МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ ДЕРЖАВНИЙ ЕКОЛОГІЧНИЙ УНІВЕРСИТЕТ

ЛОБОДА Н. С., КАТИНСЬКА І.В.

МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ПРИ ВИРІШЕННІ ГІДРОЕКОЛОГІЧНИХ ЗАДАЧ

Конспект лекцій

Одеса
Одеський державний екологічний університет
2022

УДК 504.4.054:519.2
Л68

Лобода Н.С., Катинська І.В.

Л68 Методи багатовимірного аналізу при вирішенні гідроекологічних задач : конспект лекцій. Одеса, Одеський державний екологічний університет, 2022. 134 с.
ISBN 978-966-186-208-0

В конспекті лекцій наведені найбільш поширені методи багатовимірного статистичного аналізу гідроекологічних даних. Розглянуті теоретичні основи статистичних методів, які використовуються для обґрунтування способів просторово-часового узагальнення інформації, її стиснення, фільтрації та відновлення, а також вирішення задач класифікації та альтернативних прогнозів (факторний, дискримінантний аналіз, метод головних компонентів). До змісту конспекту лекцій входять також розділи кореляційного аналізу, включаючи автокореляційні та взаємокореляційні функції, регресійний аналіз та дисперсійний аналіз, основні поняття фрактальності. Поряд із викладенням теоретичних положень методів розрахунків запропоновані приклади їх застосування у гідроекології.

Конспект лекцій призначений для професійної та практичної підготовки студентів рівня вищої освіти магістр за спеціальністю 101 «Екологія», також може використовуватись аспірантами та студентами інших освітніх програм та рівнів вищої освіти за спеціальністю Екологія денної та заочної форм навчання.

УДК 504.4.054:519.2

Рекомендовано методичною радою Одеського державного екологічного університету Міністерства освіти і науки України як конспект лекцій (протокол №10 від 30.06. 2022 р.)

ISBN 978-966-186-208-0

ЗМІСТ

ВСТУП

Дисципліна «Методи багатовимірного аналізу при вирішенні гідроекологічних задач» відноситься до дисциплін професійно-орієнтованого циклу за спеціальністю 101 «Екологія», освітня програми «Гідроекологія», рівень ВО «магістр».

Дисципліна передбачає вивчення теоретичних основ сучасних методів багатовимірного статистичного аналізу даних та їх практичне застосування до вирішення практичних задач гідроекології.

Загальний обсяг навчального часу визначається робочим навчальним планом та становить 30 годин лекцій, 30 годин практичних занять, 90 годин самостійної роботи студентів.

В результаті вивчення дисципліни «Методи багатовимірного аналізу при вирішенні гідроекологічних задач» студенти повинні знати теоретичні основи методів багатовимірного статистичного аналізу та вміти виконувати інтерпретацію отриманих результатів розрахунків для вирішення прикладних гідроекологічних задач.

Підчас засвоєння дисципліни передбачається опанування основ кореляційного, регресійного, факторного, дискримінантного аналізу, методу головних компонент з метою виявлення основних закономірностей гідрологічних та гідроекологічних процесів у часі та просторі.

Вивчення дисципліни «Методи багатовимірного аналізу при вирішенні гідроекологічних задач» ґрунтується на знаннях, одержаних студентами за такими дисциплінами учбового плану «Методи математичної статистики у гідроекологічних дослідженнях», «Вища математика», «Моделювання та прогнозування стану довкілля», «Гідроекологія», «Гідрохімія». Вивчення дисципліни потребує розуміння процесів виникнення та формування гідрологічних та гідроекологічних явищ, гідрохімічного складу вод й впливу на них антропогенних чинників. Знання, одержані в результаті вивчення дисципліни «Методи багатовимірного аналізу при вирішенні гідроекологічних задач», будуть використовуватися у курсовому проектуванні, при написанні кваліфікаційних магістерських та аспірантських робіт.

ЛЕКЦІЯ 1
«ВИПАДКОВІ ВЕЛИЧИНИ ТА ЇХ ХАРАКТЕРИСТИКИ»

1.1 Випадкова величина та закони її розподілу

Якщо в результаті випробувань розглядається не сама подія, а її кількісна характеристика, то в теорії ймовірностей використовується поняття випадкової величини. ***Випадковою називається величина, яка внаслідок випробування набуває того чи іншого значення, наперед невідомо, якого саме.*** Якщо можливі значення випадкової величини можна перерахувати, то випадкова величина відноситься до ***дискретної.*** Дискретна випадкова величина описується кінцевим числом значень. Дискретна випадкова величина X, із можливими значеннями $x_1, x_2, x_3, \ldots, x_N$, з ймовірнісної точки зору буде повністю описана, якщо кожному значенню випадкової величини поставити у відповідність значення ймовірності його появи у статистичному випробуванні. Наприклад, дискретна випадкова величина X має можливі значення $x_1, x_2, x_3, \ldots, x_N$ з відповідними ймовірностями – $p_1, p_2, p_3, \ldots, p_N$.

Законом розподілу випадкової дискретної величини називається співвідношення, яке установлює зв'язок між можливими значеннями випадкових величин та їх ймовірностями. Закон розподілу дискретної випадкової величини X представляється *таблицею розподілу*, яка має наступний вигляд

$$x_i: \quad x_1 \quad x_2 \quad x_3 \quad \ldots \quad x_N;$$

$$p_i: \quad p_1 \quad p_2 \quad p_3 \quad \ldots \quad p_N$$

Таблицю розподілу називають також *рядом розподілу* випадкової величини *X*. Про випадкову величину *X* говорять, що вона підкорюється заданому закону розподілу.

Сума ймовірностей усіх значень дискретної випадкової величини дорівнює одиниці, тобто $\sum_{i=1}^{N} p_i = 1$**.**

Якщо таблицю розподілу представити у графічному вигляді (рис.1.1), то отримаємо багатокутник розподілу. Багатокутник розподілу є графічним зображенням ряду розподілу: на вісь абсцис наносяться значення випадкової величини, на вісь ординат – ймовірності цих значень. Багатокутник розподілу також як і ряд розподілу повністю характеризує випадкову величину. Закон розподілу, представлений у вигляді таблиці або багатокутника, є вичерпною ймовірнісною характеристикою дискретної випадкової величини.

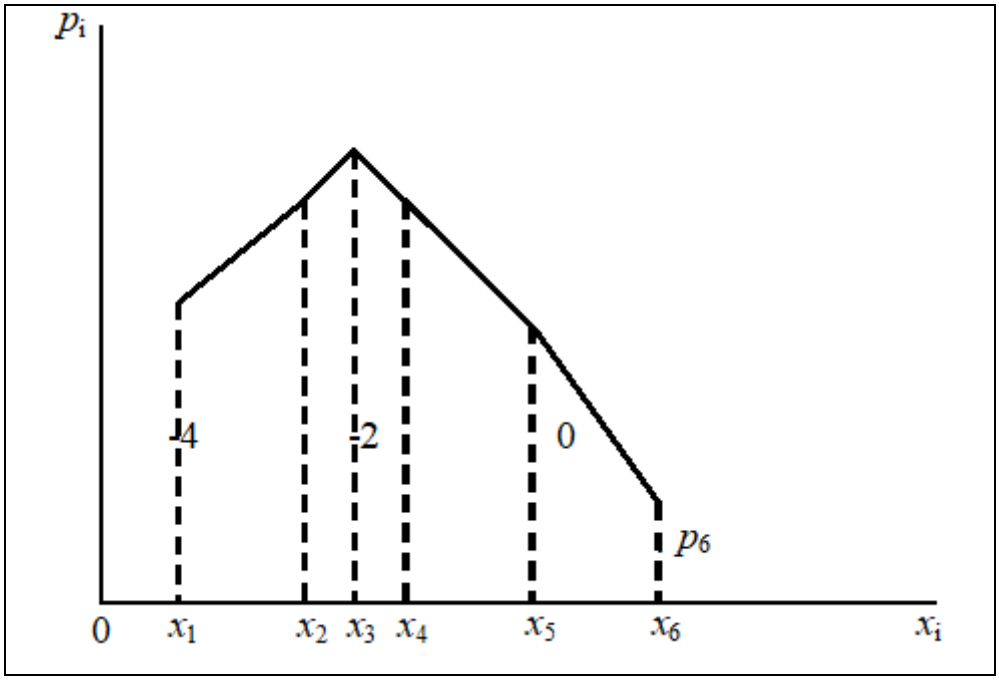

Рисунок 1.1 – Багатокутник розподілу дискретної випадкової величини

У гідроекології та пов’язаної з нею гідрології використовуються випадкові величини, які отримуються в результаті випробувань

(гідрологічних спостережень), тобто розглядаються не генеральні сукупності, а вибірки з них ($n << N$).

Якщо йдеться мова йде про випадкову величину, то ***частотою появи значення випадкової (емпіричною ймовірністю) дискретної величини X називається відношення числа випадків*** m***, коли розглядуване значення величини спостерігалося, до загальної кількості значень, які випадкова величина X приймала в результаті*** n ***випробувань***:

$$p_i^* = \frac{m}{n}, \tag{1.1}$$

де p_i^* – частота появи випадкової величини;

i – номер випробування;

m – число випадків, коли значення випадкової величини спостерігалося у результаті проведення n випробувань;

n – загальна кількість випробувань.

Поряд з дискретними випадковими величинами існують і безперервні. ***Випадкові величини, які заповнюють деякий проміжок на осі*** x***, називають безперервними.*** Кожне окреме значення безперервної випадкової величини не володіє ніякою, відмінною від нуля, ймовірністю. Для того, щоб задати закон розподілу безперервної випадкової величини, вводиться поняття інтегральної функції розподілу

$$F(x) = p(X < x), \tag{1.2}$$

де $F(x)$ не ймовірність події $X = x$, а ймовірність того, що $X < x$, де x – деяка поточна змінна.

Функція розподілу випадкової величини є самою універсальною характеристикою випадкової величини. Вона установлюється як для безперервних, так і перервних величин. Вона повністю характеризує випадкову величину з ймовірнісної точки зору і є однією з форм законів розподілу.

Функцію розподілу $F(x)$ ***також називають інтегральною функцією розподілу або інтегральним законом розподілу.***

Функція розподілу $F(x)$ має такі властивості:

1. Вона монотонно неспадна, тобто при $x_2 > x_1$, $F(x_2) > F(x_1)$ (рис.1.2).

2. На мінус нескінченності інтегральна функція розподілу дорівнює нулю:

$$\text{коли } x \to -\infty, \text{ то } F(-\infty) = 0. \tag{1.3}$$

3. На плюс нескінченності інтегральна функція розподілу дорівнює одиниці:

$$\text{коли } x \to +\infty, \text{ то } F(+\infty) = 1. \tag{1.4}$$

Інтегральна функція розподілу дискретної випадкової величини установлюється на основі таблиці розподілу

$$F(x) = p(x < X) = \sum_{x_i < x} p_i(X = x_i), \tag{1.5}$$

де нерівність $x_i < x$ указує на те, що підсумовування відбувалося не для усіх x_i, а лише для тих, які меньше x.

У гідрологічних розрахунках (СНіП 2.01.14-83) здебільшого використовується не інтегральна функція розподілу $F(x)$, а функція забезпеченості $P(x)$, яка пов'язана з інтегральною функцією співвідношенням

$$P(x) = 1 - F(x). \tag{1.6}$$

Забезпеченість випадкової величини X – це ймовірність того, що випадкова величина X більше деякого заданого значення x, тобто $P(x) = p(X > x)$.

Для дискретних випадкових величин функція забезпеченості може бути представленою у вигляді

$$P(x) = p(X > x) = \sum_{x_i > x} p_i(X = x_i), \tag{1.7}$$

де нерівність $x_i > x$ указує на те, що підсумовування відбувалося не для усіх x_i, а лише для тих, які більші x.

Інтегральна функція $F(x)$, яка представлена у графічному вигляді, має назву інтегральної кривої розподілу, а функція з $P(x)$ називається кривою забезпеченості (рис.1.2).

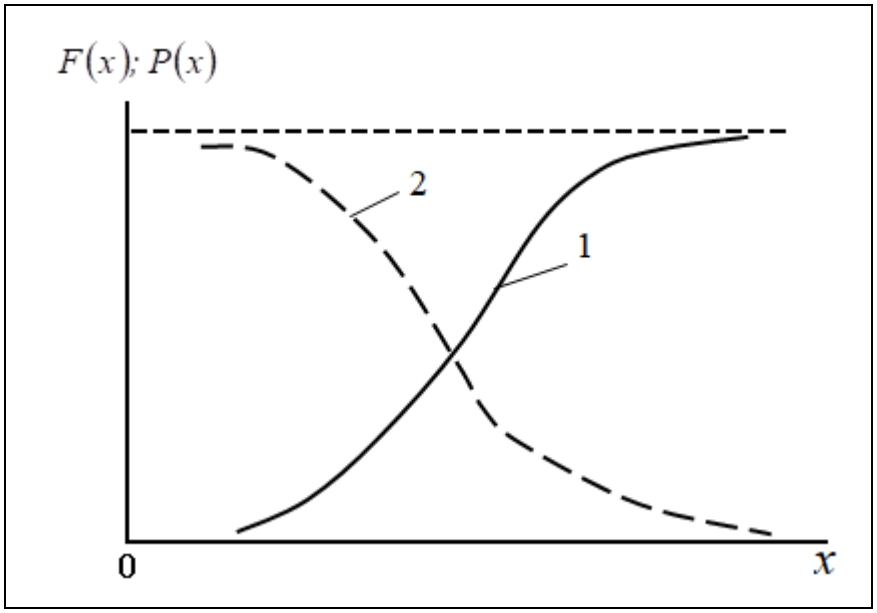

Рисунок 1.2 – Інтегральна функція розподілу $F(x)$ випадкової величини *x* (1) та функція забезпеченості $P(x)$ (2)

Похідна $F'(x) = f(x)$ ***характеризує щільність ймовірності безперервної випадкової величини*** X ***у точці*** x ***і носить назву щільності розподілу безперервної випадкової величини або диференціального закону розподілу випадкової безперервної величини.*** Ця форма закону розподілу існує тільки для безперервних випадкових величин.

Диференціальна функція розподілу $f(x)$ безперервної випадкової величини *x* має такі властивості:

1. Щільність розподілу завжди невід'ємна, тобто

$$f(x) \geq 0. \tag{1.8}$$

2. Інтеграл від щільності розподілу у межах від $-\infty$ до ∞ дорівнює одиниці

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \qquad (1.9)$$

3. Щільність розподілу є розмірною величиною і її розмірність зворотна розмірності величини x.

З геометричної точки зору перераховані властивості указують на те, що крива щільності розподілу лежить не нижче осі абсцис та повна площа, обмежена кривою $f(x)$ і віссю абсцис, дорівнює 1.

Між інтегральною, диференціальною функціями та функцією забезпеченості існують певні співвідношення, які записуються таким чином

$$p(\alpha < X < \beta) = F(\beta) - F(\alpha), \qquad (1.10)$$

$$p(\alpha < X < \beta) = \int_{\alpha}^{\beta} f(x)dx, \qquad (1.11)$$

$$P(x) = 1 - \int_{-\infty}^{x} f(x)dx = 1 - F(x) = \int_{x}^{\infty} f(x)dx. \qquad (1.12)$$

З (1.10-1.12) витікає, що через одну форму закону розподілу можна перейти до іншої.

1.2 Оцінки статистичних параметрів закону розподілу випадкової величини та вимоги до них

Якщо закон розподілу базується на даних спостережень, то він є емпіричним. Якщо закон розподілу представляється у виді математичного рівняння, то він є теоретичним (нормальний, Пірсона III, логнормальний та інші (*Лобода Н.С., Гопченко Є.Д., 2006*). ***Статистичними параметрами законів розподілу є числові характеристики, які несуть у собі певну інформацію про особливості***

розподілу випадкової величини на числовій осі. Оскільки ми маємо у розпорядженні тільки обмежені за часом результати спостережень (випадкову вибірку з усієї генеральної сукупності), то у подальшому мова може йти тільки про більш чи менш наближені до істини кількісні характеристики розподілу. Тому статистичні параметри, які розраховані по вибіркам (даним спостережень), мають назву оцінок статистичних параметрів. До оцінок статистичних параметрів висуваються такі вимоги.

1. Оцінки повинні бути незміщеними, тобто математичне сподівання кожної оцінки має дорівнювати значенню статистичного параметра генеральної сукупності.

2. Оцінки мають бути ефективними: їх розсіювання відносно значення відповідного параметру генеральної сукупності має бути найменшим з усіх можливих.

3. Оцінки повинні бути обґрунтованими, тобто сходитися по імовірності до параметру генеральної сукупності.

