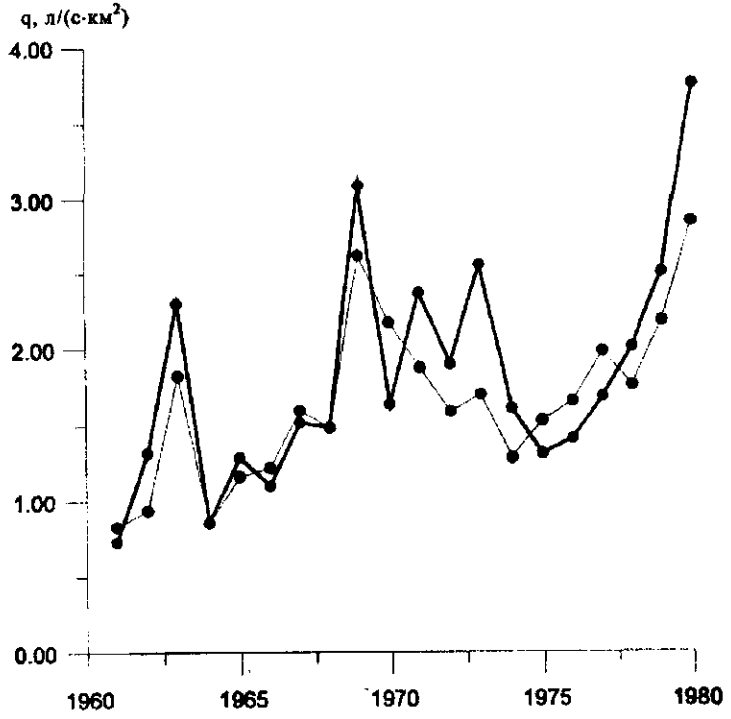

Рисунок 3.4 - Порівняння спостереженого (нижня крива) та відновленого (верхня крива) за трьома компонентами ряду річного стоку р.Когільник

Рівняння (3.19) було використано автором для розрахунків рядів природного (непорушеного водогосподарською діяльністю) річного стоку Причорноморської низовини. Як $\bar{x}_i$ розглядалася норма кліматичного (розрахованого за рівнянням водно-теплового балансу) стоку, а вагові коефіцієнти перших трьох компонент розкладання визначалися за показниками кліматичних факторів (див.3.12-3.14).

4 МЕТОД СУМІСНОГО АНАЛІЗУ ПРОСТОРОВОЇ ДИСПЕРСІЇ ГІДРОЛОГІЧНИХ ХАРАКТЕРИСТИК

4.1 Теоретичні основи методу

У гідрологічних розрахунках найчастіше використовується математична модель стоку, яка описує його ймовірністну природу. Такого роду моделі включають до себе ряд гіпотез, які дозволяють звести розрахунки до статистичної оцінки декількох параметрів моделі: середнє арифметичне значення досліджуваної гідрометеорологічної величини, коефіцієнт варіації Cv, коефіцієнт асиметрії Cs, коефіцієнт автокореляції $r(1)$.

Навіть при довгих рядах спостережень оцінки окремих статистичних параметрів визначаються з великою похибкою, тобто є статистично незначущими. До числа таких параметрів відносяться насамперед коефіцієнти автокореляції $r(1)$ й асиметрії Cs, а також розрахункове відношення Cs/Cv.

Обмеженість у часі наявних спостережень по більшості рядів стоку річок України породжує статистичну нестійкість цих параметрів, що може бути ефективно компенсовано за рахунок додаткової інформації про просторові закономірності розподілу розглядуваних характеристик річкового стоку. Для підвищення надійності оцінок статистичних параметрів за вибірковими даними рекомендується виконувати їхнє просторове узагальнення. За допомогою методу, запропонованого С.М. Крицким і М.Ф. Менкелем (1981,1982), можна обґрунтувати характер цього узагальнення. Суть методу зводиться до визначення географічної і випадкової складових загальної просторової дисперсії розглядуваного статистичного параметра A:

$$\sigma_{\Pi}^2 = \sigma_{\Gamma}^2 + \sigma_{B}^2 , \qquad (4.1)$$

де σ_{Π}^2 - повна складова дисперсії параметра;

σ_{Γ}^2 - географічна складова дисперсії параметра;

σ_{B}^2 - випадкова складова дисперсії параметра.

При цьому повна просторова дисперсія параметра оцінюється за формулою

$$\sigma_{\Pi}^2 = \frac{\sum_{j=1}^{k}(A_j - A_{CEP})^2}{k-1}, \qquad (4.2)$$

де k - число об'єктів (водозборів), об'єднаних в одну групу;

j - порядковий номер розглядуваного об'єкту (водозбору);

A_j - індивідуальна оцінка параметра (оцінка, виконана для окремого водозбору);

$A_{сер}$ - осереднена в межах виділеної групи оцінка параметра.

Випадкова складова просторової дисперсії параметра визначається як осереднена по групі виділених об'єктів дисперсія індивідуальної оцінки параметра

$$\sigma_B^2 = \frac{\sum_{j=1}^{k}\sigma_{Aj}^2}{k}, \qquad (4.3)$$

де σ_{Aj} - середнє квадратичне відхилення індивідуальної оцінки параметра A.

Географічна складова визначається за допомогою зворотного розрахунку з (4.1):

$$\sigma_{\Gamma}^2 = \sigma_{\Pi}^2 - \sigma_B^2. \qquad (4.4)$$

Якщо виконується умова

$$\frac{\sigma_B^2}{\sigma_{\Pi}^2} > \frac{\sigma_{\Gamma}^2}{\sigma_{\Pi}^2}, \qquad (4.5)$$

то можна зробити висновок, що просторовий розподіл досліджуваного параметра більшою мірою визначається випадковими властивостями поєднуваних вибірок і меншою - зміною фізико-географічних умов формування стоку по території. Таким чином, при виконанні (4.5) приймається рішення, що вибіркові оцінки параметрів можуть бути осереднені в межах досліджуваної території.

Необхідно підкреслити, що якість об'єднання тим вища, чим менший внесок географічної складової у повну просторову дисперсію параметра. Географічна складова є, власне кажучи, оцінкою статистичної неоднорідності вихідного матеріалу. Коли оцінки вибіркових параметрів дуже великі, географічна складова дисперсії, обчислена зворотним розрахунком за (4.4), може приймати негативні значення. У цьому випадку внесок випадкової складової у повну просторову дисперсію параметра може бути прийнятий рівним 100% , а географічної – 0,00%.

Середнє квадратичне відхилення осередненої у просторі оцінки статистичного параметра розраховується за співвідношенням

$$\sigma_{сер} = \sqrt{\frac{\sigma_B^2}{k} + \sigma_{\Gamma}^2} \qquad (4.6)$$

Величина σ_{CEP} поряд з умовою (4.5) також є критерієм якості об'єднання. Осереднена оцінка параметра визнається статистично достовірною, коли виконується умова

$$A_{сер} > 2\sigma_{сер} \qquad (4.7)$$

У ході послідовного об'єднання параметрів можна виявити ряди, статистичні властивості яких відрізняються від властивостей об'єднуваної сукупності: у міру збільшення числа поєднуваних об'єктів k при незначному зростанні географічної складової σ_{Γ}^2 дисперсія осередненого у межах поєднуваної сукупності параметра $\sigma_{сер}^2$ відповідно до виразу (4.6) повинна зменшуватися. "Сплеск" в убутній функції $\sigma_{сер}^2 = \varphi(k)$ свідчить про те, що досліджуваний параметр k-того ряду стоку значно відрізняється від осередненої оцінки та значень відповідного параметра інших рядів. Для таких рядів у наступних розрахунках рекомендується використовувати не осереднену, а уточнену оцінку параметра, за винятком випадків, коли ряд є статистично неоднорідним унаслідок водогосподарських перетворень або відноситься до іншого району із своїми статистичними властивостями.

Для оцінки якості розрахунків також використовуються так звані допустимі відносні середні квадратичні відхилення $\varepsilon_{доп}$, визначення параметра A за вибірковими даними.

Якщо $\varepsilon_A \le \varepsilon_{ДОП.}$, то вибіркове значення параметра приймається до розрахунку. Величина ε_A визначається за формулою

$$\varepsilon_A = \frac{\sigma_A}{A} \cdot 100\% , \qquad (4.8)$$

де σ_A - середнє квадратичне відхилення оцінки параметра A.

Для статистичних параметрів, що розраховуються по спостереженим даним з великим середньоквадратичним відхиленням, осереднена в межах поєднуваної сукупності оцінка є більш достовірною, ніж індивідуальна. Осереднені у межах статистично однорідних районів оцінки статистичних параметрів рекомендуються до використання при побудові стохастичних моделей, а також при описуванні статистичних розподілів характеристик стоку тих водозборів, на яких спостереження за стоком відсутні.

Уточнена по сукупності розглянутих об'єктів оцінка статистичного параметра розраховується на основі виразу

$$A_j^{'} = \frac{A_j \sigma_{CEP}^2 + A_{CEP} \sigma_j^2}{\sigma_{CEP}^2 + \sigma_j^2} , \qquad (4.9)$$

де $A_j^{'}$ - уточнена оцінка індивідуального значення параметра A_j з урахуванням інформації, що увійшла в поєднувану сукупність;

A_j - вхідне значення параметра по j-тому розглянутому об'єкту (водозбору);

σ_j^2 - дисперсія параметра A_j по j-тому розглянутому об'єкту (водозбору);

A_{CEP} - осереднена в межах виділеної групи об'єктів оцінка параметра A ;

σ_{CEP}^2 - дисперсія осередненої у межах виділеної групи об'єктів оцінки статистичного параметра.

Середнє квадратичне відхилення уточненого значення параметра визначається на основі наступної формули, отриманої за методом статистичних випробувань

$$\sigma_j^{'} = \frac{\sigma_j \sigma_{CEP}}{\sqrt{\sigma_j^2 + \sigma_{CEP}^2}} , \qquad (4.10)$$

де $\sigma_j^{'}$ - середнє квадратичне відхилення уточненого параметра $A_j^{'}$ по j-тому розглянутому об'єкту (водозбору);

σ_j - середнє квадратичне відхилення параметра A_j по j-тому розглянутому об'єкту (водозбору).

Таким чином, метод С.М. Крицкого та М.Ф. Менкеля дозволяє вирішувати багато задач географічного узагальнення. Наприклад, задача вибору способу географічного узагальнення може бути вирішена при розгляді умови (4.5). Якщо умова виконується, то як спосіб географічного узагальнення вибирається районування, тобто осереднення розглядуваної характеристики у межах виділеної території, якщо не виконується – картування досліджуваної характеристики у вигляді карти ізоліній. Визначення меж географічного узагальнення може спиратися на виконання умови (4.7) та аналіз залежності $\sigma_{cep}^2 = \varphi(k)$. Зростання географічної складової повної просторової дисперсії параметра буде тим інтенсивнішим, чим ширше межі просторового узагальнення.

Слід зазначити, що перед застосуванням методу бажано провести попередній аналіз вхідної інформації, використовуючи для виділення початкових угруповань вже існуючі географічні узагальнення, наприклад фізико-географічне районування або районування за синхронністю коливань стоку.

На першому етапі узагальнень може бути прийнята гіпотеза про те, що не тільки крива розподілу, а й статистичні параметри усіх розглядуваних річок належать до однієї генеральної сукупності. Надалі для окремих параметрів межі установлених статистично однорідних районів можуть розширюватися. Чим більший вплив підстильної поверхні на формування стоку, тим, як правило, меньші просторові масштаби виділених районів.

4.2 Приклади застосування методу сумісного аналізу даних

У табл. 4.1 наведений приклад розподілу складових просторової дисперсії середньобагаторічних значень (норм) річного стоку для басейну р.Уссурі [17]. Попередні угруповання водозборів по районах (Північний, Центральний, Південний) виділені згідно з результатами районування за синхронністю коливань стоку (див.розділ 2). Особливості коливань стоку у басейні р.Уссурі обумовлюються різним характером мусонних процесів на півночі та півдні водозбору.

У північній частині переважає дія мусону помірних широт, у південній – субтропічних. Виділення Центрального району обумовлено особливостями гідрогеологічної структури. Для усіх трьох виділених районів та водозбору в цілому географічна складова просторової дисперсії значно перевищує випадкову, а відносна похибка осередненого параметра перевищує допустиму (10%), що свідчить про необхідність картування норм річного стоку або пошуку рівнянь парної, або множинної регресії, які б описували географічні закономірності розподілу норм річного стоку в межах водозбору р.Уссурі.

Таблиця 4.1 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення норм річного стоку річок в басейні р. Уссурі

Район	Середнє значення параметра $\overline{q}$, л/с·км^2	Дисперсія			$\varepsilon_{сер}$, %
		повна	випадкова складова	географічна складова	
Норма річного стоку					
Північний	12,3	9,80	0,52 5%	9,28 95%	24,8
Центральний	11,2	4,99	0,36 7%	4,63 93%	19,2
Південний	8,94	2,55	0,587 23%	1,97 77%	15,8
Водозбір р.Уссурі	10,3	6,39	0,48 8%	5,91 92%	34,0

У межах басейну р. Дністер при обгрунтуванні способу узагальнення статистичних параметрів $Cv, r(1), Cs/Cv$ попередньо були виділені 4 райони за особливостями фізико-географічних умов. До першого входить гірський Дністер (правобережні притоки), до другого – передгірний Дністер (лівобережні притоки), до третього – річки Північної і Центральної Молдови, які характеризуються підвищеним підземним живленням річок за рахунок розвантаження карстових вод. Четвертий район утворюють річки Причорноморської низовини з посушливим кліматом і незначною часткою підземного живлення (5%). Просторова зміна коефіцієнтів варіації обумовлена характером зміни загальної зволоженості території: значення коефіцієнтів варіації збільшуються в міру переходу із зон надлишкового та достатнього зволоження (райони 1 і 2) до зони недостатнього зволоження (райони 3 та 4). Але на просторовий розподіл коефіцієнтів варіації впливає такий фактор підстильної поверхні як гідрогеологічна структура й пов'язане з нею підземне живлення. По перше, у межах водозбору р.Дністер існують карстові зони, які впливають на багаторічну мінливість стоку. По –друге, коефіцієнти варіації річок з площею, меншою ніж друга критична, можуть суттєво відрізнятись від коефіцієнтів варіації великих річок [3]. У зв'язку з цим, для об'єднання за методом сумісного аналізу бажано використовувати дані по водозборах з площею більшою ніж 1000 км^2 (друга критична площа для більшості розглядуваних річок). Певний інтерес являє собою район 1(Карпати та Передкарпаття), де географічна й випадкова складові близькі одна до одної. У таких випадках при практичному застосуванні для водозборів із короткими рядами спостережень рекомендується використовувати не осереднені, а уточнені за формулою (1.9) значення коефіцієнтів варіації. Як випливає з табл.1.2, коефіцієнти варіації верхньої частини водозбору р.Дністер змінюються незначно в зоні надлишкового зволоження й достатнього зволоження, унаслідок чого райони 1 і 2 можливо об'єднати в один. Для району 3 (Середній Дністер) географічна складова набуває від'ємного значення, у зв'язку з чим береться рівною 0, у той час як випадкова складова дорівнює 100%. Для нижньої частини водозбору р.Дністер через близкість значень випадкової та географічної складових також бажано використовувати уточнені оцінки параметра Cv. Відносна похибка визначення осередненого значення параметра для усіх районів меньша за допустиму (15%).

Просторовий розподіл коефіцієнтів автокореляції визначається в більшій мірі не географічною зональністю, а внеском підземного живлення у формування стоку (табл. 1.3).

У гірській частині р.Дністер (зона надлишкового зволоження) та Причорномор'ї (зона недостатнього зволоження), де внесок підземного живлення слабко виражений, коефіцієнти автокореляції близькі до нуля. У межах Волино-Подільського артезіанського басейну (район 2), а також Північної і Центральної Молдови (район 3), де за рахунок близького залягання водоносних горизонтів і наявності карстових вод підземне живлення істотне, відзначаються високі значення коефіцієнтів автокореляції, які наближаються до 0,5. Для оцінки якості об'єднання коефіцієнтів автокореляції r(1) та відношення *Cs / Cv* використовується (4.7).

Слід зазначити, що за часів колишнього СРСР для території України було рекомендовано використовувати коефіцієнт автокореляції, який дорівнює 0,22. Але по результатах, наведених у табл.1.3, можна зробити висновок, що діапазон значень r(1) набагато ширший. Так, у межах України на основі методу сумісного аналіза для коефіцієнта автокореляції річного стоку виділено 7 районів, а для коефіцієнта асиметрії –10 [14].

Що стосується відношення Cs/Cv, то в зв'язку з великими похибками розрахунку коефіцієнта асиметрії за даними, перевірка гіпотези про можливість районування виконувалося в межах районів, виділених для коефіцієнта варіації (табл.4.2). У гірській зоні басейну р.Дністер відношення Cs/Cv можна взяти рівним 2 , на лівобережжі р.Дністер - 3, а в середній течії 1,5. У нижній частині басейну р.Дністер це відношення рекомендується брати рівним 1,7 (табл. 4.4).

Таблиця 4.2 – Результати застосування методу сумісного аналізу даних до обґрунтування способу просторового узагальнення коефіцієнтів варіації річного стоку річок в басейні р. Дністер

Район	Середнє значення *параметра Cv*	Дисперсія			ε_{Cv}, %	Внесок підземного живлення, %
		повна	випадкова складова	географічна складова		
Карпати та *Прикарпаття*	0,32	0,00283	0,00164 58%	0,00119 42%	11,1	30
Лівобережний Дністер до впадіння р.Марківка	0,31	0,00215	0,00206 96%	0,00009 4%	4,70	54
Середній Дністер (Північна та Центральна Молдова)	0,56	0,00512	0,0116 100%	-0,00649 0%	4,39	25
Нижній Дністер (Причорноморська низовина)	0,75	0,0470	0,0239 51%	0,0231 49%	12,4	5

Таблиця 4.3 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення коефіцієнтів автокорреляції річного стоку річок в басейні р. Дністер

Район	Середнє значення коэффіцієнта автокореляції $r(1)$	Дисперсія			Середнє квадратичне відхилення $\sigma_{r(1)}$
		повна	випадкова складова	географічна складова	
Карпати та Прикарпаття	0,149	0,0140	0,0381 100%	-0,0241 0%	0,035
Лівобережний Дністер до р.Марківка	0,479	0,0273	0,0194 71%	0,00799 29%	0,096
Середній р.Дністер (Північна та Центральна Молдова)	0,499	0,0149	0,00234 100%	-0,00878 0%	0,036
Нижній р.Днестер (Причорноморська низовина)	0,001	0,0680	0,111 100%	-0,00431 0%	0,096

Таблиця 4.4 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення відношення Cs/Cv річного стоку річок в басейні р.Дністер

Район	Середнє значення Cs/Cv	Дисперсія			$\sigma_{Cs/Cv}$
		повна	випадкова складова	географічна складова	
Карпати та Прикарпаття	2,0	2,01	1,86 93%	0,146 7%	0,350
Лівобережні притоки р. Дністер до р.Марківка	3,0	0,400	1,72 100%	-1,32 0%	0,456
Середній Дністер (Північна та Центральна Молдова)	1,5	0,398	0,873 100%	-0,474 0%	0,616
Нижній Дністер (Причорноморська низовина)	1,7	2,29	1,41 62%	0,880 38%	1,01

5 РЕГРЕСІЙНІ МОДЕЛІ У ГІДРОЛОГІЧНИХ РОЗРАХУНКАХ

5.1 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:

$$m_{y/x} = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x) \tag{5.1}$$

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

$$m_{x/y} = m_x + r_{xy}\frac{\sigma_x}{\sigma_y}(y - m_y), \tag{5.2}$$

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 (5.1) представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = ax_i + b, \tag{5.3}$$