1.3 Розрахунки статистичних параметрів за методом моментів

В основі цього методу лежить визначення статистичних параметрів законів розподілу випадкової величини через статистичні моменти. Назва «метод моментів» прийшло із розділу фізики “Механіка”, де механічний момент являє собою добуток сили на плече. Плече – відстань від точки, у якій прикладена сила, до точки опори. Значення дискретної випадкової величини розглядається як матеріальна точка на числовій осі з масою, пропорційною ймовірності появи цієї випадкової величини. Якщо плечем є відстань від нуля числової осі до матеріальної точки, то такі статистичні моменти називаються *початковими*. Коли ж для визначення статистичного моменту береться відстань від математичного сподівання до розглядуваної матеріальної точки, то такий статистичний момент отримує назву *центрального***.**

Для будь-якої випадкової величини центральний момент першого порядку дорівнює нулю.

У математичній статистиці широко використовують початкові α та центральні μ статистичні моменти

$$\alpha_s = \sum_{i=1}^{N} x_i^s \, p_i \,, \tag{1.13}$$

$$\mu_s = \sum_{i=1}^{N} (x_i - m_x)^s \, p_i \,; \tag{1.14}$$

де x_i – значення, яке приймає дискретна випадкова величина;

p_i – ймовірність появи значення x_i;

N – кількість значень x_i, які утворюють генеральну сукупність дискретної випадкової величини X

S – порядок моменту.

Найбільше застосування у практиці знайшли статистичні моменти $\alpha_1, \mu_2, \mu_3, \mu_4$ та їх безрозмірні характеристики (нормовані моменти). Вони лягли в основу визначення статистичних параметрів.

Статистичний параметр математичне сподівання m_x може бути представленим як середнє зважене по ймовірності значення випадкової величини *X*. Кожне із значень x_i під час осереднення враховується з вагою, пропорційною імовірності появи цього значення

$$m_x = \frac{x_1 p_1 + x_2 p_2 + \ldots + x_N p_N}{p_1 + p_2 + \ldots + p_N} = \frac{\sum_{i=1}^{N} x_i p_i}{\sum_{i=1}^{N} p_i}. \tag{1.15}$$

де N – довжина генеральної сукупності.

Враховуючи, що для дискретних випадкових величин $\sum_{i=1}^{N} p_i = 1$, отримаємо

$$m_x = \sum_{i=1}^{N} x_i p_i \,. \tag{1.16}$$

Вираз (1.16), який описує математичне сподівання, є першим початковим моментом α_1. Таким чином, математичне сподівання сума добутків усіх можливих значень випадкової величини x_i помножених на ймовірність появи p_i цих значень може бути представленим як абсциса

центру тяжіння усієї системи N матеріальних точок. Отже, ***математичне сподівання*** m_x є центром статистичного розподілу випадкової величини X на числовій осі.

Середнє арифметичне є оцінкою математичного сподівання за даними спостережень (вибіркою довжиною n)

$$\hat{m}_x = \bar{x} = \sum_{i=1}^{n} x_i p_i = \frac{1}{n}\sum_{i=1}^{n} x_i , \tag{1.17}$$

де $p_i = \frac{1}{n}$.

Другий центральний момент μ_2 називають дисперсією і позначають як σ_x^2

$$\mu_2 = \sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{1.18}$$

З дисперсію пов'язані такі статистичні параметри як середнє квадратичне відхилення, коефіцієнт варіації. Другий центральний момент μ_2 або дисперсія σ_x^2 характеризує розсіювання значень випадкової величини відносно математичного сподівання. Дисперсія випадкової величини має розмірність, яка дорівнює квадрату випадкової величини. Але для більш наочної характеристики розсіювання зручно користуватися величиною, розмірність якої співпадає з розмірністю випадкової величини. Для цього з дисперсії добувають квадратний корінь. Отримана величина зветься середнім квадратичним відхиленням (стандартом) і позначується символом σ_x.

Середнє квадратичне відхилення (стандарт) σ_x розраховується за такою формулою

$$\sigma_x = \sqrt{\mu_2} , \tag{1.19}$$

або

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (1.20)$$

Стандарт, представлений у безрозмірному вигляді називається коефіцієнтом варіації C_V.

Визначення коефіцієнта варіації виконується за формулами

$$C_V = \frac{\sqrt{\mu_2}}{m_x} \quad \text{та} \quad C_V = \frac{\sigma_x}{m_x}, \qquad (1.21)$$

або

$$C_V = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}^2(n-1)}} = \sqrt{\frac{\sum_{i=1}^{n}(k_i - 1)^2}{n-1}} \qquad (1.22)$$

де $k_i = x_i / \bar{x}$ - модульний коефіцієнт.

Третій центральний момент μ_3 служить характеристикою асиметрії розподілу. Якщо розподіл випадкової величини симетричний відносно m_x, то μ_3 дорівнює нулю. Безрозмірна характеристика асиметрії називається коефіцієнтом асиметрії, який розраховується за формулами

$$C_S = \frac{\mu_3}{\sigma_x^3}, \qquad (1.23)$$

$$C_S = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(k_i - 1)^3}{C_V^3}. \qquad (1.24)$$

Для випадкової величини, яка підкорюється нормальному закону розподілу, $C_S=0$, отже, крива щільності розподілу ймовірностей випадкової величина є симетричною відносно математичного сподівання.

Четвертий центральний момент характеризує гостровершинність розподілу у порівнянні з кривою нормального розподілу (рис.1.3). Для випадкової величини з нормальним законом розподілу співвідношення μ_4 / σ_x^4 завжди дорівнює 3. Таким чином, нормування β_4 через σ_x^4 дозволяє отримати безрозмірний статистичний параметр, названий ексцесом

$$E = \frac{\beta_4}{\sigma_x^4} - 3. \tag{1.25}$$

$$E = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{n\sigma_x^4} - 3. \tag{1.26}$$

Якщо $E > 0$, то крива розподілу витягнута догори відносно нормального закону розподілу, для якого $E = 0$. Коли ж $E < 0$, крива розподілу приплюснута донизу по відношенню до кривої нормального розподілу.

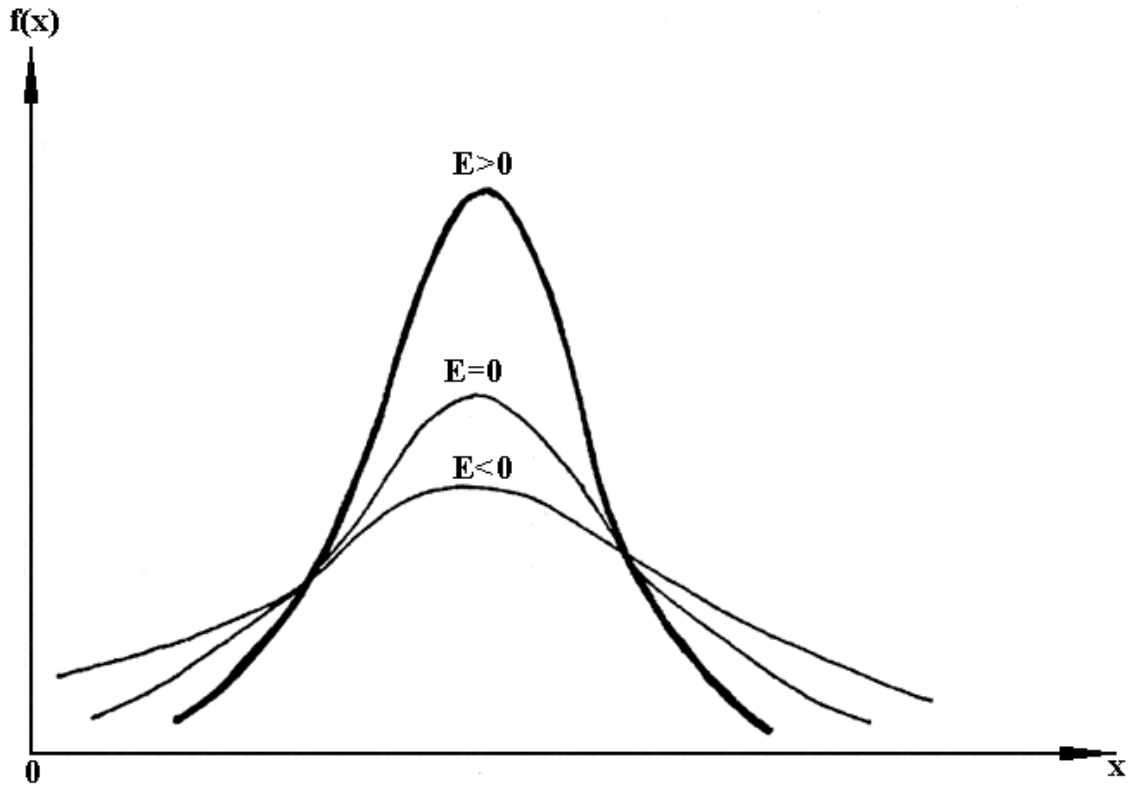

Рисунок 1.3 – Криві розподілу з різними значеннями ексцесу E

1.4 Точність оцінок статистичних параметрів стоку, розрахованих за методом моментів

Як відзначено вище, оцінка статистичного параметру за вибіркою може відрізнятися від значення цього ж параметру генеральної сукупності. Якість оцінок визначається їх систематичними та випадковими похибками. Систематичні похибки можна усунути, а випадкові лише оцінити. Вичерпне уявлення про випадкові похибки оцінок параметрів дає знання їх закону розподілу, але іноді встановити закон розподілу вибіркових оцінок неможливо. Частіше встановлюють середньоквадратичну похибку оцінок, яка є основним показником випадкових похибок. Оцінювання середньоквадратичного відхилення вибіркових оцінок параметрів від відповідних параметрів генеральної сукупності проводилося на основі метода статистичних випробувань (Монте Карло) за таким алгоритмом (табл.1.1).

1. Генерація генеральної сукупності досліджуваних величин за обраною функцією розподілу і заданими статистичними параметрами, яка розбивається на вибірки менші за об'ємом.
2. Розрахунки зміщених оцінок параметрів по вибіркам.
3. Усунення зміщення оцінок відносно параметру генеральної сукупності.
4. Розрахунки середньоквадратичного відхилення незміщених оцінок від значення параметрів генеральної сукупності.

Саме таким шляхом були розроблені формули середньоквадратичного відхилення параметрів C_S та C_V. Що стосується оцінки математичного сподівання, то її середньоквадратичне відхилення розраховується за формулою розробленою для величин, які підкоряються нормальному закону розподілу, тобто

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \ , \qquad (1.27)$$

дотримуючись при цьому припущення, що нормальний закон розподілу вибіркових середніх зберігається і для вибірок, які відхиляються від нормального розподілу.

Таблиця 1.1 – Формули для оцінки випадкових похибок розрахунків статистичних параметрів

Параметр	Середньоквадратичне відхилення параметру
$\overline{X}$	$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$
C_V	$\sigma_{C_V} = \frac{C_v}{n + 4C_v^2}\sqrt{\frac{(n(1+C_v^2)}{2}}$
C_S	$\sigma_{C_S}\sqrt{\frac{6}{n}(1 + 6C_v^2 + 5C_v^4)}$
C_S / C_V	$\sigma_{C_S/C_V} = \frac{1}{C_v}\sqrt{\frac{6}{n}}$

Відношення стандарту вибіркової оцінки до значення самого вибіркового параметру виражене у відсотках

$$\varepsilon_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}100\% = \frac{C_v}{\sqrt{n}}100\%, \quad (1.28)$$

$$\varepsilon_{C_V} = \frac{\sigma_{C_V}}{C_v}100\%, \quad (1.29)$$

$$\varepsilon_{C_S} = \frac{\sigma_{C_S}}{C_S}100\% \quad (1.30)$$

використовується як критерій визначення достатньої чи недостатньої тривалості спостережень за досліджуваною величиною. Наприклад, при розрахунках річного стоку тривалість періоду спостережень за стоком визнається достатньою, якщо ε_x<5-10%, ε_{Cv}<15%. У протилежному випадку ці оцінки статистичних параметрів уточняються по даним річок-аналогів з набагато більшим періодом спостережень (приведення рядів стоку до

тривалого періоду за методом аналогії). Вибіркові значення коефіцієнта асиметрії, як правило, мають великі середньоквадратичні відхилення і його відносне значення (ε_{Cs}) досягає 50-70% від вибіркового значення самого параметру. У зв'язку з цим оцінки параметра C_S, або співвідношення C_S/C_V осереднюються у границях однорідних за умовами формування досліджуваного процесу районах.

З формул розрахунків середньоквадратичних відхилень оцінок статистичних параметрів видно, що випадкові похибки зростають по мірі збільшення часової мінливості стоку, яка характеризується коефіцієнтом варіації C_v.

1.5 Емпірична крива забезпеченостей

У гідрологічних та гідроекологічних розрахунках закон розподілу випадкової величини задається у вигляді функції забезпеченості (див. рис. 1.2). Забезпеченістю значення x випадкової величини X є імовірність перевищення цього значення, тобто

$$P(x) = p(X > x), \tag{1.31}$$

де $P(x)$ – забезпеченість значення x;

$p(X > x)$ – ймовірність перевищення значення x.

Оцінкою забезпеченості значення x_i з вибірки дискретних випадкових величин довжиною n є відносна частота події $X \geq x_i$. Для обчислення емпіричної забезпеченості $P(x)$ використовується формула

$$P(x_i) = \frac{m}{n+1}, \tag{1.32}$$

де m – порядковий номер елемента x_i у ранжованому статистичному ряді (мається на увазі статистичний ряд, в якому всі члени розміщені в порядку зменшення $x_{i+1} < x_i$). Фактично m являє собою абсолютну частоту появи події $X \leq x_i$.

На рисунку 1.4 показана емпірична крива забезпеченості критерію чутливості до забруднення сполуками азоту. Водна Рамкова Директива та Нітратна Директива визначили граничне значення вмісту сполук азоту у

воді (50 мг/дм^3 або 11,3 мгN/дм^3). З метою кількісної оцінки чутливості річок до забруднення сполуками азоту була використана сума концентрацій іонів азоту у воді (мгN/дм^3). Сумарне значення концентрацій розглядалося як критерій чутливості:

$$k_н = NH_4^+ + NO_2^- + NO_3^-, \quad (1.33)$$

де $k_н$ – показник чутливості території до забруднення сполуками азоту;

NH_4^+ – концентрація азоту амонійного, мгN/дм^3;

NO_2^- – концентрація азоту нітритного, мгN/дм^3;

NO_3^- – концентрація азоту нітратного, мгN/дм^3.

Якщо $k_н$>11,3 мгN/дм^3, то розглядувана зона визначається як “чутлива до забруднення” сполуками азоту, тобто вона відноситься до зони ризику недосягнення водним об’єктом “доброго екологічного стану”. У прикладі ймовірність перевищення критичного значення $k_н$ становить 24%. Зона ризику попадає у область забезпеченостей від 24% до 0,001%.

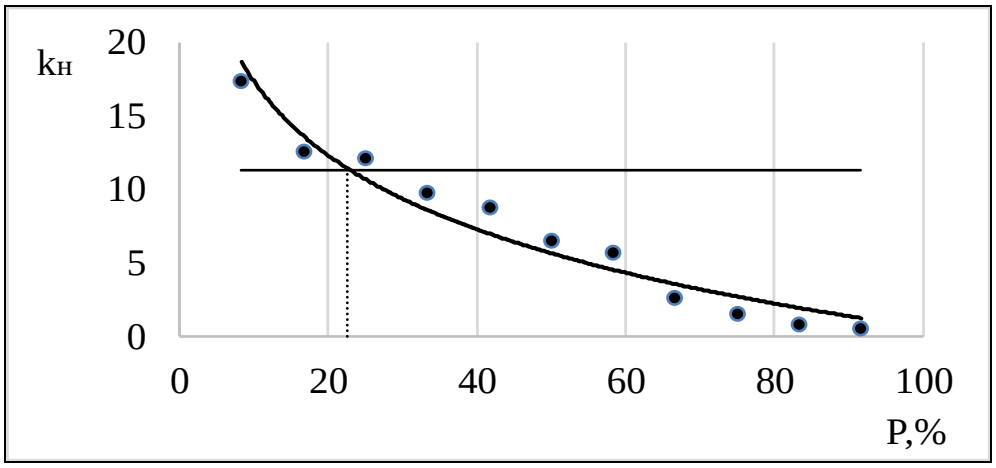

Рисунок 1.4 – Емпірична крива розподілу забезпеченостей критерію чутливості $k_н$ до забруднення сполуками азоту у пункті спостережень р. Киргиж-Китай –с. Малоярославець (*Loboda, N. & Daus, M. 2021)*

Емпіричні криві забезпеченості не дають можливості визначити величини стоку за межами вихідної інформації. Безпосередньо графічна екстраполяція емпіричної кривої забезпеченості у область малих (P < 5%) та великих (P > 95%) забезпеченостей має суб’єктивний характер і може

привести до значних похибок у розрахунках. Теоретичні закони розподілу у даному випадку "вирівнюють" емпіричний розподіл стокової величини та екстраполюють його за межі спостережених даних.