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

$$\Delta = \sum_{i=1}^{n}\left[y_i - \tilde{y}(x_i)\right]^2 = min, \tag{5.4}$$

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

$$\frac{\partial\Delta(a,b)}{\partial a} = \frac{\partial\left[\left(\sum_{i=1}^{n} y_i - ax_i - b\right)^2\right]}{\partial a} = 0; \tag{5.5}$$

$$\frac{\partial\Delta(a,b)}{\partial b} = \frac{\partial\left[\left(\sum_{i=1}^{n} y_i - ax_i - b\right)^2\right]}{\partial b} = 0 \tag{5.6}$$

Розв'язуючи рівняння (5.5) та (5.6) відносно a та b, одержуємо

$$a = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2}; \tag{5.7}$$

$$b = \bar{y} - a\bar{x}, \tag{5.8}$$

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

Чисельник дробу, який знаходиться в правій частині рівняння (5.7), є оцінкою коваріації (коваріаційного моменту) $\hat{K}_{xy}$, розрахованого за дискретною вибіркою завдовжки n

$$\hat{K}_{xy} = \hat{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \tag{5.9}$$

а знаменник – оцінкою дисперсії випадкової величини X

$$\hat{\sigma}_x^2 = S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2, \qquad (5.10)$$

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

$$\hat{r}_{xy} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \qquad (5.11)$$

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 позначених як і S_y та S_x

$$a = r\frac{S_y}{S_x}. \qquad (5.12)$$

Математична модель множинної лінійної регресії представляється рівнянням вигляду

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2) + b_3(x_{3i} - \bar{x}_3)... + b_k(x_{ki} - \bar{x}_k) \qquad (5.13)$$

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

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

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

k - число предикторів.

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 вигляд

$$b_j = \frac{\sigma}{\sigma_j}\frac{D_{0j}}{D_{00}}, \qquad (5.14)$$

де σ - оцінка середнього квадратичного відхилення досліджуваної характеристики y;

σ_j - оцінка середнього квадратичного відхилення j-того предиктора;

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 предикторами r_{ij} і коефіцієнти парної кореляції r_{oj} між предиктантом і предикторами. При цьому визначник розширеної матриці кореляцій другого порядку записується у вигляді

$$D = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{vmatrix}, \qquad (5.15)$$

а мінори цього визначника записуються таким чином

$$D_{00} = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}, \qquad (5.16)$$

$$D_{01} = \begin{vmatrix} r_{10} & r_{12} \\ r_{20} & r1 \end{vmatrix}, \qquad (5.17)$$

$$D_{02} = \begin{vmatrix} r_{10} & 1 \\ r_{20} & r_{21} \end{vmatrix} \qquad (5.18)$$

У записах вигляду (5.15-5.18) r_{0j} - коефіцієнт кореляції між предиктантом (0) та предиктором (j).

Рівняння лінійної множинної регресії для двох предикторів буде мати вигляд

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2), \qquad (5.19)$$

де

$$b_1 = \frac{\sigma}{\sigma_1}\frac{D_{01}}{D_{00}} \qquad (5.20)$$

$$b_2 = \frac{\sigma}{\sigma_2}\frac{D_{02}}{D_{00}} \qquad (5.21)$$

Рівняння (5.19) може бути записано і таким чином

$$\tilde{y}_i = b_1 x_{1i} + b_2 x_{2i} + b_0, \qquad (5.22)$$

де

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2, \qquad (5.23)$$

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

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

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

$$\sigma_{\tilde{y}} = \sigma\sqrt{1 - R^2}, \qquad (5.24)$$

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

$$R = \sqrt{1 - \frac{D}{D_{00}}}, \qquad (5.25)$$

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.

5.2 Оцінка адекватності регресійної моделі за складовими дисперсії випадкової величини

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

$$\sigma_y^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}. \qquad (5.26)$$

В основі формули (5.26) лежить відхилення спостереженої величини y_i від середнього арифметичного значення (рис. 5.1).

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

$$y_i - \bar{y} = (y_i - \tilde{y}_i) + (\tilde{y}_i - \bar{y}), \text{ або } y_i - \bar{y} = (\tilde{y}_i - \bar{y}) + (y_i - \tilde{y}) \qquad (5.27)$$

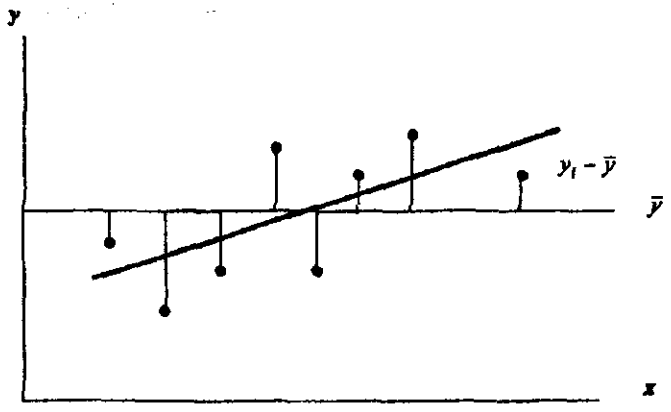

Рисунок 5.1 – Ілюстрація розсіювання спостережених значень y_i від $\bar{y}$

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

Обидві частини рівняння піднесемо до квадрата

$$(y_i - \bar{y})^2 = [(y_i - \bar{y}) + (y_i - \tilde{y}_i)]^2. \qquad (5.28)$$

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

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i) + \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad (5.29)$$

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

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-1} \qquad (5.30)$$

Рівняння (5.30) представляє собою суму дисперсій

$$\sigma_y^2 = \sigma_p^2 + \sigma_{зал}^2 \qquad (5.31)$$

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 представлено на рисунке 5.2.

Величина $(y_i - \tilde{y}_i)$ характеризує розсіювання точок, які відповідають даним спостережень, від значень, розрахованих за рівнянням лінійної регресії (рис. 5.3).

Якщо як розрахункова модель розглядається рівняння лінійної парної регресії, то гіпотеза про адекватність обраної моделі перевіряється за критерієм Фішера, який формується таким чином

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 / n-1}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n-1} = \frac{\sigma_y^2}{\sigma_{зал}^2}. \qquad (5.32)$$

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

$$F < F_{кр}(\alpha, \nu_1, \nu_2), \qquad (5.33)$$

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

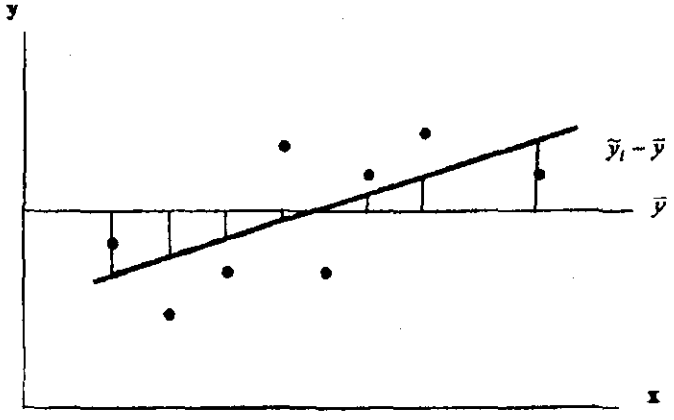

Рисунок 5.2 – Ілюстрація відхилення лінії регресії від $\bar{y}$

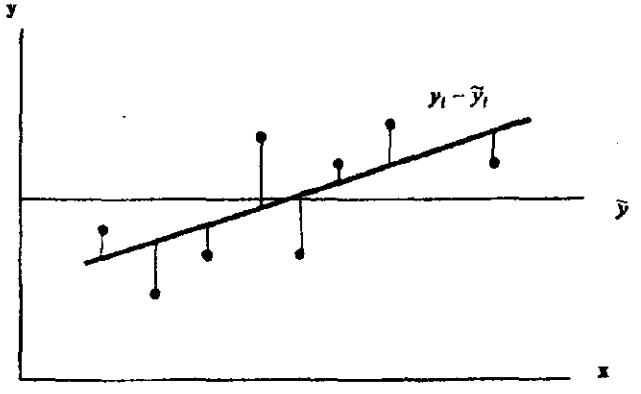

Рисунок 5.3 – Ілюстрація відхилення спостережених даних від лінії регресії

Коли ж мова йде про множинну регресію, то залишкова дисперсія - це та частина дисперсії вихідної величини Y, яка не описується залежністю предиктанта Y від предикторів $X_1, X_2, ..., X_k$. За аналогією з (5.32) перевірка гіпотези про адекватність моделі множинної лінійної регресії даним спостережень здійснюється за критерієм Фішера, який визначається за формулою

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 /(n-1)}{\sum_{i=1}^{n-k}(y_i - \tilde{y}_i)^2 /(n-k-1)}, \qquad (5.34)$$

де n - об'єм вибірок;
k - кількість предикторів;
$y_i, \tilde{y}_i$ - 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 предикторів, які відображають вплив основних стокоформуючих чинників.

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

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

Вибір оптимальних предикторів називають операцією "просіювання".

Просіювання може відбуватися за допомогою частинних коефіцієнтів кореляції.

5.3 Визначення частинних коефіцієнтів кореляції

Наведемо 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* $r_{yx_2 \bullet x_1}$ [28].