Контрольні запитання.

1. Дати визначення випадкової величини.
2. Дати визначення закону розподілу випадкової величини.
3. Дати визначення статистичних параметрів.
4. Дати визначення забезпеченості випадкової величини.
5. Дати визначення математичному сподіванню.
6. Дати визначення дисперсії.
7. Описати розрахунки емпіричної забезпеченості.

ЛЕКЦІЯ 2
«МЕТОД СУМІСНОГО АНАЛІЗУ ДАНИХ»

Статистичні параметри випадкових величин містять у собі основну інформацію про їх ймовірнісний закон розподілу (щільність ймовірності, інтегральна функція розподілу, функція забезпеченості).

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

Обмеженість у часі наявних спостережень породжує статистичну нестійкість визначення параметрів, що може бути ефективно компенсоване за рахунок додаткової інформації про просторові закономірності розподілу розглядуваних характеристик. Це означає, що статистичні параметри, визначені за рядом спостережень з великою похибкою, можуть бути уточнені шляхом приєднання додаткової інформації по іншим водним об'єктам, розташованим неподалік (звідки і пішла назва «метод сумісного аналізу даних»).

Прикладом застосування статистичних методів до просторового узагальнення гідрохімічних характеристик є гідрохімічне картування описане у роботі (*Пеляшенко, 1979*). Для 22 гідрохімічних полів наведені осереднені оцінки вмісту хімічних компонентів, які визначені за об'єднаною вибіркою для кожної компоненти окремо. Таке просторове узагальнення має назву районування. Можливість об'єднання даних по різних точках поля була обґрунтована за допомогою перевірки гіпотез про статистичну однорідність рядів (їх приналежність до однієї генеральної сукупності).

У роботі (*Крицкий С.Н., Менкель М.Ф., 1982*) запропонований підхід, який базується на розгляді складових просторової дисперсії досліджуваної характеристики. Цей підхід був застосований Лободою Н.С. для теоретичного обґрунтування способу просторового узагальнення (районування чи картування) статистичних параметрів річного стоку (*Лобода Н.С., 2005*). Суть методу зводиться до визначення географічної і випадкової складових загальної просторової дисперсії розглядуваного статистичного параметра A:

$$\sigma_{П}^2 = \sigma_{Г}^2 + \sigma_{B}^2, \qquad (2.1)$$

де $\sigma_{П}^2$ – повна складова дисперсії параметра;

$\sigma_{Г}^2$ – географічна складова дисперсії параметра;

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

При цьому повна просторова дисперсія параметра оцінюється за формулою

$$\sigma_{П}^2 = \frac{\sum_{j=1}^{k} (A_j - A_{CEE})^2}{k-1}, \qquad (2.2)$$

де k – число об’єктів, об'єднаних в одну групу;

j – порядковий номер розглядуваного об’єкту;

A_j – індивідуальна оцінка параметра (оцінка, виконана для окремого об’єкту);

A_{CEP} – осереднена в межах виділеної групи оцінка параметра.

Випадкова складова просторової дисперсії параметра визначається як осереднена за групою виділених об'єктів дисперсія індивідуальної оцінки параметра

$$\sigma_{B}^2 = \frac{\sum_{j=1}^{k} \sigma_{Aj}^2}{k}, \qquad (2.3)$$

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

Географічна складова знаходиться за допомогою зворотного розрахунку з (2.1):

$$\sigma_{Г}^2 = \sigma_{П}^2 - \sigma_{B}^2. \qquad (2.4)$$

Якщо виконується умова

$$\frac{\sigma_B^2}{\sigma_П^2} > \frac{\sigma_Г^2}{\sigma_П^2}, \qquad (2.5)$$

то можна зробити висновок, що просторовий розподіл досліджуваного параметра в більшій мірі визначається випадковими властивостями поєднуваних вибірок і в меншій – зміною фізико-географічних умов по території. Таким чином, при виконанні (2.5) 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.

Вибір способу узагальнення досліджуваних характеристик здійснюється на основі отриманих результатів:

- якщо внесок випадкової складової у просторову дисперсію більше або дорівнює 70%, тобто $\frac{\sigma_B^2}{\sigma_П^2} \geq 70\%$, то робиться висновок про можливість районування (що означає осереднення досліджуваного параметру);

- якщо внесок географічної складової у просторову дисперсію більше або дорівнює 70%, тобто $\frac{\sigma_Г^2}{\sigma_П^2} \geq 70\%$, то робиться висновок про можливість картування (що означає представлення досліджуваного параметру у вигляді карт ізоліній);

- якщо внесок випадкової та географічної складових у просторову дисперсію менше 70%, то робиться висновок про необхідність уточнення даних.

Коли оцінки вибіркових параметрів дуже великі, географічна складова дисперсії, обчислена зворотним розрахунком за (2.4), 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

$$\sigma_{CEE} = \sqrt{\frac{\sigma_B^2}{k} + \sigma_\Gamma^2} \tag{2.6}$$

Величина σ_{CEP} поряд з умовою (2.5) також є критерієм якості об'єднання. *Осереднена оцінка параметра визнається статистично достовірною, коли виконується умова*

$$A_{CEE} > 2\sigma_{CEE} \tag{2.7}$$

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_δ^2 дисперсія осередненого у межах поєднуваної сукупності параметра σ_{CEE}^2 відповідно до виразу (2.6) повинна зменшуватися. "Сплеск" в убутній функції $\sigma_{CEP}^2 = \varphi(k)$ свідчить про те, що досліджуваний параметр k-того ряду стоку значно відрізняється від осередненої оцінки та значень відповідного параметру 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

$$\varepsilon_A = \frac{\sigma_A}{A} \cdot 100\%, \tag{2.8}$$

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.

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

$$A_j^{'} = \frac{A_j \sigma_{CEP}^2 + A_{CEE} \sigma_j^2}{\sigma_{CEP}^2 + \sigma_j^2} , \qquad (2.9)$$

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;

σ_j^2 – дисперсія параметра A_j по j-тому розглянутому об'єкту;

A_{CEE} – осереднена в межах виділеної групи об'єктів оцінка параметра A;

σ_{CEE}^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

$$\sigma_j^{'} = \frac{\sigma_j \sigma_{CEP}}{\sqrt{\sigma_j^2 + \sigma_{CEP}^2}} , \qquad (2.10)$$

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

σ_j – середнє квадратичне відхилення параметра A_j по j-тому розглянутому об'єкту.

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 (2.5). 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 (2.7) та аналіз залежності $\sigma^2_{CEE} = \varphi(k)$. Зростання географічної складової повної просторової дисперсії параметра буде тим інтенсивнішим, чим ширше межі просторового узагальнення.

У прикладі (табл.2.1) показані результати застосування методу сумісного аналізу до гідрохімічних показників річок нижньої течії Дністра, звідки витікає, що можуть бути осереднені у межах досліджуваної території дані по $БСК_5$, гідрокарбонатам, нафтопродуктам, нітритам, СПАР. Середні багаторічні значення концентрацій таких гідрохімічних компонентів як завислі речовини, калій, кальцій, магній, сульфати, хлориди, фосфати можуть бути представлені у вигляді карт ізоліній. Інші компоненти мають бути уточнені з використанням формули (2.9).

Таблиця 2.1 – Оцінки складових просторової дисперсії для гідрохімічних компонент нижньої течії Дністра (2000-2012 рр.)

Показник	Складові дисперсії, %	
	$\frac{\sigma^2_B}{\sigma^2_П}$	$\frac{\sigma^2_Г}{\sigma^2_П}$
1	**2**	**3**
Водневий показник	56,5	43,5
Азот амонійний	42,1	57,9
$БСК_5$	100	0
Гідрокарбонати	100	0
Завислі речовини	23,6	76,4
Калій	6,62	93,4
Кальцій	6,28	93,72
Магній	2,30	97,70
Натрій	2,31	97,69
Нафтопродукти	100	0
Нітрити	87,17	12,82
Сухий залишок	4,06	95,93
Розчинений кисень	52,4	47,6
СПАР	84,13	15,87

Продовження таблиці 2.1

1	2	3
Сульфати	1,90	98,1
Фосфати	14,6	85,4
Хлориди	1,92	98,07
ХСК	8,02	91,98

Контрольні запитання

1. Записати загальну просторову дисперсію параметра А через її складові.
2. Визначення випадкової складової просторової дисперсії параметра А.
3. Визначення географічної складової просторової дисперсії параметра А.
4. При яких співвідношеннях між географічною та випадковою складовою приймається рішення про можливість картування параметра А.
5. При яких співвідношеннях між географічною та випадковою складовою приймається рішення про можливість районування параметра А.
6. При яких співвідношеннях між географічною та випадковою складовою приймається рішення про уточнення значень параметра А по об’єктах, які увійшли до об’єднуваної виборки.

ЛЕКЦІЯ 3
«ЗАГАЛЬНІ УЯВЛЕННЯ ПРО ТЕОРІЮ ВИПАДКОВИХ ПРОЦЕСІВ»

3.1 Поняття про випадкову функцію

На практиці доводиться оперувати випадковими величинами, статистичний розподіл яких змінюється у процесі проведення випробувань. Такі величини одержали назву випадкових функцій. Вивченням подібних випадкових явищ, в яких випадковість набуває форми процесу, займається спеціалізований розділ теорії ймовірностей – теорія випадкових функцій або стохастичних процесів (*Лобода Н.С., Гопченко Є.Д., 2006*).

Випадкова функція у найпростішому виді може бути представленою як система випадкових величин. Таку систему розглядають як єдине ціле, яке описується диференціальним або інтегральним законом розподілу.

Випадковою функцією називається функція, яка в разі випробувань може набрати того чи іншого конкретного вигляду, наперед невідомо, якого саме. Конкретний вигляд, якого набуває випадкова функція в результаті випробувань, називається реалізацією випадкової функції.

Значення випадкової функції встановлюються за допомогою випробувань і можуть бути різними в залежності від ходу випробувань.

Якщо аргументом випадкової функції є час, то для її позначення використовується термін «випадковий процес».

Випадковою функцією аргументу t ***називається функція, ординати якої для будь-яких фіксованих значень*** t ***є випадковими величинами. У загальному випадку випадкова функція складається з системи реалізацій*** $x_i(t)$ $(i=\overline{1,n})$***, які відображають результати***

окремих експериментів. Випадкові функції при їх перетині $t = t_j$ ***утворюють сукупність точок, які є випадковими величинами.***

Випадкову функцію вивчають аналітичним способом, заснованим на визначенні багатовимірного закону її розподілу, і статистичним, заснованим на визначенні тільки числових характеристик такої функції. Перший спосіб застосовують у теоретичних дослідженнях. Другий спосіб широко використовують для вирішення різних прикладних задач у теорії випадкових функцій.

Розглянемо випадкову функцію $X(t)$, над якою проведено n незалежних випробувань і отримано n реалізацій. Кожна реалізація є звичайною, тобто невипадковою функцією (рис.3.1).

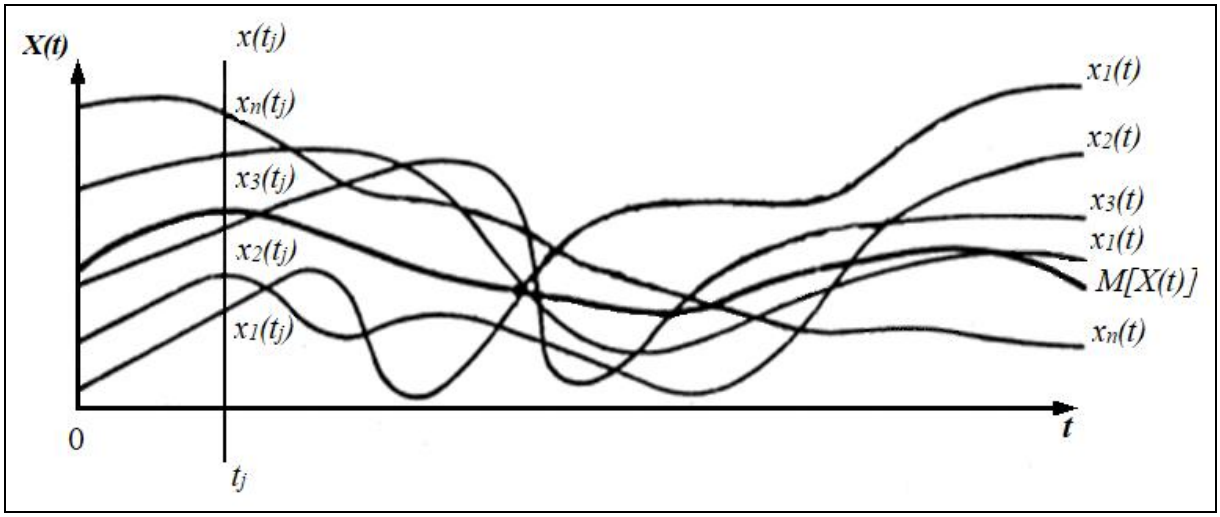

Рисунок 3.1 – Сімейство реалізацій випадкової функції

Реалізації випадкової функції можна отримати за допомогою самописного приладу, осцилограми запису змін процесу в часі, кінозйомки, періодичної фотозйомки та ін.

3.2 Закон розподілу випадкової функції

Якщо зафіксувати деяке значення аргументу t_j, *то при заданому значенні аргументу* t_j *(рис.3.1) випадкова функція перетворюється на випадкову величину* $X(t_j)$, *яка називається перетином випадкової функції,*

що відповідає моменту t_j. Значеннями цієї випадкової величини будуть значення $x_1(t_j), x_2(t_j), ..., x_n(t_j)$, де n – кількість реалізацій.

Таким чином, випадкова функція містить у собі риси випадкової величини і функції. Якщо зафіксувати значення аргументу, то отримуємо випадкову величину. В результаті кожного випробування випадкова функція, у свою чергу, перетворюється на звичайну невипадкову функцію, яка має назву реалізації.

3.3 Характеристики випадкових функцій та їх визначення

Основними числовими характеристиками випадкового процесу є: ***математичне сподівання*** $m_x(t)$, яке визначає таку функцію, навколо якої групуються реалізації випадкової функції; ***дисперсія*** $D_x(t)$, що показує ступінь розсіювання цих реалізацій відносно математичного сподівання $m_x(t)$, ***коваріаційна*** $K_{xx}(t_i, t_j)$ та ***кореляційна*** $r_{xx}(t_i, t_j)$ ***функції***, які визначають внутрішню структуру випадкового процесу і характеризують ступінь взаємного зв'язку між перетинами випадкової функції $X(t)$.

Якщо стоїть задача дослідження взаємозв'язку між двома випадковими функціями $X(t)$ та $Y(t)$, то його характер визначають взаємна коваріаційна $K_{xy}(t_i, t_j)$ та взаємна кореляційна функції $r_{xy}(t_i, t_j)$.

На відміну від числових характеристик випадкових величин, характеристики випадкових функцій являють собою не числа, а функції.

Математичне сподівання випадкової величини визначається в такий спосіб. Розглянемо перетин випадкової функції $X(t)$ при фіксованому t. У цьому перетині матимемо звичайну випадкову величину; визначимо її математичне сподівання. Очевидно, у загальному випадку воно залежить від t, тобто являє собою деяку функцію t:

$$m_x(t) = M[X(t)]. \tag{3.1}$$

Таким чином, *математичним сподіванням випадкової функції* $X(t)$ *називається невипадкова функція* $m_x(t)$, *яка при кожному значенні аргументу* t *дорівнює математичному сподіванню відповідного перетину*

випадкової функції. На рис.3.1 жирною лінією показаний хід у часі математичного сподівання.

Дисперсією випадкової функції $X(t)$ *називається невипадкова функція* $D_x(t)$*, значення якої для кожного аргументу* t *дорівнюють дисперсії відповідного перетину випадкової функції*

$$D_X(t) = D[X(t)]. \tag{3.2}$$

Дисперсія $D_x(t)$ є невід'ємною функцією. Добуваючи з неї квадратний корінь, одержимо функцію $\sigma_x(t)$ – середнє квадратичне відхилення випадкової функції

$$\sigma_x(t) = \sigma[X(t)] = \sqrt{D[X(t)]}. \tag{3.3}$$

Досить часто користуються центрованою характеристикою випадкової функції

$$\delta[X(t)] = X(t) - M[X(t)]. \tag{3.4}$$

Очевидно, що

$$M\delta[X(t)] = 0. \tag{3.5}$$

Нормована випадкова функція представляється у вигляді

$$X_0(t) = \delta[X(t)] / \sigma[X(t)], \tag{3.6}$$

для якої справедливі рівняння

$$M[X_0(t)] = 0; \quad D[X_0(t)] = 1. \tag{3.7}$$

Дисперсія випадкової функції характеризує розсіювання реалізацій відносно функції математичного сподівання. Не виключена можливість однакової дисперсії для всіх значень аргументу t, тобто $D[X(t)] = D(X)$. Якщо така дисперсія дорівнює нулю, то можна вважати, що випадкова функція $X(t) = M[X(t)]$ має ймовірність появи, яка дорівнює одиниці.