Будемо вважати, що нам відома матриця кореляцій

$$R_x = \begin{vmatrix} 1 & r_{x_2} \\ r_{x_2 x_1} & 1 \end{vmatrix} \tag{5.35}$$

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

$$R_{yx} = \begin{vmatrix} r_{yx_1} \\ r_{yx_2} \end{vmatrix} \tag{5.36}$$

На їх основі сформуємо розширену матрицю кореляцій

$$\widetilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1 x_2} \\ r_{yx_2} & r_{x_1 x_2} & 1 \end{vmatrix} \tag{5.37}$$

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, викреслюються перші рядок і стовпець. Очевидно:

$$D_{yx_1} = r_{yx_1} - r_{yx_2} r_{x_1 x_2}, \tag{5.38}$$

$$D_{yx_2} = r_{yx_2} - r_{yx_1} r_{x_1 x_2} \tag{5.39}$$

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

$$D_{\overline{yx_1}} = \begin{vmatrix} 1 & r_{yx_2} \\ r_{yx_2} & 1 \end{vmatrix} = 1 - r_{yx_2}^2, \tag{5.40}$$

$$D_{\overline{yx_2}} = \begin{vmatrix} 1 & r_{yx_1} \\ r_{yx_1} & 1 \end{vmatrix} = 1 - r_{yx_1}^2 \tag{5.41}$$

Частинні коефіцієнти кореляції визначаються таким чином:

$$r_{yx_1 \bullet x_2} = \frac{D_{yx_1}}{\sqrt{|R_x| D_{\overline{yx_1}}}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_2}^2\right)}}, \tag{5.42}$$

$$r_{yx_2 \bullet x_1} = \frac{D_{yx_2}}{\sqrt{|R_x| D_{\overline{yx_2}}}} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} \tag{5.43}$$

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: $r_{yx_1} = 0.72; r_{yx_2} = 0.91; r_{x_1x_2} = 0.84$. 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 (5.42) і (5.43).

Розрахунки дають такі їхні 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 врахувати 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

$$\widetilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \dots & r_{yx_k} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} & \dots & r_{x_2x_n} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} & \dots & r_{x_3x_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 & \dots & r_{x_kx_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & r_{x_nx_3} & \dots & r_{x_nx_k} & \dots & 1 \end{vmatrix} \qquad (5.44)$$

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

$$r_{yx_k \bullet x_1, x_2, \dots, x_{k-1}, x_{k+1}, \dots x_m} = \frac{D_{yx_k}}{\sqrt{|R_x| D_{\overline{yx_k}}}}, \qquad (5.45)$$

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 предикторами.

Вибір оптимальних предикторів виконується за наступною схемою.

1. З числа предикторів вибирається той, який має найтіснішій зв'язок з предиктантом і йому привласнюється номер 1.

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

3. З числа частинних коефіцієнтів кореляції, які характеризують зв'язок між предиктантом 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 частинні 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 $\upsilon = m - k$.

5.4 Приклад аналізу розрахунків за моделлю множинної лінійної регресії з покроковим добором предикторів

Розглянуто 40 водозборів у басейні р. Уссурі (табл. 5.1). До розрахунків залучений пакет статистичних програм „Microstat".

Необхідно добрати оптимальні предиктори та отримати розрахункове рівняння множинної лінійної регресії для визначення середньобагаторічної величини річного стоку невивчених у гідрологічному відношенні водозборів басейну р. Уссурі. Як потенційні предиктори 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 предиктантом та усіма предикторами. Обирається предиктор, який має найбільш тісний зв'язок з предиктантом $(\overline{q})$. У розглянутому випадку таким предиктором визнається умовна довгота λ : $r_{\overline{q}\lambda} = 0{,}692$.

Таблиця 5.1 – 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
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

Продовження таблиці 5.1

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
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

$$\bar{q} = 0.440\lambda + 8.03; \qquad \sigma_{\tilde{q}} = 1.61 \tag{5.46}$$

Підраховуються повна дисперсія предиктанта й регресійна та залишкова складові повної дисперсії

$$\sigma^2_{\bar{q}} = \frac{\sum_{i=1}^{n}(q_i - \bar{q})^2}{n-1} = \frac{\sum_{i=1}^{40}(q_i - \bar{q})^2}{39} = \frac{189}{39} = 4.85 \tag{5.47}$$

$$\sigma^2_p = \frac{\sum_{i=1}^{n}(\tilde{q}_i - \bar{q})^2}{n-1} = \frac{90.4}{39} = 2.32 \tag{5.48}$$

$$\sigma^2_{зал} = \frac{\sum_{i=1}^{n}(q_i - \tilde{q}_i)^2}{n-1} = \frac{98.5}{39} = 2.51 \tag{5.49}$$

Підраховується критерій Фішера за (5.34)

$$F = \frac{189/39}{98.5/38} = \frac{4.85}{2.59} = 1.87 , \tag{5.50}$$

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

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

Надалі розраховуються частинні коефіцієнти кореляції між предиктантом $\bar{q}$ та предикторами, які не увійшли до рівняння (5.46):

$$\begin{aligned} &r_{\bar{q},lg(F+1)\bullet\lambda} = 0.16; \\ &r_{\bar{q},\bar{x}\bullet\lambda} = 0.44; \\ &r_{\bar{q},H\bullet\lambda} = 0.40; \\ &r_{\bar{q},f_б\bullet\lambda} = 0.45; \\ &r_{\bar{q},f_л\bullet\lambda} = 0.46; \\ &r_{\bar{q},\varphi\bullet\lambda} = 0.26 \end{aligned} \tag{5.51}$$

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

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

$$\bar{q} = 0.105 f_л + 0.420\lambda - 1.66; \qquad \sigma_{\tilde{q}} = 1.44; \quad R = 0.768 \tag{5.52}$$

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 набуває значення

$$F = \frac{189/39}{77.3/37} = \frac{4.85}{2.09} = 2.32 \qquad (5.53)$$

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 предиктантом $\bar{q}$ та предикторами, які не увійшли до рівняння (5.52)

$$\begin{aligned} r_{\bar{q},lg(F+1)\bullet\lambda,f_л} &= 0.25; \\ r_{\bar{q},\bar{X}\bullet\lambda,f_л} &= 0.27; \\ r_{\bar{q},H\bullet\lambda,f_л} &= 0.13; \\ r_{\bar{q},f_\delta\bullet\lambda,f_л} &= 0.14; \\ r_{\bar{q},\varphi\bullet\lambda,f_л} &= 0.00 \end{aligned} \qquad (5.54)$$

Найбільший частинний коефіцієнт кореляції за умови, що вплив довготи та залісеності врахований, 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 , необхідно перевірити його значущість за формулами

$$t = \frac{r_{xy}}{\sqrt{\sigma^2_{r_{xy}}}}, \qquad (5.55)$$

$$\sigma_{r_{xy}} = \frac{1 - r^2_{xy}}{\sqrt{n-1}}, \qquad (5.56)$$

де t - критерій Стьюдента.

$$\sigma_{r_{\bar{q},\bar{x}\bullet\lambda,f_л}} = \frac{1-0.27^2}{\sqrt{40-1}} = \frac{1-0.0729}{6.24} = \frac{0.927}{6.24} = 0.149 \qquad (5.57)$$

$$t = \frac{0.27}{0.149} = 0.85 \qquad (5.58)$$

Критичне значення t для рівня значущості 0.05 та числа степенів вільності 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 предикторів 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 береться 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 5.2 – 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

продовження таблиці 5.2

N	$\bar{q}_i$	$\tilde{q}_i$	$\Delta\bar{q} = \bar{q}_i - \tilde{q}_i$	$\delta_i = \frac{\Delta\bar{q}_i}{\bar{q}_i} \cdot 100\%$
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
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

6 ДИСКРИМІНАНТНИЙ АНАЛІЗ ЯК МЕТОД ПРИЙНЯТТЯ АЛЬТЕРНАТИВНИХ РІШЕНЬ ПРИ ВИРІШЕННІ ЗАДАЧ РАЙОНУВАННЯ

6.1 Схема побудови розв'язувального правила

Дискримінантний аналіз – це статистичний метод, який дозволяє вивчати різницю між двома або більше групами об'єктів по декількох змінних одночасно.

У гідрології часто виникає задача щодо віднесення того чи іншого об'єкта або спостереження до одного з відомих класів. Такими класами можуть бути різні гідрологічні райони, до яких має бути віднесений той чи інший об'єкт. В практиці гідрологічного прогнозування часто виникає потреба складати прогноз здійснення або нездійснення того чи іншого гідрологічного явища. Такий прогноз називають альтернативним. Поставлені задачі називаються задачами прийняття альтернативних рішень або задачами класифікації. Питання про віднесення об'єкта до тієї чи іншої сукупності (групи) вирішується у такому випадку шляхом порівнювання ознак, характерних для кожної із розглядуваних сукупностей, із ознаками самого розглядуваного об'єкта.

У теорії розпізнавання образів задача класифікації формулюється таким чином: на основі відомостей про окремих представників різних класів навчаючої системи із характерними для них ознаками (предикторами) необхідно знайти вирішальне правило, за яким той чи інший об'єкт може бути віднесеним до одного з класів.

Сформульована задача є типовою задачею розпізнавання образів. Суть її полягає у тому, що, по-перше, необхідно поділити весь простір образів на два підпростори, у першому з яких явище відбувається, а в другому – ні. По-друге, треба побудувати правило, за допомогою котрого можна віднести образ, який підлягає розпізнаванню, до того чи іншого з підпросторів.

Нехай ми маємо множину V векторів-предікторів (образів), що складають простір зображень Rv. Припустимо, що цей простір поділяється на два підпростори R_{v1} і R_{v2}. У першому з них розташовується множина V_1 образів X, при яких явище відбувається, а у другому - множина V_2 образів X, коли явище не відбувається. Ясно, що

$$V_1 \cup V_2 = V, V_1 \cap V = 0 \qquad (6.1)$$

Насамперед, як зазначалося вище, треба побудувати поверхню, яка б

розділяла підпростори R_{V1} і R_{V2}. Наведемо для пояснення прості приклади.

Нехай простір буде двовимірним $R_V = R_V(x_1, x_2)$ (рис.6.1). Тоді ми маємо на площині (x_1, x_2) лінію $x_1 = f(x_2)$, що відділяє підпростір R_{V1} від підпростору R_{V2}. Досить простим буде і випадок трьохвимірного простору

$R_V = R_V(x_1, x_2 x_3)$ (рис. 6.2). У цьому випадку підпростори R_{V1} і R_{V2} розділяє деяка поверхня у тривимірному просторі, рівняння якої має вигляд $x_3=\varphi(x_1, x_2)$.