Бувають випадки, коли математичні сподівання випадкових функцій однакові, а їхні дисперсії різні. Наприклад, на рис. 3.2 показані реалізації випадкових функцій $X(t), Y(t), Z(t)$, що мають однакові математичні сподівання, але різні дисперсії.

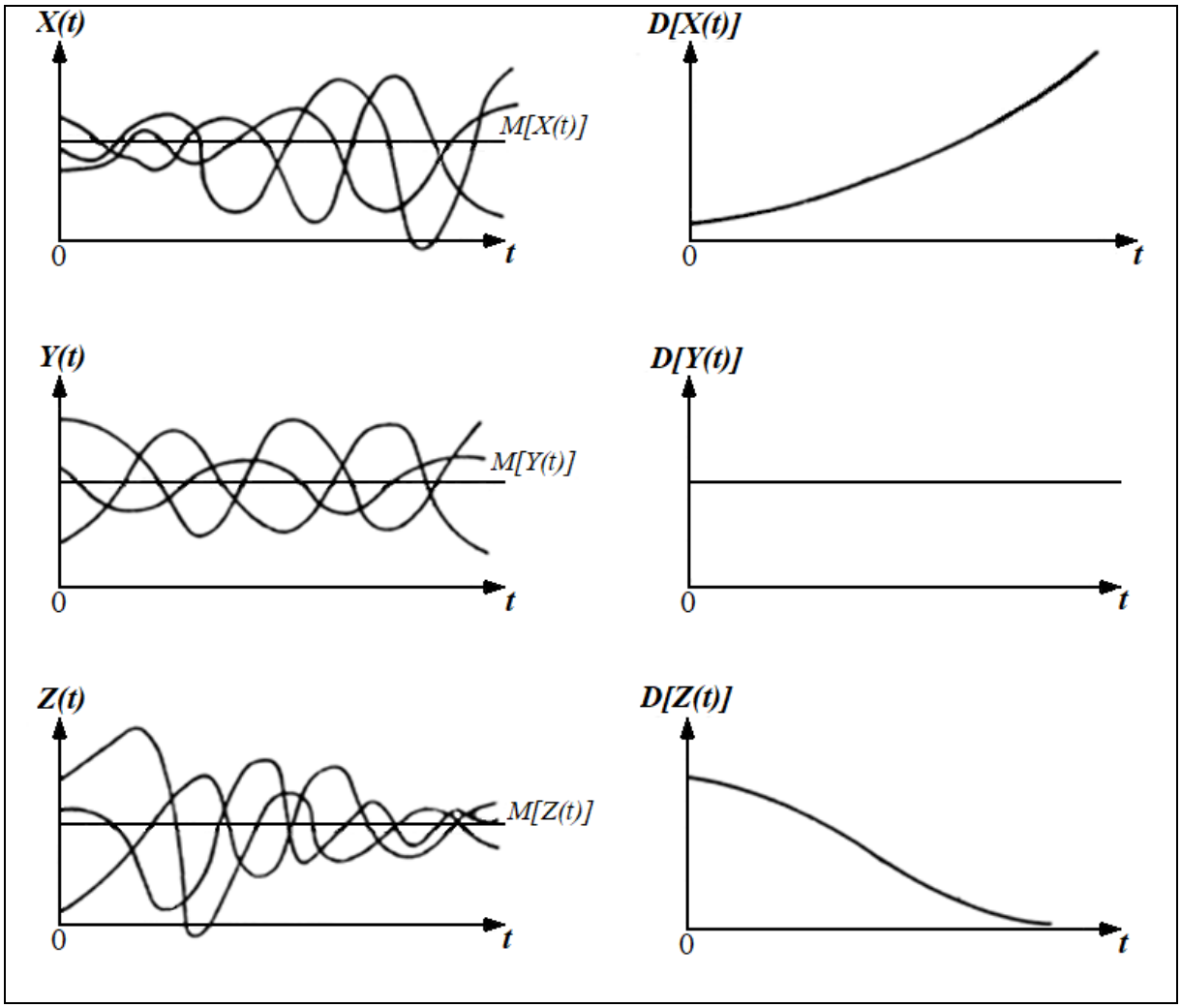

Рисунок 3.2 – Зміна дисперсії у часі для різних випадкових процесів (*Н.Н. Кузьменко, Ю.В. Поліщук, Л.А. Шаповалова, 1980*)

Математичне сподівання і дисперсія визначають тільки смугу можливих реалізацій випадкової функції, але не поводження реалізацій усередині такої смуги, як це показано на рисунках 3.3 та 3.4. Математичне сподівання та дисперсія цих випадкових функцій однакові, але характер реалізацій та зв'язок між ними різний.

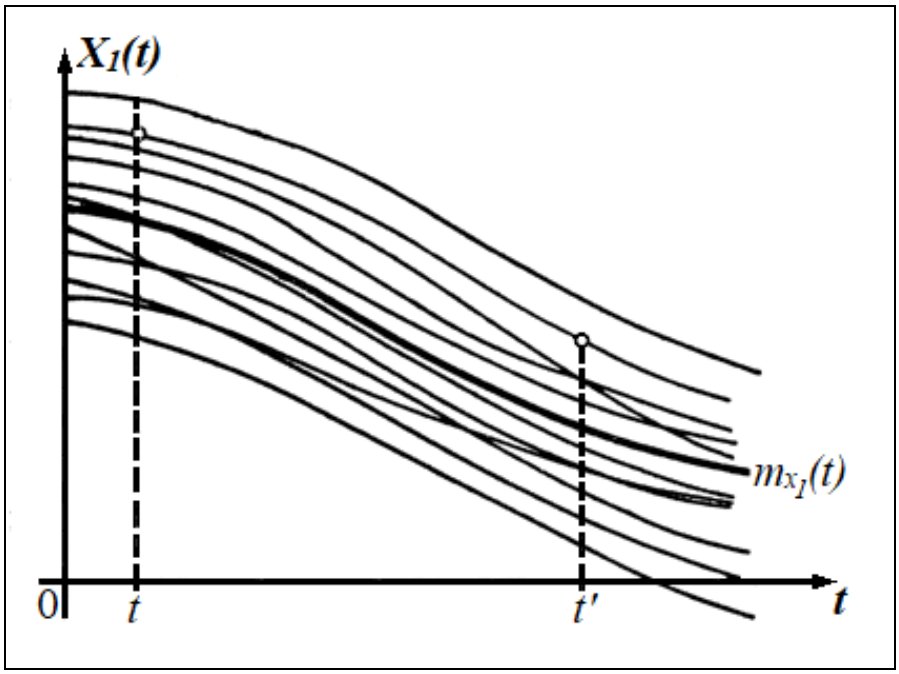

Рисунок 3.3 – Реалізації випадкової функції $X_1(t)$

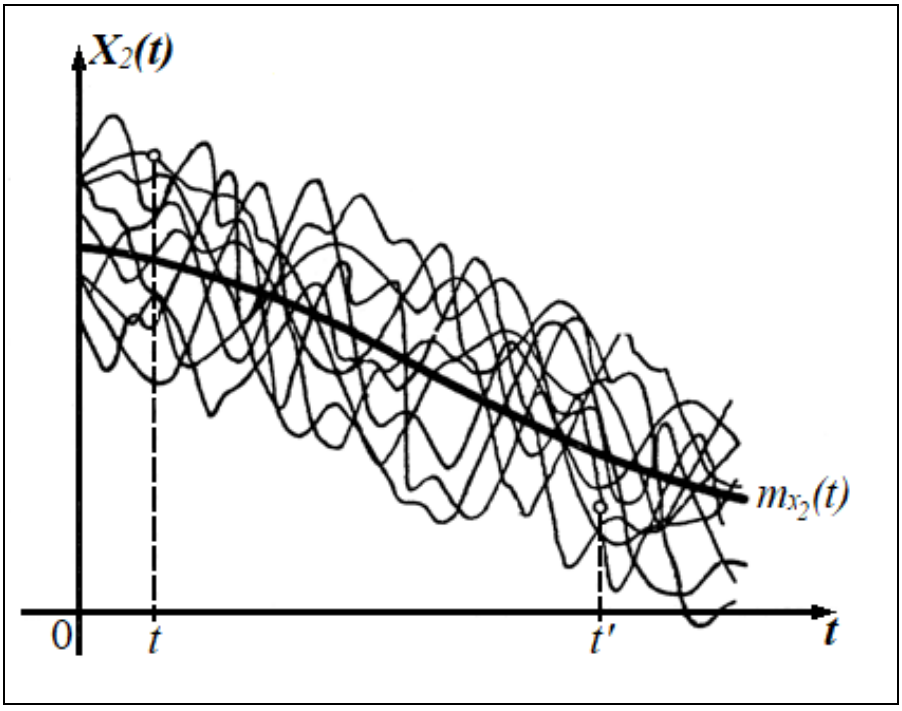

Рисунок 3.4 – Реалізації випадкової функції $X_2(t)$

Зв'язок між випадковими величинами $X(t_i)$ і $X(t_j)$ у різних перетинах випадкової функції можна виразити за допомогою коваріаційного моменту чи коефіцієнта кореляції. Ступінь залежності між перетинами випадкової функції при різних аргументах t характеризується коваріаційною або кореляційною функціями.

Коваріаційним моментом випадкової функції (коваріаційною функцією) називається невипадкова функція двох аргументів $K_{xx} = K_{xx}(t_i, t_j) = K_x$*, яка для кожної пари перетинів випадкової функції відповідних аргументів* t_i *і* t_j *дорівнює коваріаційному моменту між випадковими величинами у цих двох перетинах*

$$K_x = K\left[X(t_i), X(t_j)\right] = M\left\{\delta\left[X(t_i)\right] \cdot \delta\left[X(t_j)\right]\right\}. \tag{3.8}$$

На практиці також використовують кореляційну функцію випадкової функції, яка являє собою нормований кореляційний момент

$$r_x = r\left[X(t_i), X(t_j)\right] = \frac{K\left[X(t_i), X(t_j)\right]}{\sigma(t_i) \cdot \sigma(t_j)}. \tag{3.9}$$

Кореляційну функцію випадкового процесу часто називають ***автокореляційною***, що означає кореляцію процесу із самим собою.

При $t_i = t_j = t$ одержимо

$$K\left[X(t), X(t)\right] = M\left[\delta^2(t)\right] = D_x(t); \tag{3.10}$$

$$r_x = r\left[X(t), X(t)\right] = \frac{K\left[X(t), X(t)\right]}{\sigma(t) \cdot \sigma(t)} = \frac{D_x(t)}{D_x(t)} = 1, \tag{3.11}$$

тобто при $t_i = t_j$ коваріаційна функція стає дисперсією випадкової функції, а кореляційна функція дорівнює 1.

Таким чином, при $t_i = t_j$ необхідність в дисперсії, як в окремій характеристиці випадкової функції, відпадає: як основну характеристику випадкової функції достатньо розглядати її математичне сподівання й кореляційну функцію.

Коваріаційний момент двох випадкових величин $X(t)$ і $X(t')$ не залежить від послідовності, в якій ці величини розглядаються, отже, коваріаційна функція симетрична відносно своїх аргументів, тобто не змінюється при зміні аргументів місцями

$$K_x(t,t') = K_x(t',t). \tag{3.12}$$

Інтервал $(t_i - t_j)$ називають «запізнюванням» чи «зсувом».

Очевидно, що значення коваріаційної і автокореляційної функцій залежать від інтервалу між t_i і t_j. Вони зменшуються із збільшенням цього інтервалу. При одних і тих же математичних сподіваннях і дисперсіях дві випадкові функції можуть мати різні автокореляційні функції. Автокореляційна функція може бути залежною не від окремих значень t_i і t_j, а від різниць $t_i - t_j$. Якщо при одному із значень t_i чи t_j випадкова функція $X(t)$ стає невипадковою величиною, то $K[X(t_i), X(t_j)] = 0$ при будь-якому значенні іншого аргументу.

Нехай над випадковою функцією $X(t)$ зроблено m незалежних випробувань (спостережень), у результаті отримано m реалізацій випадкової функції. Якщо число реалізацій дорівнює m, то для кожного перетину t_i можна обчислити емпіричні (статистичні) оцінки характеристик випадкової функції $X(t)$.

Так, оцінкою математичного сподівання в перетині t_i є просте середнє арифметичне

$$\bar{x}(t_i) = \frac{\sum_{j=1}^{m} x(t_i)}{m}. \tag{3.13}$$

Оцінка дисперсії надається за формулою

$$\hat{\sigma}_x^2(t_i) = S_x^2(t_i) = \frac{1}{m}\sum_{j=1}^{m}\{\delta[X(t_i)]\} = \frac{\sum_{j=1}^{m}\left[x(t_i) - \bar{x}(t_i)\right]^2}{m-1}. \tag{3.14}$$

Для коваріаційних та кореляційних функцій розрахунок виконується за такими формулами:

$$\hat{K}_x(t_i,t_k)=\frac{\sum_{j=1}^{m}\left[x(t_i)-\bar{x}(t_i)\right]\cdot\left[x(t_k)-\bar{x}(t_k)\right]}{(m-1)};\qquad(3.15)$$

$$\hat{r}_x(t_i,t_k)=\frac{\sum_{j=1}^{m}\left[x(t_i)-\bar{x}(t_i)\right]\cdot\left[x(t_k)-\bar{x}(t_k)\right]}{(m-1)\sigma_x(t_i)\sigma_x(t_k)}.\qquad(3.16)$$

3.4 Стаціонарність випадкових процесів

Випадкові процеси, що протікають у часі приблизно однорідно, називають стаціонарними. Вони мають вигляд безперервних коливань відносно деякого середнього значення. З часом середня амплітуда і характер коливань істотно не змінюються.

Математичне сподівання стаціонарного випадкового процесу не залежить від аргументу t ***і є постійною величиною.***

Коваріаційна функція стаціонарного випадкового процесу є функцією тільки одного аргументу $\tau=t_2-t_1$ (зсуву).

Стаціонарний випадковий процес повинен задовольняти трьом умовам:

$$M[X(t_s)]=const,\ D[X(t_s)]=const;\ K_X[X(t_s),(t_k)]=K(t_i-t_j)=K_\tau.\quad(3.17)$$

Таким чином, *випадкова функція називається стаціонарною, якщо усі її статистичні характеристики не залежать від* t, точніше, не змінюються при будь-якому зсуві аргументів, від яких вони залежать.

Виконання другої умови з (3.17) є частинним випадком третьої. Для стаціонарного випадкового процесу характерна незалежність значень коваріаційної функції від положення відрізка τ на осі t. Коваріаційна функція повинна залежати тільки від довжини проміжку τ. Покладемо $t_s=t;\ t_k=t+\tau$. Якщо $\tau=0$, то

$$D_x(t)=K_X[X(t),X(t+0)]=K_X(\tau=0)=const.\qquad(3.18)$$

Іншими словами, у випадку стаціонарного випадкового процесу його дисперсія є ніщо інше, як значення коваріаційної функції при $\tau = 0$, тобто вона не є функцією аргументу. Слід зазначити, що з властивості симетричності коваріаційної і кореляційної функцій

$$K_x(\tau) = K_x(-\tau); \tag{3.19}$$

$$r_x(\tau) = r_x(-\tau) \tag{3.20}$$

випливає незалежність цих функцій від знака величини τ, тобто можна вважати $\tau = |t_2 - t_1|$. Тому коваріаційну і кореляційну функції знаходять тільки для додатного аргументу.

3.5 Ергодичність стаціонарних випадкових процесів

Переважній більшості стаціонарних випадкових функцій властива ергодичність. ***Властивість ергодичності полягає у тому, що кожна окрема реалізація стаціонарної випадкової функції є повноправним представником усієї сукупності можливих реалізацій. Одна реалізація достатньої довжини може замінити при статистичній обробці всі реалізації в цілому. Статистичні характеристики кожної реалізації є одними й тими ж для усіх реалізацій.***

Розглянемо дві стаціонарні випадкові функції $X(t)$ і $Y(t)$, наведені на рис.3.5. Для функції $X(t)$ кожна із реалізацій має своє математичне сподівання і дисперсію, які не співпадають із математичним сподіванням процесу у цілому, Така функція не є ергодичною. Кожна з реалізацій випадкової функції $Y(t)$ може бути охарактеризована однаковим середнім арифметичним, дисперсією. Така функція може розглядатися як ергодична.