Більш складні умови виникають, коли розглядаються образи із багатовимірного простору $R_V = R_V(x_1,x2,\ldots,x_n)$. Поверхня, що поділяє цей простір на підпростори R_{V1} і R_{V2} називається роздільною гіперповерхнею, а її рівняння має вигляд:

$$F(x_1,x2,\ldots,x_n) = 0. \qquad (6.2)$$

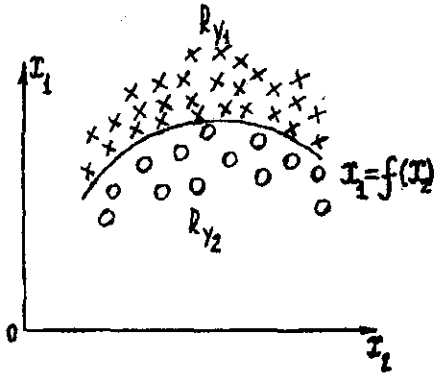

Рисунок 6.1 - Образи і роздільна функція в двовимірному просторі [28]

Надалі отримується правило, за допомого якого є підстава віднести вектор X, що підлягає розпізнаванню, до підпростору R_{V1} або підпростору R_{V2}. Це правило називають розв'язувальним правилом. Якщо відповідно до нього приймається рішення, що $X \in R_{V1}$, то явище прогнозується, якщо приймається рішення, що $X \in R_{V2}$, то явище не прогнозується. Етап,

який складається з побудови розділяючої гіперповерхні та розв'язувального правила, носить назву етапу навчання.

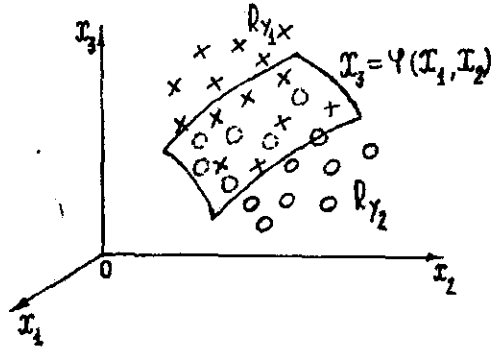

Рисунок 6.2.-Образи та роздільна поверхня в тривимірному просторі [28]

Прийняття рішення про належність вектора X до підпросторів R_{V1} чи R_{V2} називають етапом розпізнавання. Множина V векторів-*предикторів, на основі якої реалізуються перелічені етапи*, називається навчаючою сукупністю. Крім неї, створюється ще й перевірна сукупність, яка використовується для перевірки адекватності моделі альтернативного прогнозу. Саме розв'язувальне правило може бути представлено у вигляді деякої математичної функції, яку назвемо дискримінантною.

Функції, які забезпечують можливість віднесення об'єкта, що підлягає класифікації, до однієї з виділених груп, називають дискримінантними.

Застосування дискримінантного аналізу, з одного боку, передбачає принципову роздільність класів, а з іншого боку, допускає існє часткове "перекривання" (рис. 6.3). Таким чином, рішення про віднесення явища або об'єкта до того чи іншого класу приймається з тією чи іншою долею похибки. "Перекривання" класів виражається у тому, що інтервал числових значень ознак у об'єктів, які належать до різних класів, перекривається. Це утруднює віднесення об'єкта або явища до визначеного класу, якщо їхня кількісна ознака лежить у області перекриття. Наприклад, виділені два гідрологічних райони, які відрізняються один від одного кількісними характеристиками стоку. Але розподіл стокових величин у просторі підпорядковується закону географічної зональності. Отже існують "приграничні" водозбори, на яких розглядувані характеристики

можуть попадати у область "перекриття". Коли класифікація виконується за однією з ознак, то водозбір буде віднесений до першого класу (району А), якщо ознака x буде меншою за x_0, і до другого (району В), якщо $x > x_0$. Неправильні рішення позначимо через δ_a та δ_b. Як x_0 можна вибрати середину інтервалу перекриття або абсцису точки перетину кривих щільності розподілу x_0. Проте більш доцільно вибрати таку точку x_0^*, для якої буде виконуватись $\delta_a = \delta_b$. За цієї умови максимальна похибка (найбільша з двох) має мінімальне значення. Таким чином, коли розглядається одна ознака, то задача зводиться до побудови точки, яка розділяє дві сукупності. Результат класифікації визначається відносно цієї точки. Як уже відзначалося, при використанні двох ознак розділ сукупностей відбувається відносно прямої, трьох – відносно поверхні, і т.д.

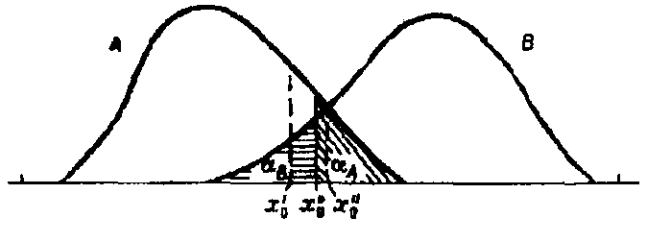

Рисунок 6.3 – Вирішальні правила при розмежуванні одновимірних сукупностей

Розглянемо основні ідеї теорії розпізнавання образів [28]. Позначимо через H_1 гіпотезу, що образ $X \in V_1$. Альтернативною буде гіпотеза H_2, про те, що $X \in V_2$. Задача розпізнавання полягає у тому, що треба знайти правило, яке дозволяє обгрунтовано прийняти гіпотезу H_1 або H_2. Всіляка процедура перевірки гіпотез передбачає, що приймаючи те чи інше рішення, ми можемо припустити помилку 1-го чи 2-го роду. Нагадаємо, що помилку 1-го роду ми припускаємо, коли відкидаємо правильну гіпотезу. Помилка 2-го роду пов'язана з прийняттям невірної гіпотези. Помилку другого роду ще називають "похибкою хибної тривоги".

Будемо вважати, що відомими є умовні **ймовірності класів** V_1 і V_2:

$$P(x_1, x_2, ..., x_n / V_1) \tag{6.3}$$

та

$$P(x_1, x_2, ..., x_n / V_2) \tag{6.4}$$

Позначимо ймовірність помилки 1-го роду через P_a, а 2-го роду через P_b. Знаючи ймовірності (6.3) та (6.4), а також апріорні ймовірності $P(V_1)$ і $P(V_2)$ класів V_1 і V_2, можна розрахувати ймовірності помилок 1-го й 2-го роду.

Розв'язувальне правило або дискримінантна функція будується на основі функції подібності

$$\lambda(x_1, x_2, ..., x_n) = \frac{P(x_1, x_2, ..., x_n / V_1)}{P(x_1, x_2, ..., x_n / V2)}, \tag{6.5}$$

а величину

$$\frac{\delta_b P(V_2)}{\delta_a P(V_1)} = \theta \tag{6.6}$$

називають порогом.

Таким чином, ми прийшли до такого розв'язувального правила:

вектор $X \in V_1$, якщо $\lambda(x_1, x_2, ..., x_n) > \theta$ (6.7)

вектор $X \in V_2$, якщо $\lambda(x_1, x_2, ..., x_n) < \theta$. (6.8)

Якщо є підстави вважати, що $\delta_a = \delta_b$ і $P(V_1) = P(V_2)$, то $\theta = 1$ й розв'язувальне правило має вигляд:

$$\begin{aligned} &\text{вектор } X \in V_1, \text{ якщо } \lambda(x_1, x_2, ..., x_n) > 1, \\ &\text{вектор } X \in V_2, \text{ якщо } \lambda(x_1, x_2, ..., x_n) < 1 \end{aligned} \tag{6.9}$$

Отже розв'язувальне правило базується на нерівності

$$\frac{P(x_1, x_2, ..., x_n / V_1)}{P(x_1, x_2, ..., x_n / V_2)} > \frac{\delta_b P(V_2)}{\delta_a P(V_1)}, \tag{6.10}$$

яку можна представити у логарифмічному вигляді

$$\ln\frac{P(x_1,x_2,....,x_n/V_1)}{P(x_1,x_2,...,x_n/V_2)} > \ln\frac{\delta_b P(V_2)}{\delta_a P(V_1)} \qquad (6.11)$$

Функцію вигляду

$$F(x_1,x_2,...,x_n)=\ln P(x_1,x_2,...,x_n/V_1)-\ln P(x_1,x_2,...,x_n/V_2)+\frac{\delta_a P(V_1)}{\delta_b P(V_2)} \qquad (6.12)$$

називають дискримінантною функцією.

Якщо використовується дискримінантна функція, то розв'язувальне правило набуває вигляду:

$$X \in V_1, якщо F(x_1,x_2,...,x_n)>0; \qquad (6.13)$$

$$X \in V_2, якщо F(x_1,x_2,...,x_n)<0 \qquad (6.14)$$

Ясно, що рівняння $F(x_1,x_2,...,x_n)=0$ є рівнянням роздільної поверхні для підпросторів R_{V1} і R_{V2}.

Методи, основані на теорії статистичних рішень, мають такі обмеження: для їхньої реалізації необхідно знати щільності умовних розподілів образів у класах V_1 і V_2. Ці закони розподілів є багатовимірними, і на основі множини векторів-предикторів класів V_1 і V_2, одержання їхнього аналітичного вигляду - це дуже складна задача. Тому вважають, що вид законів розподілу є відомим. У такому разі задача зводиться до необхідності на основі вибірок векторів-предікторів отримати оцінки параметрів цих законів. Ця процедура носить назву відновлення закону розподілу. При практичних реалізаціях цих методів найбільш часто вважають, що класи векторів-предикторів підпорядковуються умовним нормальним законам розподілу. В дійсності ці припущення строго не виконуються. Дуже добре відомо, що для багатьох метеорологічних величин, які виступають у ролі предикторів, нормальний закон розподілу не виконується. Але, як показує досвід, це не вносить суттєвих похибок, якщо більшість предикторів має одномодальний розподіл. Ця умова у більшості випадків виконується [28].

6.2 Побудова розв'язувального правила на основі *багатовимірного нормального розподілу*

Будемо вважати, що вектори-предиктори

$$X_j=\begin{pmatrix} x_{1j} \\ x_{2j} \\ ... \\ x_{nj} \end{pmatrix} \qquad (6.15)$$

підпорядковуються багатовимірному нормальному розподілу. Параметрами його, як відомо, є вектори математичних сподівань

$$\mu=\begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix} \qquad (6.16)$$

і матриці коваріацій розміром $n \times n$. Тому щільності нормальних умовних розподілів для класів V_1 і V_2 мають вигляд:

$$P(x_1,x_2,...,x_n/V_1)=\frac{1}{(2\pi)^{n/2}|K_1|^{1/2}}\exp\left[-\frac{1}{2}(X-\mu_1)K_1^{-1}(X-\mu_1)\right], \qquad (6.17)$$

$$P(x_1,x_2,...,x_n/V_2)=\frac{1}{(2\pi)^{n/2}|K_2|^{1/2}}\exp\left[-\frac{1}{2}(X-\mu_2)K_2^{-1}(X-\mu_2)\right], \qquad (6.18)$$

де μ_1,μ_2,K_1,K_2 - вектори математичних сподівань і матриці коваріацій для першого та другого класів.

Після підстановки (6.17) і (6.18) до дискримінантної функції (6.12) прийдемо до такого рівняння:

$$F(x)=-\frac{n}{2}\ln 2\pi-\frac{1}{2}\ln|K_1|-\frac{1}{2}(X-\mu_1)'K_1^{-1}(X-\mu_1)+\frac{n}{2}\ln 2\pi+$$
$$+\frac{1}{2}\ln|K_2|+\frac{1}{2}(X-\mu_2)'K_2^{-1}(X-\mu_2)+\ln\frac{P(V)}{P(V_2)} \qquad (6.19)$$

Після скорочень дискримінантна функція набуде вигляду

$$F(x)=\frac{1}{2}\left[(X-\mu_2)K_2^{-1}(X-\mu_2)-(X-\mu_1)K_1^{-1}(X-\mu_1)+\ln\frac{|K_2|}{|K_1|}\right] \qquad (6.20)$$
$$+\ln\frac{P(V_1)}{P(V_2)}$$

Вважається, що ціни помилок першого і другого роду однакові $\delta_a=\delta_b$. Дискримінантна функція F(x), яка визначається формулою (6.20), носить назву квадратичної дискримінантної функції. Така назва пов'язується з тим, що перші два члени у квадратних дужках являють собою квадратичні форми, тобто багаточлени степеня не більше другого.

Дискримінантна функція (6.20) утримує операції обернення коваріаційних матриць K_1 і K_2. Ця операція може призвести до негативних наслідків, коли матриці коваріацій погано обумовлені. В такому разі похибки, що містяться в коваріаціях, можуть призвести до великих похибок коефіцієнтів квадратичних форм і, таким чином, до помилок на етапі розпізнавання. У ряді випадків ці похибки можуть бути більшими ніж ті похибки, які ми робимо, приймаючи умову:

$$K_1=K_2=K\ , \qquad (6.21)$$

тобто вважаючи, що матриці коваріацій двох класів однакові. Частіше за все приймається умова:

$$K=\frac{K_1+K_2}{2} \qquad (6.22)$$

Якщо прийняти умову (6.22) у дискримінантній функції (6.20) і вважати, що $P(V_1)=P(V_2)$, то отримаємо:

$$F(x)=\frac{1}{2}\left[(X-\mu_2)'K^{-1}(X-\mu_2)-(X-\mu_1)'K^{-1}(X-\mu_1)\right]=$$
$$=(\mu_1-\mu_2)'K^{-1}X+\frac{1}{2}\left[\mu_2'K^{-1}\mu_2-\mu_1'K^{-1}\mu_1\right] \qquad (6.23)$$

Дискримінантна функція (6.23) є лінійною дискримінантною функцією.

Використання дискримінантного аналізу значно спрощується, якщо є підстави вважати коваріації рівними нулю. Це можна зробити, якщо всі недіагональні елементи матриць коваріацій класів V_1 і V_2 значно менші від діагональних, тобто дисперсій, і ними можна знехтувати. Тоді

$$K=\begin{pmatrix}\sigma_1^2 & 0 & \dots & 0\\ 0 & \sigma_2^2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \sigma_n^2\end{pmatrix}\ , \qquad (6.24)$$

а її обернена матриця

$$K^{-1}=\begin{pmatrix}\frac{1}{\sigma_1^2} & 0 & \dots & 0\\ 0 & \frac{1}{\sigma_2^2} & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \frac{1}{\sigma_n^2}\end{pmatrix}\ , \qquad (6.25)$$

тобто операція обернення матриць коваріацій значно спрощується. При цьому спрощується й процедура розрахування коефіцієнтів квадратичних форм у квадратичній дискримінантній функції (6.12).

Коли виконується умова (6.21) і, крім того, можна вважати коваріаційну матрицю діагональною, значно спрощується вигляд лінійної дискримінантної функції (6.23). У цьому випадку вона дорівнює :

$$F(x)=\sum_{i=1}^{n}\frac{\mu_{1i}-\mu_{2i}}{\sigma_i^2}x_i+\sum_{i=1}^{n}\frac{\mu_{2i}^2-\mu_{1i}^2}{\sigma_i^2} \qquad (6.26)$$

Звичайно, при практичному використанні розглянутих дискримінантних функцій замість складових векторів математичних сподівань і дисперсій предикторів використовують їхні статистичні оцінки, тобто середні значення і вибіркові дисперсії предикторів.

Критерієм якості проведення роздільної поверхні є число Махаланобіса, яке характеризує відстань між центрами класів. При побудові лінійної дискримінантної функції число Махаланобіса визначається за матричним рівнянням

$$\Delta = (\mu_1 - \mu_2)' K^{-1} (\mu_1 - \mu_2), \qquad (6.27)$$

а при використанні (6.26)

$$\Delta = \sum_{i=1}^{n} (\mu_{1i} - \mu_{2i})^2 / \sigma_i^2. \qquad (6.28)$$

Чим більше число Махаланобіса, тим меньше ймовірність похибки класифікації. При Δ=11 ймовірність похибки класифікації досягає 5%.

6.3 Приклад застосування лінійної дискримінантної функції до гідрологічного районування

Розглянемо водозбори лівобережжя річки Дністер. Більшість статистичних параметрів річок Поділля, що впадають у р.Дністер з лівого боку значно відрізняється від статистичних параметрів річного стоку лівобережних річок, які впадають у р.Дністер нижче Могильова-Подільського.

Перша задача, яка стоїть перед дослідниками, - це задача інтерпретації. При вирішенні цієї задачі необхідно дати відповідь на питання: чи можливо, використовуючи заданий набір змінних, виділити позицію між районами.

Друга задача – це задача класифікації. Тобто, необхідно одержати класифікаційне правило, щоб уточнити межі між районами.

Виділені ознаки (предиктори): середня висота водозборів $H_{сер}$, ухил I, заболоченість $f_б$, залісеність $f_л$.

До першого класу (район А) увійшли 7 водозборів. До другого (район В) – 18 водозборів.

Коваріаційні матриці мають вигляд:

Коваріаційна матриця класу А

	$H_{сер}$	I	$f_б$	$f_л$
$H_{сер}$	428	401	-19.0	134
I	401	913	-35.1	131
$f_б$	19.0	-35.1	4.81	-2.4
$f_л$	135	131	-2.40	49.6

Коваріаційна матриця класу В

	$H_{сер}$	I	$f_б$	$f_л$
$H_{сер}$	505	777	-65.2	196
I	777	315	-11.6	458
$f_б$	65.2	-11.6	3.08	-0.41
$f_л$	196	458	-0.41	29.0

Виходячи з аналізу коваріаційних матриць, можемо прийняти умову

$$K = K_1 = K_2$$

або

$$K = \frac{K_1 + K_2}{2}$$

Таким чином, до виведення розв'язувального правила можна застосувати лінійну дискримінантну функцію.

Після застосування пакету стандартних програм „Statistic", отримуємо функцію

$$F(x) = -9.52 - 0.001 H_{сер} + 0.01 I + 1.14 f_б + 0.37 f_л$$

Число Махаланобіса дорівнює $\Delta = 11.1$.

Таким чином, можна зробити висновок, що отримане розв'язувальне правило можна застосувати до описування різниці між двома районами.

Вирішимо задачу класифікації. Розглянемо водозбір р. Стрвяж – Луки, який не увійшов до навчальної вибірки даних.

Класифікаційні ознаки цього водозбору

$$H_{сер} = 500\text{м};\quad I = 156\%;\quad f_б = 0/5\%;\quad f_л = 44\%$$

Отже, водозбір р.Стрвяж – с.Луки належить до району А.

7 САМОПОДІБНІСТЬ (ФРАКТАЛИ) У ПРОСТОРОВОМУ РОЗПОДІЛІ СТАТИСТИЧНИХ ПАРАМЕТРІВ

7.1 Поняття про фрактали

Термін "фрактал" був введений Мандельбротом, співробітником дослідницького центру імені Томаса Дж.Уотсона корпорації IBM в Йорктаун-Хейтсі (шт.Нью-Йорк) і походить від латинського слова fractus, що означає ламати, розбивати. Більшість структур мають фундаментальну властивість масштабної регулярності, відомої як інваріантність по відношенню до масштабу або властивість "самоподібності". Іншими словами, якщо розглядати об'єкти в різному масштабі, то постійно виявляються одні й ті ж фундаментальні елементи (фрактали). Важливо підкреслити, що спочатку фрактали застосовувалися як своєрідна мова геометрії. Проте, на відміну від звичних об'єктів евклідової геометрії (пряма лінія, коло тощо) вони не можуть бути безпосередньо спостереженими. Фрактали виражаються не в первинних геометричних формах, а в алгоритмах, наборах математичних процедур. Саме їхній пошук й обгрунтовування є центральною задачею сучасної теорії фракталів.

Більшість природних процесів, не дивлячись на те, що вони відповідають певним детерміністичним законам, на достатньо великих часових інтервалах є непередбачуваними і проявляють схожі закономірності у варіаціях в різних часових масштабах подібно до того, як об'єкти, що мають інваріантність у просторових масштабах, проявляють схожі структурні закономірності в просторі.