Розглянемо стаціонарну, ергодичну функцію *X(t)*. Осереднюючи значення реалізації ергодичної функції за часом, одержують оцінку математичного сподівання

$$\hat{m}_x \approx \frac{1}{T}\int_0^T x(t)dt, \tag{3.21}$$

коваріаційної функції

$$\hat{K}_x(\tau) \approx \frac{1}{T-\tau} \int_0^{T-\tau} [x(t) - \hat{m}_x] \cdot [x(t+\tau) - \hat{m}_x] dt \ . \qquad (3.22)$$

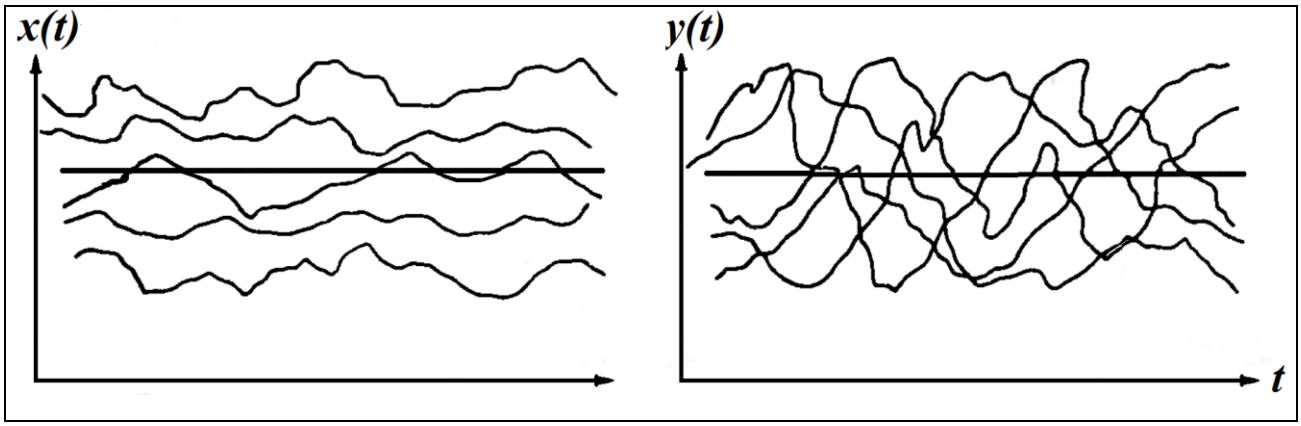

Рисунок 3.5 – Ергодична ***x(t)*** та неергодична ***y(t)*** стаціонарні випадкові процеси

Властивість ергодичності має велике практичне значення, оскільки при її виконанні для визначення статистичних характеристик випадкових функцій нема необхідності у великій кількості реалізацій.

Для дискретних випадкових величин визначення коваріаційної, кореляційної та взаємної кореляційної функції відбувається за такими формулами

$$K_{xx}(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{(n-\tau-1)}, \qquad (3.23)$$

$$R_{xx}(\tau) = r_{xx}(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{(n-\tau-1)\sigma_x^2}, \qquad (3.24)$$

$$R_{xy}(\tau) = r_{xy}(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y})}{(n-\tau-1)\sigma_x \sigma_y}, \qquad (3.25)$$

де $K_{xx}(\tau)$ – коваріаційна функція;

$R_{xx}(\tau) = r_{xx}(\tau)$ – кореляційна функція;

$R_{xy}(\tau) = r_{xy}(\tau)$ – взаємна кореляційна функція;

n – число членів вихідного ряду спостережень;

τ – зсув ряду по відношенню до самого себе (запізнювання);

x_i – член ряду від x_i до $x_{n-\tau}$;

$x_{i+\tau}$ – члени ряду від $x_{i+\tau}$ до x_n;

$\bar{y}, \sigma_y$ – відповідно середнє і стандарт вихідного ряду Y;

$\bar{x}, \sigma_x$ – відповідно середнє і стандарт вихідного ряду X.

Контрольні запитання

1. Дати визначення випадкової функції.
2. Дати визначення стаціонарного випадкового процесу.
3. Дати визначення стаціонарного ергодичного процесу.
4. Записати рівняння коваріаційної функції $K_{xx}(\tau)$.
5. Яку властивість стаціонарного ергодичного процесу характеризує коваріаційна функція.

ЛЕКЦІЯ 4
«ВНУТРИРЯДНА КОРЕЛЯЦІЯ, АВТОКОРЕЛЯЦІЙНА ФУНКЦІЯ, КОЕФІЦІЄНТ АВТОКОРЕЛЯЦІЇ»

У гідроекологічних розрахунках річковий стік розглядається як ергодичний процес з річним періодом. Така гіпотеза зручна, оскільки по кожному річковому створу є лише одна реалізація процесу стоку, яка представлена гідрологічними спостереженнями. Однак, цю гіпотезу не можна прийняти, якщо вимірювати час геологічними масштабами. В межах гідрологічних спостережень за умови незначних антропогенних навантажень, порушення ергодичності стоку маловірогідні.

Одним із традиційних методів дослідження зв`язків між послідовними членами ряду річного стоку є *кореляційний аналіз.* ***Ступінь залежності між суміжними членами ряду характеризується коефіцієнтом автокореляції.***

Ординати кореляційної функції за вибірковими даними обчислюються з урахуванням зсуву у часі за такою формулою:

$$r_{xx}(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sigma_x^2 (n - \tau - 1)}, \qquad (4.1)$$

де n – число членів початкового ряду спостережень;

τ – зсув ряду по відношенню до самого себе при підрахунку коефіцієнта кореляції (запізнювання);

x_i – член ряду від x_i до $x_{n-\tau}$;

$x_{i+\tau}$ – члени ряду від $x_{i+\tau}$ до x_n;

$\bar{x}, \sigma_x$ – відповідно середнє і стандарт початкового ряду.

Кореляційна функція оцінює зв’язки між послідовними членами одного й того ж самого ряду, а тому має назву **автокореляційної** функції.

При τ=0 формула розрахунку автокореляційної функції набуває виду

$$r_{xx}(\tau = 0) = \frac{\sum_{i=1}^{n-0} (x_i - \bar{x})(x_{i+0} - \bar{x})}{\sigma_x^2 (n - 0 - 1)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}{\sigma_x^2 (n - 1)} = \frac{\sigma_x^2}{\sigma_x^2} = 1. \qquad (4.2)$$

Чим більший зсув між членами ряду, тим менше відповідне значення автокореляційної функції. Це свідчить про послаблення лінійних зв'язків між членами ряду. Із збільшенням відстані між членами ряду значення автокореляційної функції наближаються до нуля.

На рисунку 4.1 показана автокореляційна функція миттєвих швидкостей потоку (n=300). Вона використана для визначення розміру турбулентного потоку. Для опису турбулентних структур потоку використовується уявлення про те, що турбулентний потік є сукупністю накладених на усереднений рух води збурень (вихорів) різних масштабів. Масштаби великих вихорів мають порядок довжини, яка відповідає шляху перемішування, вони спільномірні з лінійними розмірами потоків.

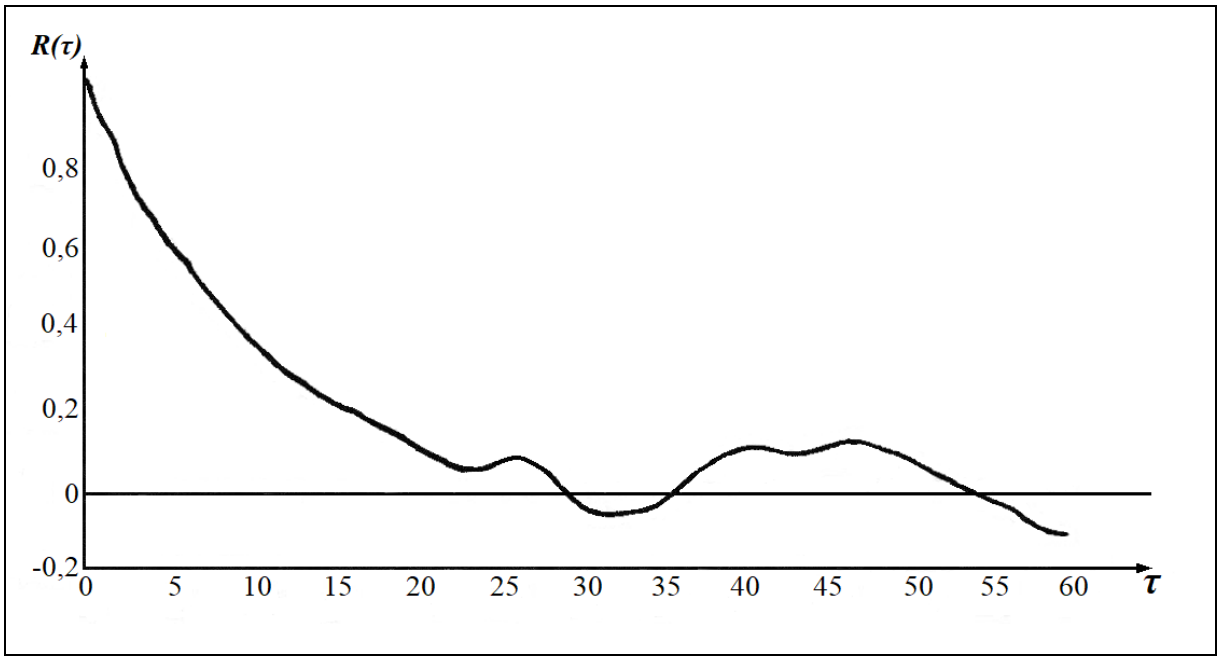

Рисунок – 4.1 Автокореляційна функція миттєвих швидкостей потоку

Зовнішній масштаб (розмір) турбулентних збурень визначається за формулою

$$l = \bar{u}\,\tau_0, \qquad (4.3)$$

де l – зовнішній масштаб турбулентності; $\bar{u}$ - усереднена швидкість;

τ_0 – час нульової кореляції (час, при якому кореляційна функція дорівнює нулю).

Використовуючи графік на рис.4.1 можна визначити τ_0 як відстань від початку координат до першого перетину кривої $R(\tau)$. Визначена величина τ_0 називається ***радіусом кореляції***.

Навіть при відносно тривалих рядах стоку (50-60 років) недостовірні як окремі ординати автокореляційної функції, так і її контури. Тому при розрахунках річного стоку його ряди розглядаються як простий ланцюг Маркова, тобто враховуються кореляційні зв`язки тільки між стоком суміжних років. Виходячи з вище сказаного, з усіх ординат автокореляційної функції береться до уваги тільки перша, яка відповідає $\tau = 1$.

При $\tau = 1$ автокореляційна функція розраховується наступним чином

$$r_{xx}(\tau = 1) = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sigma^2 (n-1)}, \tag{4.4}$$

$$r_{xx}(\tau = 1) = r_{(1)} = \frac{\sum_{i=1}^{n-1}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n-1}(x_i - \bar{x})^2}. \tag{4.5}$$

Значення автокореляційної функції при зсуві $\tau = 1$ оцінює тісноту лінійного зв'язку між суміжними членами ряду і називається коефіцієнтом автокореляції.

Середня квадратична похибка розрахованого коефіцієнта автокореляції визначається за рівнянням

$$\sigma_{r(1)} = \frac{1 - r(1)^2}{\sqrt{n-1}}. \tag{4.6}$$

Розрахункове значення коефіцієнта автокореляції є значущим, якщо виконується умова

$$r(1) \geq 2\sigma_{r(1)}. \tag{4.7}$$

Контрольні запитання

1. Яку властивість стаціонарного ергодичного процесу характеризує автокореляційна функція.
2. Як змінюється автокореляційна функція із збільшенням часового зсуву τ
3. Дати визначення коефіцієнту автокореляції
4. Яку властивість ряду спостережень характеризує коефіцієнт автокореляції.

ЛЕКЦІЯ 5
«СТРУКТУРНА ФУНКЦІЯ»

Структурну функцію визначають як математичне сподівання квадрата різниці перерізів випадкової стаціонарної ергодичної функції. Якщо реалізація ергодичної випадкової функції задана в дискретних точках вимірювання, то структурна функція має вигляд

$$M(s)=\frac{1}{(n-s)}\sum_{i=1}^{n-s}\left(\Delta x_i-\Delta x_{i+s}\right)^2, \qquad (5.1)$$

де n – число вимірювань;

$\Delta x_i = x_i - \bar{x}$; $\Delta x_{i+1} = x_{i+1} - \bar{x}$ – центровані значення ряду x;

σ_x^2 – дисперсія;

s – зсув (відстань між перерізами) у часі або просторі;

Представимо структурну функцію у такому записі

$$M(s)=\frac{1}{(n-s)}\sum_{i=1}^{n-s}\left[(x_i-\bar{x})-\left(x_{i+s}-\bar{x}\right)\right]^2, \qquad (5.2)$$

звідки можна отримати такий вираз

$$M(s)=\frac{1}{(n-s)}\sum_{i=1}^{n-s}\left(\Delta x_i-\Delta x_{i+s}\right)^2=2\left[\sigma_x^2-K(s)\right] \qquad (5.3)$$

де $K(s)$ – коваріаційна функція, яка розраховується за формулою:

$$K(s)=\frac{1}{(n-s)}\sum_{i=1}^{n-s}\left(\Delta x_i\cdot\Delta x_{i+s}\right). \qquad (5.4)$$

Нормована автоковаріаційна функція називається автокореляційною і має вигляд

$$R(s)=\frac{K(s)}{\sigma_x^2} \qquad (5.5)$$

Коли $s=0$, $R(s)=1$, оскільки автоковаріаційна функція при $s=0$ дорівнює дисперсії $R(0)=\sigma_x^2$. Автокореляційна функція стаціонарної випадкової функції симетрична $R(s)=R(-s)$ і є функцією лише різниці двох аргументів $t_1-t_2=s$.

Нормована структурна функція записується у вигляді

$$m(s)=\frac{K(s)}{\sigma_x^2}=2\left[1-R(s)\right] \tag{5.6}$$

й змінюється від 0 до 2.

Якщо розглянути область від $R(s)=1$ до $R(s)=0$, коли s змінюється від 0 до $+\infty$, то справедливим буде такий запис нормованої структурної функції

$$m(s)=\frac{\frac{1}{2}K(s)}{\sigma_x^2}=1-R(s). \tag{5.7}$$

Тоді при $R(s)=1$ нормована структурна функція $m(s)$ буде дорівнювати нулю, й при $R(s)=0$ – одиниці.

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 А.Н. Колмогоровим для масштабування турбулентних утворень, де швидкість потоку розглядалася як розрахункова характеристика. У такому разі структурна функція являє собою дисперсію різниці швидкостей потоку u у двух точках, які знаходяться на відстані Δz одна від одної

$$M(\Delta z)=\overline{[u(z)-u(z)+\Delta z]^2}\ . \tag{5.8}$$

Якщо замінити $\Delta z=\bar{u}\,\tau$, то можна перейти до часової структурної функції виду

$$M(\tau)=\frac{1}{(n-\tau)}\sum_{i=1}^{n-\tau}\left(\Delta u_i-\Delta u_{i+\tau}\right)^2=2\left[\sigma_x^2-K(\tau)\right], \qquad (5.9)$$

де $\Delta u_i=u_i-\overline{u}$; $\Delta u_{i+1}=u_{i+1}-\overline{u}$ – центровані значення ряду швидкостей потоку, так звані пульсації.

Нормована структурна та автокореляційна часові функції пов'язані між собою

$$m(\tau)=2\left[1-r(\tau)\right]. \qquad (5.10)$$

Установлено, що структурна функція зростає до певного часу $\tau_{\max}$ (рис.5.1). При $\tau>\tau_{\max}$ структурна функція мало змінюється, досягаючи стану "насичення". Якщо структурна функція нормована, то при стані насичення вона досягає значення 2.