Незалежно від природи або методу побудови у всіх фракталів є одна важлива загальна властивість: ступінь складності їхньої структури може бути зміряний якимось характеристичним числом - фрактальною розмірністю. Таким чином, фрактал є математичним об'єктом, що має дробову розмірність на відміну від традиційних математичних фігур цілої розмірності. Визначення фрактальної розмірності може відбуватися різними методами. В найпростішому випадку, якщо множина розбивається на N підмножин, кожна з яких в k раз менше від всієї множини, то фрактальна розмірність дорівнює $d = \frac{\ln N}{\ln k}$. Проте, в більшості випадків масштабні множники неоднорідні, тобто у фракталів є цілий спектр скейлингів. Такі фрактали називаються мультифракталами і характеризуються спектром розмірностей.

Фрактали надають можливість надзвичайно компактного способу опису об'єктів і процесів. Фрактальний підхід до опису різних *гідрологічних величин останнім часом активно розвивається в прикладній* гідрології. Ідея масштабної самоподібності топографії земної поверхні, вперше показана в роботах Мандельброта, знайшла свій розвиток в роботах, де розглядається гідравліко-геометрична подібність річкових систем. Наприклад, фрактальна природа річкових систем може бути описаною таким степеневим рівнянням (Mandelbrot,1977)

$$L^{1/d} \approx A^{0.5}, \tag{7.1}$$

де L - довжина річки уздовж продольної осі;
A - площа водозбору;
d - фрактальна розмірність.

Зміна величини d може бути основою до районування річкових систем.

Серед російських і українських вчених, що вивчають мультифрактальні властивості гідрометеорологічних об'єктів, слід відзначити А.Г. Бершадського [2] – дослідника турбулентних структур, С.С. Іванова [8], який розглядає фрактальні властивості глобального рельєфу. Сучасний погляд на генезис фрактальних розмірностей турбулентних пульсацій в атмосфері, представлений ученими В.Д. Русовим і О.В. Глушковим.

Важливо відзначити, що фрактальні розмірності несуть в собі певну інформацію про просторово-часовий розподіл досліджуваних величин, *подібно до статистичних параметрів, які широко застосовуються в* гідрології.

Якщо часові ряди стаціонарні, то найпростішим способом їхнього масштабування є стандартний спектральний аналіз, на основі якого визначається енергетичний спектр $E(f)$ в залежності від частоти f. Для стаціонарних часових рядів за наявності внутрішньорядних кореляційних зв'язків повинна існувати залежність вигляду

$$E(f) \sim f^{-\beta}, \tag{7.2}$$

де f - частота.

При цьому показник ступеня β для функції спектральної щільності пов'язаний із показником степеня відповідної автокореляційної функції $r(s)$

$$\gamma = 1 - \beta; \tag{7.3}$$

$$r(s) \sim s^{-\gamma}, \tag{7.4}$$

де s - крок в часі.

При цьому γ та β можуть відігравати роль масштабуючих множників.

Більш ніж півстоліття назад вчений Хурст показав, що в рядах річного стоку різних річок виявляється певна степенева залежність, яка вказує на існування властивостей самоподібності в коливаннях стоку.

У загальному випадку встановлення властивостей самоподібності передбачає наявність степеневої залежності між статистичним моментом $F_q(s)$ порядку q та масштабом s

$$F_q(s) = s^{h(q)}, \tag{7.5}$$

де $h(q)$ - масштабуючий ступінь Рені або фрактальна розмірність, яку по відношенню до часових рядів називають також узагальненим показником Хурста.

Якщо $h(q)$ не залежить від q (для усіх q значення $h(q)$ постійні), то це розглядається як властивість монофрактальності. Для монофрактальних об'єктів $h(q) = H$ зазвичай вивчається флуктуаційна функція другого порядку (так званий другий статистичний момент) вигляду

$$F_2(s) = s^H. \tag{7.6}$$

Пошук фрактальної розмірності відбувається шляхом побудови емпіричної залежності $F_2(s)$ від s, де H - степеневий показник кривої, який визначається як тангенс кута нахилу після логарифмування обох осей.

7.2 Визначення фрактальних розмірностей за просторовою структурною функцією

В сучасній стохастичній гідрології масштабування і пов'язане з ним встановлення фрактальних розмірностей більшості гідрометеорологічних величин можна одержати в результаті дослідження часової або

просторової варіації величини, яка вивчається, в часі або просторі з кроком s. Під варіацією розуміють зміну досліджуваної характеристики на деякому часовому інтервалі або заданій відстані, яку найчастіше представляють у вигляді статистичної функції

$$\sqrt{M(s)} \approx F_2(s) \sim s^H \qquad (7.7)$$

Функція $F_2(s)$ представляє собою квадратний корінь з відомої в статистичній гідрології структурної функції $M(s)$, яка застосовується при дослідженні гідрометеорологічних величин, чия стаціонарність носить локальний характер і зберігається на порівняно невеликих інтервалах зміни аргументу [23]. Структурну функцію визначають як математичне сподівання квадрата різниці перетинів випадкової фукції. Якщо реалізація ергодичного випадкового процесу задана в дискретних точках вимірювання, то структурна функція може бути представленою у вигляді

$$M(s) = \frac{1}{(n-s)} \sum_{i=1}^{n-s} (x_i - x_{i+s})^2 = 2\left[\sigma_x^2 - R(s)\right] \qquad (7.8)$$

де n - число вимірювань;

σ_x^2 -дисперсія;

$R(s)$-автоковаріаційна функція.

Структурні функції можуть бути як часовими, так і просторовими. Остання є математичним сподіванням квадрата різниці досліджуваної характеристики в двох точках, які знаходяться на відстані ΔL одна від одної. Просторова структурна функція використовувалася О.М. Колмогоровим для масштабування турбулентних утворень, при цьому як розрахункова характеристика розглядалася швидкість потоку.

Використання структурної функції покладено в основу методу варіацій Марка та Аронсона [8], розробленого для вивчення самоподібних об'єктів. Суть методу полягає в дослідженні масштабної поведінки просторової варіації досліджуваної величини. Просторову варіацію V можна представити структурною функцією

$$M(s) = \langle (Z_i - Z_j)^2 \rangle, \qquad (7.9)$$

де - Z_i, Z_j значення досліджуваної величини в точках і та j, а трикутні дужки означають усереднювання по всьому ансамблю точок.

Якщо варіація масштабована з показником самоподібності H, тобто можна записати $V \sim L^{2H}$, то остаточно виходять результати, які зазвичай представляються у вигляді залежності $V^{1/2}$ від відстані L, в подвійному логарифмічному масштабі.

Метод варіацій Марка та Аронсона був застосований до виявлення масштабної інваріантної в просторовому розподілі статистичних параметрів річного стоку. Розглянемо просторовий розподіл середніх багаторічних значень річного стоку $\bar{q}$ для правобережної України (79 гідрологічних постів). Просторова варіація оцінювалася як просторова структурна функція, представлена значеннями величини $\bar{q}$ в кожній точці простору

$$M(\Delta L) = \frac{\sum (A_i - A_j)^2}{m-1}, \qquad (7.10)$$

де $M(\Delta L)$- значення просторової структурної функції на відрізку ΔL, віднесене до його центру;

A_i, A_j - досліджувані характеристики в точках i та j (в цьому прикладі $A_i = \bar{q}_i$); m- число пар розглядуваних значень, які потрапили у відрізок ΔL.

В межах України була задана координатна сітка у вигляді квадратів із сторонами 75 км. Відстані між центрами тяжіння виділених водозборів визначалися за допомогою їхніх умовних координат, розраховуваних за теоремою Піфагора:

$$L_{i,j} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \qquad (7.11)$$

де L - відстань між об'єктами, яка оцінюється для кожної пари об'єктів i, j за умовними координатами просторового положення об'єкта x_i, y_i та x_j, y_j відповідно.

З метою обчислення просторової структурної функції були задані сегменти, що не перекриваються (градації) ΔL. На основі матриці відстаней вибиралися пари водозборів, які потрапляють в задану градацію ΔL. Для кожної з пар водозборів визначалася різниця $(A_i - A_j)$ й для кожної градації – відповідне значення структурної функції, розрахованої за (7.10). Результати розрахунків представляються у виді графіків (рис.7.1).

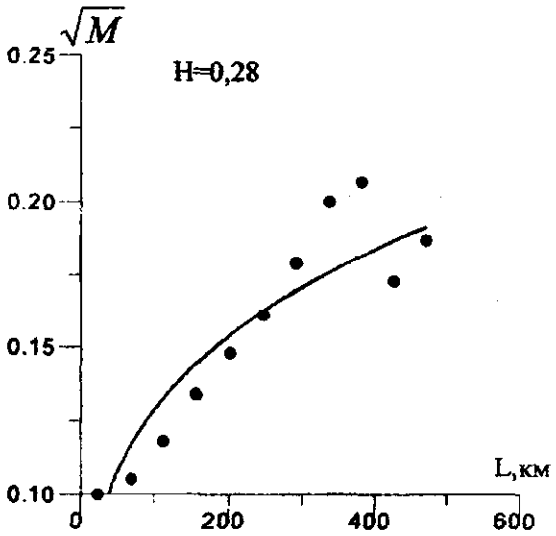

Рисунок 7.1 – Залежність значень флуктуаційної функції другого порядку від відстаней між центрами тяжіння водозборів

З рисунка видно, що при $L \le 400$км залежність варіації від відстані носить степеневий характер з показником ступеня, який визначався на основі подвійного логарифмування осей. При $L \le 400$км, дістаємо шукане значення фрактальної розмірності $H = 0{,}28$. При $L > 400$км структурна функція $M(s)$ й відповідна їй флуктуаційна функція $F_2(s) = \sqrt{M(s)}$ досягають стану насичення, що розглядається як ознака відсутності кореляційних зв'язків. Аналогічним чином були установлені властивості інваріантності у просторовому розподілі коефіцієнта варіації Cv річного стоку, для якого фрактальна розмірність складає $H = 0{,}37$. На відміну від середньоарифметичного значення просторова скорельованість значень Cv простежується тільки на відстані $L \le 200$км.

$$z_n = \sum_{i=1}^{i=n} \varphi_i \qquad n = 1,2,\ldots,N. \tag{7.12}$$

Другий статистичний момент (флуктуаційна функція) визначається за формулою

$$\sqrt{M(s)} = F_2(s) = \left\{ \frac{1}{N_s} \sum_{v=1}^{Ns} \left[z_{vs} - z_{v-1,s} \right]^2 \right\}^{\frac{1}{2}}, \tag{7.13}$$

де v- порядковий номер виділених відрізків, що не перекриваються, розміром s; $N_S = \text{int}(\frac{N}{s})$- число значень досліджуваної величини, які потрапляють в кожний відрізок розміром s;
різниця $z_{vs} - z_{v-1,s}$ є приростом z_n на кожному відрізку завдовжки s.