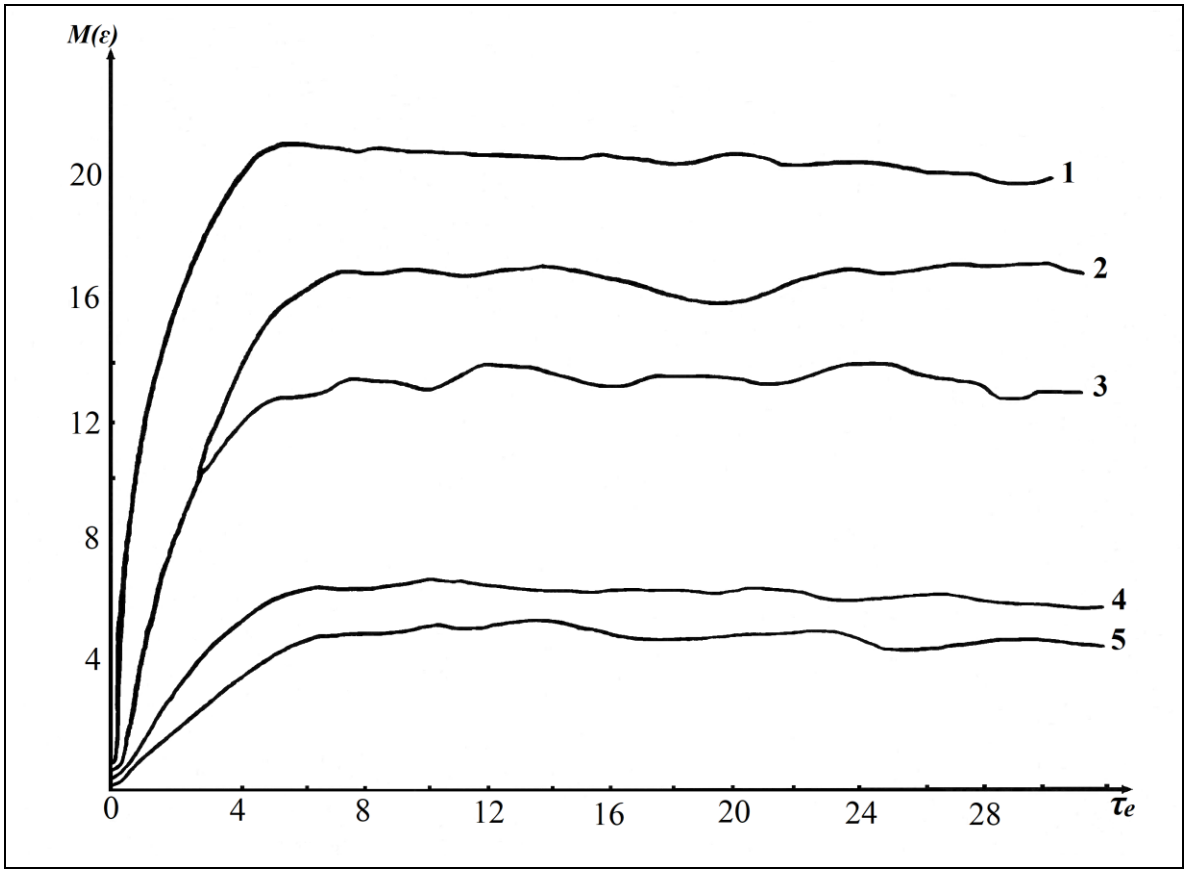

Рисунок – 5.1 Часові ненормовані структурні функції швидкостей потоку: 1 – дно; 2 – 0.8*H*; 3 – 0.6*H*; 4 – 0.2*H*; 5 – поверхня

Контрольні запитання

1. Яку властивість стаціонарного ергодичного процесу характеризує структурна функція?
2. Як структурна функція пов’язана із автокореляційною аціонарного ергодичного процесу характеризує структурна функція?
3. Як змінюється нормована структурна функція із часовим зсувом τ?
4. Якого значення досягає нормована структурна функція при своєму насиченні τ?

ЛЕКЦІЯ 6
«ФРАКТАЛЬНІСТЬ ГІДРОЕКОЛОГІЧНИХ ОБ'ЄКТІВ»

Термін "фрактал" був введений Мандельбротом, співробітником дослідницького центру імені Томаса Дж.Уотсона корпорації IBM в Йорктаун-Хейтсі (штат Нью-Йорк) і походить від латинського слова ***fractus***, що означає ламати, розбивати. Багато які структури володіють фундаментальною властивістю 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 і ті ж фундаментальні елементи (фрактали).* Важливо підкреслити, що спочатку фрактали застосовувалися як своєрідна мова геометрії. Проте, на відміну від звичних об'єктів евклидової геометрії (пряма лінія, коло і т.д.) фрактали не можуть бути безпосередньо спостереженими. Фрактали виражаються не в первинних геометричних формах, а в алгоритмах, наборах математичних процедур. Саме їх пошук й обгрунтування є центральною задачею сучасної теорії фракталів.

Багато які природні процеси, не дивлячись на те, що вони підпорядковуються певним детерміністичним законам, на достатньо великих часових інтервалах є непередбачуваними і проявляють схожі закономірності у варіаціях в різних часових масштабах подібно тому, як об'єкти, що володіють інваріантністю (самоподібністю) 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 їх структури може бути виміряна деяким характеристичним числом – фрактальною розмірністю. Таким чином, *фрактал є математичним об'єктом, який 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 всієї множини, то фрактальна розмірність дорівнює

$$d = \frac{\ln N}{\ln k}, \tag{6.1}$$

або

$$N = n^d, \qquad (6.2)$$

де d – фрактальна розмірність.

Розмірність d, визначену за (6.2) ще називають розмірністю самоподібності, де N – кількість об'єктів, подібних даному, але з розмірами в n разів меншими за даний об'єкт. Коефіцієнт подібності дорівнює $1/n$.

Проте, в більшості випадків масштабні множники (такі як d) неоднорідні, тобто у фракталів є цілий спектр скейлингів (масштабів). Такі фрактали називаються мультифракталами й характеризуються цілим спектром розмірностей.

За допомогою теорії фракталів стало можливим пояснити давно відомі закономірності будови річкової мережі. Рисунок річкової мережі має основні типи: радіальний відцентровий або радіальний доцентровий; деревоподібний; перистий; решетчатий; паралельний або субпаралельний; концентричний. Для достатньо великих річкових систем самоподібність досить легко простежується: будова кожної з приток подібна будові головної річки.

Річкову систему можна представити сукупність самоподібних індивідуальних потоків (рис.6.1с) – кривої Коша (рис.6.1а) та деревоподібної кривої (рис.6.1б), або як ієрархічну (порядкову) модель, у якій річкова система розглядається як сукупність відрізків (притоків) різної довжини без урахування фрактальності індивідуальних водотоків (рис.6.2). Р.Є. Хортон вніс пропозицію розглядати річкову мережу як відкриту деревоподібну систему, що складається з ієрархії різних підсистем. Самою дрібною одиницею вважається елементарний нерозгалужений водотік, якому надається перший порядок. Річки другого рівня ієрархії приймають до себе тільки притоки першого порядку, річки третього порядку утворюються при злитті річок другого порядку й т.д.

Розглядаючи річкову мережу як відкриту деревоподібну систему (рис.6.2), що складається з приток різних порядків, і, використовуючи як топологічний параметр порядок водотоку n (порядок може визначатися різними способами), Р.Є. Хортон (1943) сформулював ряд закономірностей, які стали основою сучасної гідрографічної науки.

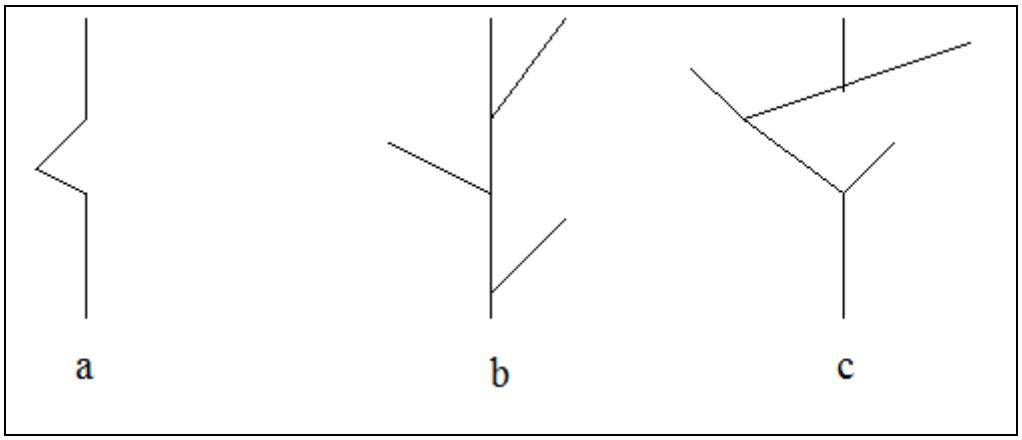

Рисунок 6.1 – Конструювання самоподібної фрактальної річкової мережі із самоподібними індивідуальними потоками

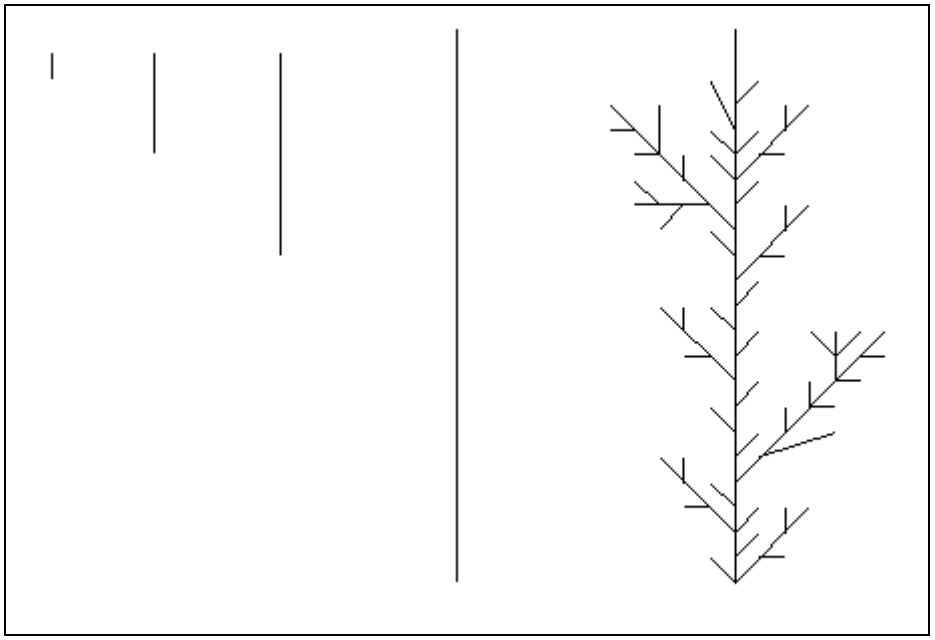

Рисунок 6.2 – Приклад порядкової фрактальної моделі річкової мережі

Перша із закономірностей носить назву закономірності кількості приток: в кожній річковій системі співвідношення між кількістю приток суміжних порядків є величина постійна, тобто:

$$\sigma_0 = \frac{N_{n-1}}{N_n}\ , \tag{6.3}$$

де σ_0 – називається коефіцієнтом біфуркації;

N_{n-1}, N_{n-2} – кількість приток суміжних порядків.

Коефіцієнт біфуркації є, одночасно, константою R_B.

Друга закономірність визначає таке положення: співвідношення між довжинами приток річок суміжних порядків залишається в середньому постійним:

$$\lambda_0 = \frac{L_n}{L_{n-1}}. \tag{6.4}$$

Третя закономірність полягає у тому, що площі водозборів приток суміжних порядків також знаходяться в певному співвідношенні:

$$\varphi_0 = \frac{F_n}{F_{n-1}}, \tag{6.5}$$

де φ_0 – коефіцієнт площі;

F_n та F_{n-1} – площі водозборів приток суміжних порядків n та n–1.

Четверта закономірність Хортона встановлює співвідношення між середніми уклонами річок суміжних порядків:

$$I_0 = \frac{I_{n-1}}{I_n}, \tag{6.6}$$

де I_0 – коефіцієнт уклонів приток I_{n-1} і I_n суміжних порядків n–1 та n.

Пізніше закономірності, установлені Р.Є. Хортоном, були пояснені фрактальною поведінкою річкової мережі. Вперше фрактальна поведінка річкової мережі була виявлена Хаком (1957) у вигляді степеневої залежності довжини річки від площі та описана Мандельбротом (1977)

$$L_r^{1/d} \sim F^{0.5}\ , \tag{6.7}$$

де L_r – довжина річки уздовж продольної осі.

Установлені закономірності Р.Є. Хортона увійшли до формул визначення фрактальної розмірності річкової мережі, де використовуються константи, які одночасно є коефіцієнтами Р.Є. Хортона

$$R_B = \frac{N_{n-1}}{N_n} = \sigma_0; \tag{6.8}$$

$$R_L = \frac{L_n}{L_{n-1}} = \lambda_0; \tag{6.9}$$

$$R_A = \frac{F_n}{F_{n-1}} = \varphi_0; \tag{6.10}$$

$$R_I = \frac{I_{n-1}}{I_n} = I_0, \tag{6.11}$$

де N_n, L_n, I_n, F_n – кількість приток n-го порядку та їх середня довжина, уклон та площа;

$N_{n-1}, L_{n-1}, I_{n-1}, F_{n-1}\, L_{n-1}, N_{n-1}$ – кількість приток $n-1$-го порядку та їх середня довжина, уклон та площа.

Константи R_B, R_L, R_A, R_I можуть визначатися за даними про притоки різних суміжних порядків. Фрактальна розмірність розглянутої річкової мережі визначається за формулою як тангенс кута нахилу залежності $\ln R_B = f(\ln R_L)$ або $\ln R_B = f(\ln R_A)$

$$d = \frac{\ln R_B}{\ln R_L}; \tag{6.12}$$

або

$$d = \frac{\ln R_B}{\ln R_L} = 2\frac{\ln R_B}{\ln R_A}. \tag{6.13}$$

Властивість самоподібності може проявлятися і у структурі математичних об’єктів. У загальному випадку установлення властивостей

самоподібності передбачає наявність степеневої залежності між статистичним моментом $F_q(s)$ порядку q та масштабом s

$$F_q(s) = s^{h(q)}\ , \tag{6.14}$$

де $F_q(s)$ – флуктуаційна функція;

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

Якщо $h(q)$ не залежить від q (для усіх q значення $h(q)$ постійні), то ця обставина розглядається як властивість монофрактальності. Для монофрактальних об’єктів $h(q) = H$, показник Хурста.

Зазвичай вивчається флуктуаційна функція другого порядку (так званий другий статистичний момент) виду

$$F_2(s) = s^H\ . \tag{6.15}$$

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.

Другий статистичний момент $F_2(s)$ пов'язаний із нормованою структурною функцією наступним рівнянням

$$\sqrt{M^0(s)} = F_2(s) \sim s^H\ , \tag{6.16}$$

або

$$M^0(s) \sim s^{2H}\ , \tag{6.17}$$

де H розглядається як фрактальна розмірність.

На практиці для розрахунків нормованої структурної функції залучається автокореляційна функція за рівнянням

$$M^0(\tau) = 2[1 - r(\tau)], \qquad (6.16)$$

$M^0(\tau)$ - нормована структурна функція (рис.6.3).