Флуктуаційна функція, що використовується у фрактальному аналізі часових рядів, по суті виконує розбиття множини N на підмножини N_s. На основі цього розкладання робляться висновки про існування ознак масштабної самоподібності. З метою встановлення таких ознак розглядаються залежності узагальненої функції $F_2(s)$ від s, яка має вигляд степеневої функції, схожої на наведену на рис. 7.1.

На великих часових масштабах фрактальна розмірність менша за 1. Якщо $H = 1$, це є свідченням самоподібності. Якщо $H < 1$, то мова йде про "самоафінність". Фрактальна розмірність рядів річного стоку змінюється від 0,5 до 0,95. Величина H пов'язана з характеристиками автокореляційних та спектральних функцій наступним рівнянням

$$H = 1 - \gamma/2 = (1 + \beta)/2. \tag{7.14}$$

Таким чином, можна зробити висновок, що районування за особливостями коливань стоку рядів може відбуватися на основі установлених фрактальних розмірностей, подібно до того, як відбувається районування за коефіцієнтами автокореляції.

$$z_n = \sum_{i=1}^{i=n} \varphi_i \qquad n = 1,2,...,N. \tag{7.12}$$

Другий статистичний момент (флуктуаційна функція) визначається за формулою

$$\sqrt{M(s)} \approx F_2(s) = \left\{ \frac{1}{N_S} \sum_{v=1}^{Ns} \left[z_{vs} - z_{v-1,s} \right]^2 \right\}^{\frac{1}{2}}, \tag{7.13}$$

де v- порядковий номер виділених відрізків, що не перекриваються, розміром s; $N_S = \mathrm{int}(\frac{N}{s})$- число значень досліджуваної величини, які потрапляють в кожний відрізок розміром s;
різниця $z_{vs} - z_{v-1,s}$ є приростом z_n на кожному відрізку завдовжки s.

Флуктуаційна функція, що використовується у фрактальному аналізі часових рядів, по суті виконує розбиття множини N на підмножини N_s. На основі цього розкладання робляться висновки про існування ознак масштабної самоподібності. З метою встановлення таких ознак розглядаються залежності узагальненої функції $F_2(s)$ від s, яка має вигляд степеневої функції, схожої на наведену на рис. 7.1.

На великих часових масштабах фрактальна розмірність менша за 1. Якщо $H = 1$, це є свідченням самоподібності. Якщо $H < 1$, то мова йде про "самоафінність". Фрактальна розмірність рядів річного стоку змінюється від 0,5 до 0,95. Величина H пов'язана з характеристиками автокореляційних та спектральних функцій наступним рівнянням

$$H = 1 - \gamma/2 = (1 + \beta)/2. \tag{7.14}$$

Таким чином, можна зробити висновок, що районування за особливостями коливань стоку рядів може відбуватися на основі установлених фрактальних розмірностей, подібно до того, як відбувається районування за коефіцієнтами автокореляції.

ЛІТЕРАТУРА

1. Багров Н.А. Аналитическое представление последовательности метеорологических полей посредством естественных ортогональных составляющих // Труды ЦИП.- 1959. -Вып.74. - С.133-138.
2. Бершадский А.Г. Крупномасштабные фрактальные структуры в лабораторной турбулентности, океане и астрофизике // Успехи физических наук. – 1990. – Т.160. – Вып.12. – С.189-194.
3. Бефани А.Н., Мельничук О.П. Расчет нормы стока временных водотоков и горных Украинских Карпат // Труды УкрНИГМИ. - Л.: Гидрометеоиздат. - 1967. - вып. 69. - С. 105 -131.
4. Вентцель Е.С. Теория вероятностей. – М. ФМ, 1962, 264 с.
5. Жук В.А., Евстигнеев В.М. Исследование синхронности колебаний годового стока отдельных регионов приемами факторного анализа: Труды ВНИИГМИ-МЦД. - М.: Гидрометиздат. С. 78-91с.
6. Евстигнеев В.М. Речной сток и гидрологические расчеты. – М.: Изд-во МГУ, 1990. – 304 с.
7. Иберла К. Факторный анализ: Пер. с англ. - М.: Статистика,1980. - 397с.
8. Иванов С.С. Определение фрактальной размерности глобального рельефа // Океанология. – 1994. – Т.34,№1. – С.102-106.
9. *Исследования и расчеты речного стока* / Под ред. В.Д.Быкова. - М.: Изд-во МГУ, 1981. -.228 с.
10. Карасев И.Ф., Савельева Л.Н. Разложение гидрологических полей на естественные ортогональные составляющие и расчет слоев стока весеннего половодья неизученных рек // Моделирование и прогнозы гидрологических процессов. – Л.: РГГМИ, 1992. – Вып.113. – С.76-84.
11. Крицкий С.Н., Менкель М.Ф. Гидрологические основы управления речным стоком.- М. Наука,1981. - 235с.
12. Крицкий С.Н., Менкель М.Ф. Гидрологические основы управления водохозяйственными системами.- М. Наука,1982. – 271 с.
13. Лобода Н.С., Горобец Т.В. Фрактальные свойства в многолетних колебаниях годового стока рек Украины // Вісник Одеського державного екологічного університету. – Вип.1. - К:КНТ. –2005. - С. 174 -182.
14. Лобода Н.С. Расчеты и обобщения характеристик годового стока рек Украины в условиях антропогенного влияния: Монография. – Одесса: Экология, 2005. – 208 с.

15. Лобода Н.С. Формализм функций памяти и мультифрактальный подход в задачах моделирования годового стока рек и его изменения под влиянием факторов антропогенной деятельности // Міжвід. наук. зб. України. - Метеорологія, кліматологія та гідрологія. - Одеса. - 2002. – Вип. 45. - С. 140 -146.
16. Лобода Н.С., Гопченко Є.Д. Стохастичні моделі у гідрологічних розрахунках. - Навчальний посібник.– Одеса: Екологія, 2006. – 200 с.
17. Лобода Н.С., Нгуен Ву Ань Статистическая структура полей годового стока в бассейне р.Уссури и стокоформирующие факторы // Український гідрометеорологічний журнал. - №1. – 2006. – С.170-176.
18. Лобода Н.С. Ландшафтна різноманітність та районування характеристик стоку Українських Карпат // Науковий Вісник Чернівецького університету. – Вип.305. , Географія. – 2006. - С. 12-19.
19. Мещерская А.В., Руховец Л.В., Юдин М.И., Яковлева Н.И. Естественные составляющие метеорологических полей. - Л.: Гидрометеоиздат, 1970.- 200 с.
20. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. - М.: Финансы и статистика. - 1982. - 120 с.
21. Пространственно-временные колебания стока рек СССР / Под ред.А.В. Рождественского. – Л.: Гидрометеоиздат,1988. – 376 с.
22. Раткович Д.Я. Многолетние колебания речного стока. - Л.: Гидрометеоиздат. - 1976. - 255 с.
23. Рождественский А.В.,Чеботарев А.И. Статистические методы в гидрологии. - Л.: Гидрометеоиздат, 1974. - 424с.
24. Руководство по определению расчетных гидрологических характеристик. – Л. Л.: Гидрометеоиздат, 1973. - 111 с.
25. Смирнов Н.П., Скляренко В.Л. Методы многомерного статистического анализа в гидрологических исследованиях. - Л.: Ленингр. ун-т, 1986. - 192 с.
26. Факторный, дискриминантный и кластерный анализ. Пер. с англ. /Дж.-О.Ким,Ч.У.Мьюллер,У.Р. Клекка и др.- Финансы и статистика, 1989, - 215с.
27. Христофоров А.В. Надежность расчетов речного стока. – М.: Изд.-во МГУ. – 1993. – 168 с.
28. Школьний Є.П., Лоєва І.Д., Гончарова Л.Д. Обробка та аналіз гідрометеорологічної інформації: навчальний підручник. - К.: Міносвіти України, 1999. - 600 с.

Додаток А

Значення критерію Фішера F для рівня значущості 0.05

ν_1 - число степенів вільності для більшої дисперсії

ν_2 - число степенів вільності 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 степенів вільності	Рівень значущості 0.05	Число степенів вільності	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		

Навчальне видання

ЛОБОДА Наталія Степанівна

МЕТОДИ ПРОСТОРОВОГО УЗАГАЛЬНЕННЯ ГІДРОЛОГІЧНОЇ ІНФОРМАЦІЇ

Конспект лекцій

Директор видавництва *Д. М. Островеров*
Технічний редактор *М. М. Бушин*

Підписано до друку 29.08.2008. Формат 60x84/16. Папір офсетний.
Друк офсетний. Ум. друк. арк. 5,12.
Тираж 200 прим. Зам. № 188.

Надруковано з готового оригінал-макета

Видавництво і друкарня "Екологія"
(Свідоцтво ДК № 1873 від 20.07.2004 р.)
65045, м. Одеса, вул. Базарна, 106.
Тел.: 33-07-18, 37-07-95, 37-14-25.

Лобода Н. С.

Л 68 Методи просторового узагальнення гідрологічної інформації: Конспект лекцій. — Одеса: Екологія, 2008 — 88 с.

Конспект лекцій призначений для магістрів, які навчаються за спеціальністю "Гідрологія та гідрохімія", напрям підготовки "Гідрометеорологія". В конспекті розглянуто методи багатовимірного статистичного аналізу стосовно до їх застосування при географічних узагальненнях стоку річок.

Конспект лекцій може бути використаний при виконанні дипломних, магістерських та аспірантських робіт.

ББК 26.22
УДК 556.16