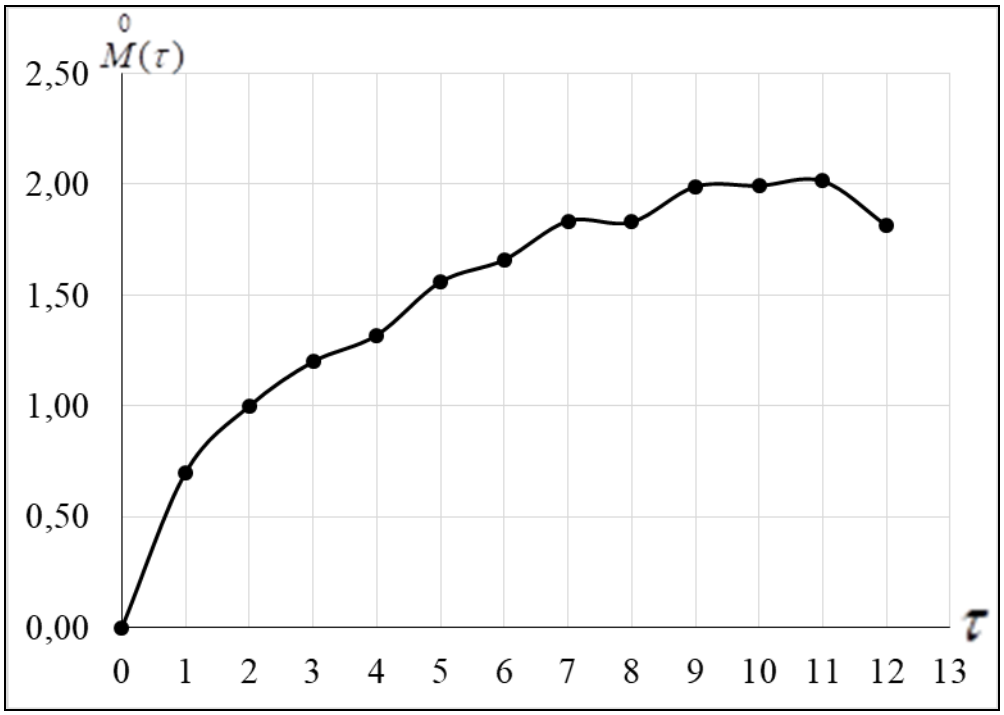

Рисунок 6.3 – Нормована структурна функція

Після логарифмування флуктуаційної функції другого порядку $\ln F_2(\tau) = \ln M^0(\tau)$ отримуємо графік (рис.6.4), де тангенс кута нахилу дорівнює $2H$. З наведеного прикладу виходить, що $2H = 0{,}232$, звідки $H = 0{,}116$

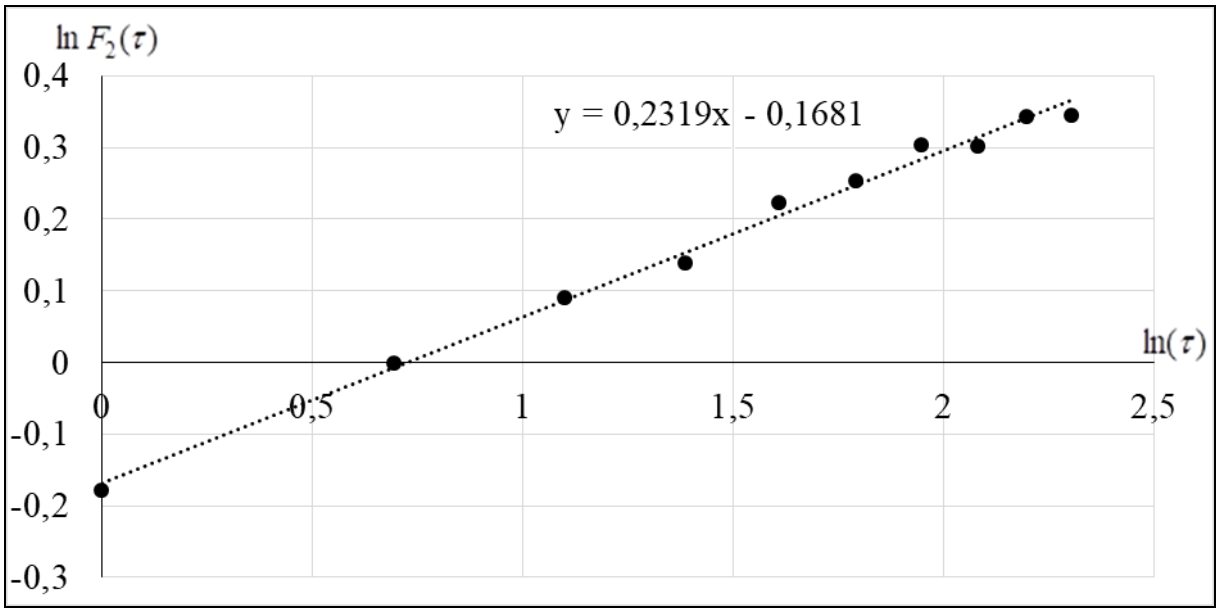

Рисунок 6.4 – Графік зв’язку логарифмованої флуктуаційної функції другого порядку $\ln F_2(\tau)$

Контрольні запитання

1. Дати визначення фрактальності (самоподібності).
2. Дати визначення фрактальності.
3. Флуктуаційна функція та її зв'язок із структурною функцією.
4. Коефіцієнти Хортона як характеристики подібності річкової мережі.
5. Як визначити фрактальну розмірність через структурну функцію?

ЛЕКЦІЯ 7
«ВЗАЄМНА КОРЕЛЯЦІЙНА ФУНКЦІЯ ТА ЇЇ ЗАСТОСУВАННЯ ДО РОЗРАХУНКІВ САМООЧИЩЕННЯ ВОДИ НА ДІЛЯНЦІ РУСЛА»

Сукупність усіх процесів, спрямованих на відновлення початкового хімічного складу води відповідно до існуючої раніше рівноваги, називається самоочищенням водного об'єкта. Більшість забруднювальних речовин є нестійкими і з часом виводяться з розчину під впливом різних процесів, які сприяють самоочищенню. Самоочищення і встановлення біологічної рівноваги відбувається в результаті сукупної дії гідравлічних, фізичних, хімічних і біологічних чинників. У річках процеси самоочищення обумовлені фізичними та хімічними чинниками. У озерах та ставках самоочищення відбувається переважно за допомогою живих організмів. Під фізичними чинниками самоочищення слід розуміти гідравлічні процеси: осідання нерозчинних осадів, вплив ультрафіолетового випромінювання та інше. Серед хімічних чинників зазначають біохімічне окиснення та окиснення неорганічних речовин. Сильно забруднена річка може шляхом самоочищення перейти із стану сапробної зони у мезосапробну і навіть у олігосапробну при відсутності постійного поповнення забруднювальними речовинами (*Осадчий В.І., 2013*).

В процесі самоочищення відбувається поліпшення фізичних властивостей води за рахунок адсорбції завислими частками органічних речовин, важких металів, мікроорганізмів, коагуляції та седиментації завислих неорганічних і органічних речовин, мінералізації нестійкої органічної речовини. При самоочищенні вміст кисню має зростати завдяки аерації та дії водної рослинності. Патогенні бактерії будуть при цьому відмирати, наявність сапрофітних мікроорганізмів різко зменшиться. Завдяки самоочищенню невелике забруднення не може змінити природного стану водойми. Але кожна водойма має певну межу самоочисної здатності від забруднень, після якої відбувається різке погіршення всіх характеристик санітарного стану. Процеси самоочищення протікають більш сприятливо завдяки більшій проточності у річках, ніж в озерах і водосховищах. При вивченні процесів самоочищення важливе значення мають співвідношення кількості забруднювальних речовин і об’єму водної маси, швидкість течії, турбулентності водних мас, глибини, умови вітрового перемішування, температурний режим тощо.

Характеристиками самоочищення є кількісні показники, які визначаються для хімічних та біологічних процесів. При забрудненні річок господарсько-побутовими водами, провідним процесом у самоочищенні є розкладання органічної речовини, кінцевим продуктом якого є мінеральні сполуки. В анаеробних умовах розкладання викликає посилене споживання кисню і корелює з показниками біологічного споживання кисню. Ці показники характеризують ступінь розкладання нестійкої органічної речовини. Константа швидкості розкладання К залежить від складу забруднювальних речовин і має різні значення. Для господарсько-побутових вод вона становить близько 0,1 доб$^{-1}$ і має приблизно таке саме значення при розкладанні фітопланктону. Промислові забруднювальні речовини обумовлюють більші коливання характеристики К. Швидкості розпаду органічних речовин донних відкладів у 20-50 разів нижчі, ніж господарсько-побутових стічних вод. Для характеристики самоочищення велике значення має швидкість зміни кількості бактерій. Протягом перших 15 годин відмирає 70% від початкової величини бактеріального зараження, а на п’яту добу їх залишається лише частки відсотка.

Процеси самоочищення вивчають на спеціальних пунктах контролю. Ці дослідження виконуються в зоні забруднення річки чи водойми, де у зв’язку з надходженням забруднювальних речовин порушуються природні біохімічні процеси і концентрація забруднювальних речовин за санітарними чи іншими показниками перевищує встановлені норми. Дослідження проводяться у трьох створах (фоновому, головному і замикальному). Фоновий створ повинен бути розташований вище створу скиду забруднених вод. Концентрації забруднювальних речовин у цьому створі не повинні перевищувати гранично допустимі. Додаткові створи встановлюються між головним (контрольним) і замикальним створами. Кількість таких створів залежить від завдань спостережень, місцевих умов та складу забруднювальних речовин. Вони розміщуються нижче випуску забруднених вод з послідовним збільшенням відстані між ними (шість – дев’ять створів). Додаткові створи можуть призначатися за наявності на досліджуваній ділянці бокового припливу. Такі створи встановлюються вище і нижче притоки та в її гирлі. Визначення самоочисної здатності водних об’єктів виконується для специфічних забруднювальних речовин, наявність яких встановлено у воді під час експедиційних досліджень, а також за такими показниками забруднення води як хімічне споживання кисню, біохімічне споживання кисню.

Визначення температури, pH, вмісту розчиненого кисню є обов'язковим. Ці показники характеризують умови і хід процесів самоочищення. Спостереження за забрудненням води виконуються кілька разів на рік у характерні фази гідрологічного і гідробіологічного режимів. Тривалість спостережень визначається необхідністю отримання надійних матеріалів щодо характеристики самоочисної здатності водотоків у роки з різним ступенем водності (багатоводні, маловодні, середні) (*Хільчевський В.К., Осадчий В.І., Курило С.М., 2012*).

Самоочисна здатність води на ділянці обчислюється за таким рівнянням

$$СЗ = \frac{C_B - C_H}{C_B} 100\% \quad , \tag{7.1}$$

де C_B – концентрація забруднювальної речовини у верхньому створі, мг/дм^3;

C_H – концентрація забруднювальної речовини у нижньому створі, мг/дм^3.

Ступінь самоочищення визначають за зниженням концентрації забруднювальної речовини на певному відрізку водотоку при відсутності додаткового забруднення між точками спостережень.

$$Sm = Q(C_B - C_H), \tag{7.2}$$

де $C_в$ – концентрація забруднювальної речовини у верхньому створі, мг/дм^3;

$C_н$ – концентрація забруднювальної речовини у нижньому створі, мг/дм^3;

Q – витрата води, м^3/с.

Швидкість самоочищення визначається за зниженням концентрації забруднювальної речовини за одиницю часу

$$Sr = \frac{dC}{dt} = \frac{C_B - C_H}{\tau}. \tag{7.3}$$

де $C_в$ – концентрація забруднювальної речовини у верхньому створі, моль/дм^3;

$C_н$ – концентрація забрудної речовини у нижньому створі, моль/дм^3;

τ – час добігання між верхнім та нижнім створами.

Під молем розуміють кількість речовини, маса якої, виражена в грамах, чисельно дорівнює її молекулярній масі, вираженій в атомних одиницях маси (а.о.м), що визначається як 1/12 частина маси атома ізотопу ^{12}C, яка дорівнює 1,66056*10^{-24} г. Моль поширений на будь-які види реальних і умовних частинок. Під реальними частинками розуміють молекули, атоми , іони , електрони, радикали, під умовними – еквіваленти. Відповідно можна говорити про моль молекул, моль іонів, моль електронів.

Щоб розрахувати молярну концентрацію хімічної речовини в розчині необхідно масу речовини в грамах поділити на її атомну масу.

Сумарний коефіцієнт самоочищення визначається за формулою В.Г. Стрітера

$$K_C = \frac{2.3}{\tau} \lg \frac{C_B}{C_H} = \frac{1}{\tau} \ln \frac{C_B}{C_H} \quad , \qquad (7.4)$$

де $C_в$ – концентрація забруднювальної речовини у верхньому створі, мг/дм^3;

$C_н$ – концентрація забруднювальної речовини у нижньому створі, мг/дм^3;

τ – час добігання, доба.

Значення K_c отримують в установах Державної гідрометеорологічної служби, фонові концентрації розраховують за рівнянням типу $C = f(Q)$, а сумарні коефіцієнти самоочищення визначають по завчасно встановленим зв'язкам типу $K = f(Q)$ або $K = f(t)$, або $\kappa = f(Q,t)$. За відсутності таких зв'язків розраховують середньоарифметичні значення відповідних коефіцієнтів за багаторічними даними за той місяць, для якого складається прогноз.

Якщо відомі коефіцієнт самоочищення та час добігання від верхнього створу до нижнього, можна установити концентрацію забруднювальної речовини у нижньому створі, із завчасністю яка дорівнює τ

$$C_H = C_B 10^{-\frac{\tau K_C}{2.3}} = C_B e^{-\tau K_C}, \qquad (7.5)$$

де $C_в$ – концентрація забруднювальної речовини у верхньому створі;

$C_н$ – концентрація забруднювальної речовини у нижньому створі;

τ – час добігання;

K_C – сумарний коефіцієнт самоочищення (за В.Г.Стріттером).

Визначення часу добігання між верхнім та нижнім створами можна здійснити на основі розрахунків взаємної кореляційної функції:

$$r_{xy}(\tau) = \frac{K_{xy}(\tau)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(y_{i+\tau} - \bar{y})}{\sigma_x \sigma_y (n - \tau_{зсув} - 1)}, \qquad (7.6)$$

де $r_{xy}(\tau)$ – взаємна кореляційна функція при зсуві у часі $\tau_{зсув}$;

$\bar{x}, \bar{y}$ – середні арифметичні значення рядів X, Y;

σ_x, σ_y – середні квадратичні відхилення двох рядів спостережень довжиною n.

При цьому ряд Q_H зсувається у часі відносно ряду Q_B, тобто формулу (7.6) можна представити у вигляді:

$$r_{xy}(\tau) = \frac{K_{Q_B Q_H}(\tau)}{\sigma_{Q_B} \sigma_{Q_H}} = \frac{\sum_{i=1}^{n-\tau} (Q_{Bi} - \overline{Q_B})(Q_{Hi+\tau} - \overline{Q}_H)}{\sigma_{Q_B} \sigma_{Q_H} (n - \tau_{зсув} - 1)}, \qquad (7.7)$$

де Q_{Bi} – витрати води у верхньому створі в час i;

Q_{Hi} – витрати води у нижньому створі в часі $i+\tau$;

$\sigma_{Q_B}, \sigma_{Q_H}$ – середні квадратичні відхилення для витрат води у верхньому та нижньому створах, відповідно.

Час добігання мас води від верхнього створу до нижнього визначався як такий, при якому $r_{xy}(\tau)$ досягала свого максимального значення (рис.7.1, рис. 7.2).

Під рядами випадкових величин X та Y слід розуміти ряд стоку (витрат води) у верхньому створі Q_B та нижньому – Q_H.

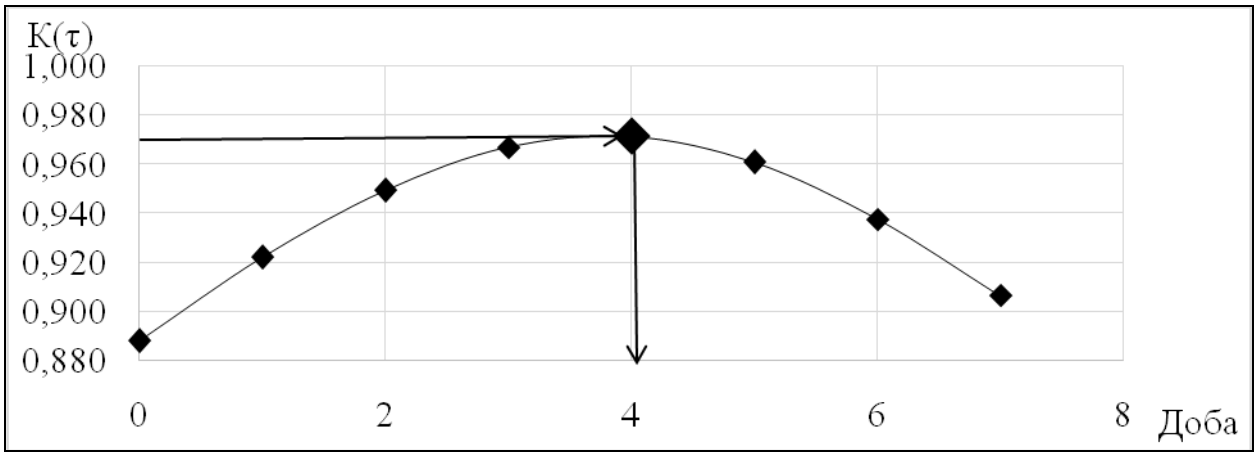

Рисунок 7.1 – Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Псел – м. Суми відносно ряду стоку р. Псел – с. Крупець (*Loboda N.S., Pilipyuk V.V., 2014*)

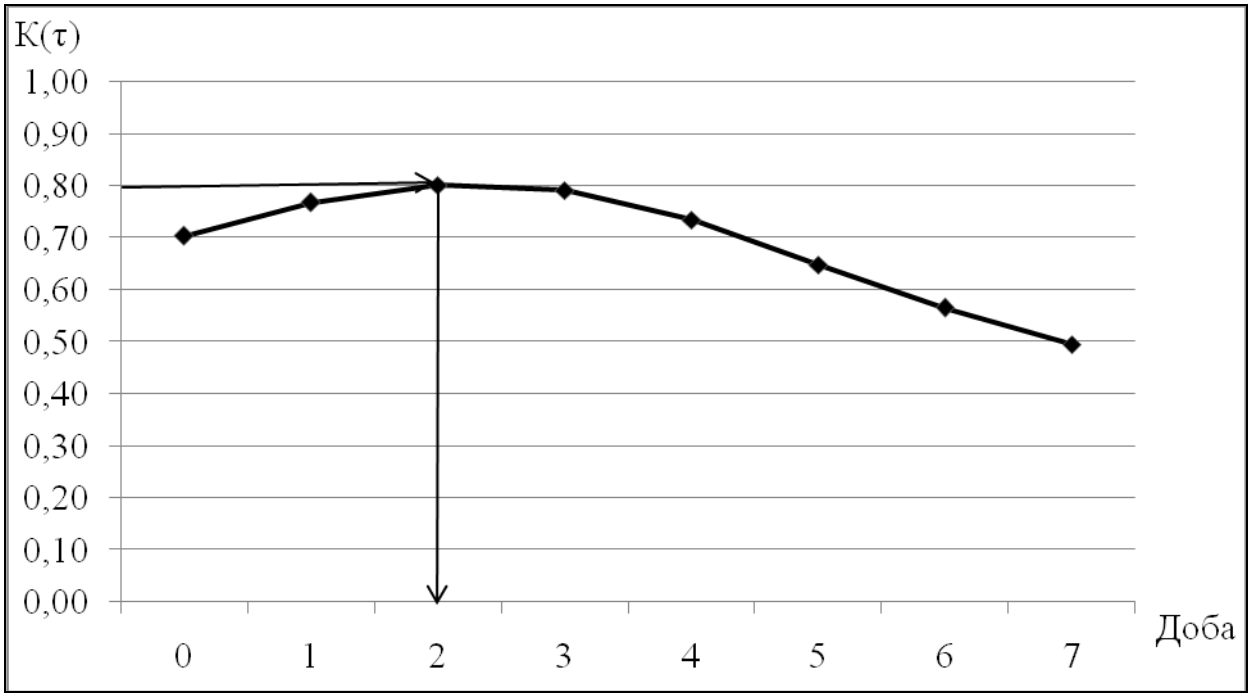

Рисунок 7.2 – Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Ворскла – с. Чернеччина відносно ряду стоку р. Ворскла – с. Козинка (*Loboda N.S., Pilipyuk V.V.*,2015)

Контрольні запитання

1. Що розуміють під самоочищенням водних об’єктів?
2. Як визначається швидкість самоочищення?
3. Який зв'язок характеризує взаємна кореляційна функція?
4. За якою ознакою визначається час добігання від верхнього створу до нижнього?
5. Як зростання часу добігання з верхнього створу до нижнього впливає на самоочищення води на ділянці річки?

ЛЕКЦІЯ 8
«МАТРИЦІ ТА ДІЇ НАД НИМИ. МАТРИЦІ КОВАРІАЦІЙ І КОРЕЛЯЦІЙ»

Матрицею називається система елементів, розташованих у визначеному порядку і утворюючих таблицю. Якщо в цій таблиці m рядків та n стовбців, а її елементи позначаються як a_{ij}, де i – номер рядка, а j – номер стовпця, на перетині яких знаходиться елемент a_{ij}, то матриця записується у наступному вигляді

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \tag{8.1}$$

Ця матриця може бути представленою також таким чином $A=\lfloor a_{ij} \rfloor$, $(i=1,2,3,\dots,m;\ j=1,2,3,\dots,n)$. Наведена матриця має розмір $m\times n$.

Якщо кількість рядків m не дорівнює кількості стовбців n, то матриця зветься ***прямокутною***.

Якщо в матриці тільки один стовпець, тобто $n=1$, то така матриця називається матрицею-стовпцем і має наступний вигляд

$$A=\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \dots \\ a_{m1} \end{bmatrix} \tag{8.2}$$

Матриця – стовпець також називається ***вектором-стовпцем.*** Якщо в матриці тільки один рядок, тобто $m=1$, то вона називається матрицею-рядком або ***вектором-рядком***.

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \end{bmatrix} \tag{8.3}$$

Якщо в матриці поміняти місцями рядки та стовбці, то отримаємо нову матрицю, яка у порівнянні з матрицею A називається транспонованою й позначається як A'.

Наприклад, якщо

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}, \tag{8.4}$$

то транспонована матриця A' буде записуватися у виді

$$A'=\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \\ a_{14} & a_{24} & a_{34} \end{bmatrix} \tag{8.5}$$

Матриця A має розмір 3×4, а транспонована матриця 4×3. Коли в матриці стільки ж рядків, скільки і стовпців, то така матриця зветься ***квадратною***. Наприклад,

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \tag{8.6}$$

Наведена квадратна матриця має розмір $n\times n$.

Квадратна матриця, в якій всі елементи, що розташовані поза головною діагоналлю, дорівнюють нулю, зветься ***діагональною***. Наприклад,

$$A=\begin{bmatrix} a_{11} & 0 & 0 & ... & 0 \\ 0 & a_{22} & 0 & ... & 0 \\ 0 & 0 & a_{33} & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & a_{nn} \end{bmatrix} \quad (8.7)$$

В діагональній матриці всі елементи з індексами, які не дорівнюють одне одному, дорівнюють нулю.

Діагональна матриця, всі елементи якої, що розташовані на головній діагоналі, дорівнюють одне одному, називається ***скалярною***. Наприклад,

$$B=\begin{bmatrix} a_{11} & 0 & 0 & ... & 0 \\ 0 & a_{11} & 0 & ... & 0 \\ 0 & 0 & a_{11} & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & a_{11} \end{bmatrix} \quad (8.8)$$

Одинична матриця – це діагональна матриця, у якій всі елементи головної діагоналі дорівнюють одиниці. Вона позначається символом E и має такий вигляд:

$$E_n=\begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{bmatrix} \quad (8.9)$$

Індекс n у символі E показує порядок одиночної матриці.

Верхня трикутна матриця – це квадратна матриця, в якій елементи, розташовані нижче головної діагоналі, дорівнюють нулю. Наприклад,

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix} \qquad (8.10)$$

Нижня трикутна матриця – це квадратна матриця, в якій всі елементи, розташовані вище головної діагоналі, дорівнюють нулю. Наприклад,

$$A=\begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix} \qquad (8.11)$$

Симетричною матрицею називається квадратна матриця, в якій всі елементи, розташовані симетрично до головної діагоналі, дорівнюють одне одному. Для симетричної матриці має місце рівність $a_{ij}=a_{ji}$. Наприклад,

$$\begin{bmatrix} 1 & 4 & 2 \\ 4 & 3 & 6 \\ 2 & 6 & 5 \end{bmatrix}.$$

Визначником квадратної матриці називається визначник, складений з елементів матриці, розташованих таким самим чином, як і в матриці. Визначник матриці A позначається символом $|A|$.

Якщо визначник матриці дорівнює нулю, то вона називається ***виродженою***, або ***сингулярною***.

У прямокутній матриці можна викреслити декілька рядків та декілька стовбців таким чином, щоб елементи, які залишились невикресленими, утворили квадратну матрицю порядку k. Визначник такої матриці називається ***мінором*** вихідної матриці. Якщо взяли матрицю розміром 3×4, то викреслюючи один будь-який стовпець можна отримати чотири квадратні матриці 3-го порядку. Викреслюванням двох стовпців та одного рядка можна отримати 18 матриць другого порядку.

Кількість матриць порядку k ($k \le min(s,r)$), яке можна отримати з матриці розміром $s \times r$, підраховується за формулою

$$n = C_s^k C_r^k, \tag{8.12}$$

де C_s^k та C_r^k є кількість комбінацій з s елементів по k у кожній та з r елементів по k у кожній. Кількість комбінацій з n елементів по m підраховується за формулою

$$C_n^m = \frac{n!}{m!(n-m)!} \quad (n \ge m) \tag{8.13}$$

Найбільший порядок r мінору матриці, визначник якої не дорівнює нулю, називається ***рангом*** матриці.

Ранг діагональної квадратної матриці дорівнює її порядку, зрозуміло, за умови, що жоден з її діагональних елементів не дорівнює нулю.

Дві матриці A і B називаються ***рівними***, якщо:

- обидві мають один і той же розмір, тобто кількість рядків матриці A дорівнює числу рядків матриці B, а кількість стовбців матриці A дорівнює кількості стовбців матриці B;
- відповідні елементи цих матриць рівні між собою. (Під відповідними елементами слід розуміти елементи з одними й тими ж індексами).

Таким чином, якщо матриця A має вигляд

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix}, \tag{8.14}$$

а матриця B-

$$B = \begin{bmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & b_{m3} & \dots & b_{mn} \end{bmatrix} \tag{8.15}$$

й $a_{ij} = b_{ij} \quad (i = 1,2,...,m;\ \ j = 1,2,...,n)$, то $A = B$.

Припустимо, що n величин $x_1, x_2, x_3, ..., x_n$ пов'язані з m величинами $y_1, y_2, y_3, ..., y_m$ такими залежностями

$$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n \\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + ... + a_{2n}x_n \\ &..... \\ y_m &= a_{m1}x_1 + a_{m2}x_2 + a_{m3}x_3 + ... + a_{mn}x_n \end{aligned}\right\} \qquad (8.16)$$

За цими формулами величини $y_1, y_2, y_3, ..., y_m$ виражаються лінійно й однорідно через n інших величин $x_1, x_2, x_3, ..., x_n$.

Відбувається перетворення величин $x_1, x_2, x_3, ..., x_n$ у величини $y_1, y_2, y_3, ..., y_m$. Таке перетворення має назву ***лінійного*** й скорочено записується таким чином

$$y_i = \sum_{j=1}^{n} a_{ij}x_j, \quad (i = 1,2,3,..., m)\ . \qquad (8.17)$$

Коефіцієнти лінійного перетворення можуть бути записані у вигляді матриці

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ ... & ... & ... & ... & ... \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mn} \end{bmatrix}, \qquad (8.18)$$

яка називається ***матрицею лінійного перетворення*** величин $x_j \quad (j = 1,2,...,n)$ у величини $y_i \quad (i = 1,2,...,m)$.

Сумою двох прямокутних матриць A та B рівних розмірів $m \times n$ називається матриця C того ж розміру, елементи якої C_{ij} дорівнюють сумі відповідних елементів матриць A та B

$$A + B = C\ , \qquad (8.19)$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \dots & \dots & \dots & \dots \\ c_{m1} & c_{m2} & \dots & c_{mn} \end{bmatrix}, \tag{8.20}$$

де $c_{ij} = a_{ij} + b_{ij} \quad (i = 1,2,\dots,m;\ j = 1,2,\dots,n)$.

Різниця матриць A та B, які мають однаковий порядок, є матриця того ж порядку D, елементи якої дорівнюють різниці відповідних елементів матриць A й B, тобто

$$d_{ij} = a_{ij} - b_{ij} \tag{8.21}$$

Сума матриць підкорюється наступним законам (сполучному або асоціативному та переставному або комутативному):

$$\begin{aligned} A + B &= B + A \\ (A + B) + C &= A + (B + C) \end{aligned}. \tag{8.22}$$

Щоб помножити матрицю A на число α, потрібно помножити на це число кожен елемент матриці A. Нова матриця записується таким чином

$$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \alpha a_{13} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{23} & \dots & \alpha a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \alpha a_{m1} & \alpha a_{m2} & \alpha a_{m3} & \dots & \alpha a_{mn} \end{bmatrix} \tag{8.23}$$

Добуток BA двох матриць можливо отримати у тому й тільки у тому випадку, коли кількість стовбців матриці B дорівнює кількості рядків матриці A. Коли розмір матриці B дорівнює $s \times m$, а розмір матриці A дорівнює $m \times n$, то матриця BA має розмір $s \times n$.

У символічному виді це записується так

$$(s \times m)(m \times n) = s \times n \tag{8.24}$$

Якщо $C = BA$, то кожний елемент c_{ij} матриці C дорівнює сумі добутків елементів i-того рядка матриці B на відповідні елементи j-того стовпця матриці A. Наприклад, елемент C_{14} представляє собою суму добутків першого рядка матриці B на четвертий стовбець матриці A, тобто

$$c_{14} = b_{11}a_{14} + b_{12}a_{24} + b_{13}a_{34} + ... + b_{1m}a_{m4} \tag{8.25}$$

У загальному випадку

$$c_{ij} = b_{i1}a_{1j} + b_{i2}a_{2j} + b_{i3}a_{3j} + ... + b_{im}a_{mj} = \sum_{k=1}^{m} b_{ik} \cdot a_{kj} \tag{8.26}$$

Добуток двох матриць не має властивості комутативності

$$BA \neq AB \tag{8.27}$$

та асоціативності

$$A(BC) \neq (AB)C \ . \tag{8.28}$$

Розв'язок чисельних задач сучасної гідрометеорології потребує знань про статистичну структуру метеорологічних полів, таких як, наприклад, поля опадів, температур, стоку, тиску та вологості повітря, швидкості вітру, тощо. Сукупність m гідрометеорологічних полів, які відносяться до визначених термінів спостереження, можна, як відзначалося вище, зобразити матрицею порядку $n \times m$ вигляду

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{ni} & \cdots & x_{nm} \end{bmatrix} \tag{8.29}$$

Таким же чином можна скласти інформацію, наприклад, про m вертикальних профілів метеорологічної величини, яка вимірюється на n рівнях атмосфери або про множину m векторів-предікторів у випадку, коли вирішується проблема побудови статистичної моделі гідрометеорологічного прогнозу. Можна навести приклади й інших задач. Щоб не перелічувати кожний раз ці задачі, будемо у подальшому інколи іменувати поля, вертикальні профілі гідрологічних величин, вектори-предіктори й т.п. гідрометеорологічними об'єктами.

Матриця (8.29) утримує великий об'єм інформації. У матриці (8.29) концентрується інформація про n об'єктів. Рядки матриці являють собою, як вже зазначалося, часові ряди відповідної гідрометеорологічної величини. Таке матричне зображення гідрометеорологічних об'єктів є дуже раціональним, оскільки дає можливість побудувати прості алгоритми дослідження їх статистичної структури. Найбільш важлива інформація про статистичну структуру гідрометеорологічних об'єктів міститься у матриці коваріацій. Алгоритм побудови цієї матриці складається з декількох етапів. Знайдемо, насамперед, вектор середніх значень

$$\overline{X} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_i \\ \dots \\ \overline{x}_n \end{bmatrix}, \tag{8.30}$$

де

$$\overline{x}_i = \frac{1}{m} \sum_{j=1}^{m} x_{ij} \quad (i = \overline{1,m}) \,. \tag{8.31}$$

Підкреслимо ще раз, що коли йдеться про поля гідрометеорологічних величин, то вектор (8.30) є осереднене поле гідрометеорологічної величини.

На основі матриці (8.29) та вектора (8.30) визначимо відповідну матрицю центрованих елементів. Для цього від елементів кожного рядка

матриці (8.29) віднімемо відповідне середнє арифметичне значення. У такому випадку отримаємо:

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{i1} & \Delta x_{i2} & \dots & \Delta x_{ij} & \dots & \Delta x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{n1} & \Delta x_{n2} & \dots & \Delta x_{ni} & \dots & \Delta x_{nm} \end{bmatrix}, \tag{8.32}$$

де

$$\Delta x_{ij} = x_{ij} - \bar{x}_i \tag{8.33}$$

Операція, що проведена над матрицею (8.29), називається, як зазначалось вище, операцією центрування. Матриця коваріацій K_x визначається за таким матричним рівнянням:

$$K_x = \frac{1}{m} \Delta X \Delta X', \tag{8.34}$$

де $\Delta X'$ – транспонована матриця центрованих величин.

Покажемо справедливість цього твердження. За цією метою розглянемо більш просту матрицю

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \end{bmatrix} \tag{8.35}$$

Для такої матриці рівність (8.34) в координатній формі прийме вигляд:

$$K_x = \frac{1}{m}\begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \end{bmatrix} \times \begin{bmatrix} \Delta x_{11} & \Delta x_{21} \\ \Delta x_{12} & \Delta x_{22} \\ \dots & \dots \\ \Delta x_{1j} & \Delta x_{2j} \\ \dots & \dots \\ \Delta x_{1m} & \Delta x_{2m} \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{1}{m}\sum_{j=1}^{m} \Delta x_{1j}^2 & \frac{1}{m}\sum_{j=1}^{m} \Delta x_{1j}\Delta x_{2j} \\ \frac{1}{m}\sum_{j=1}^{m} \Delta x_{2j}\Delta x_{1j} & \frac{1}{m}\sum_{j=1}^{m} \Delta x_{2j}^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & K_{12} \\ K_{21} & \sigma_2^2 \end{bmatrix}. \qquad (8.36)$$

Узагальнюючи проведені обчислювання, отримаємо:

$$K_x = \begin{bmatrix} \sigma_1^2 & K_{12} & \dots & K_{1j} & \dots & K_{1n} \\ K_{21} & \sigma_2^2 & \dots & K_{2j} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{i1} & K_{i2} & \dots & K_{ij} & \dots & K_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nj} & \dots & \sigma_n^2 \end{bmatrix} \qquad (8.37)$$

Елементи матриці (8.37) розраховуються за формулами:

$$\sigma_i^2 = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}^2 \quad , \qquad (8.38)$$

$$K_{ij} = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}\Delta x_{js} \qquad (8.39)$$

Отже, як випливає з формул (8.38) і (8.39), на головній діагоналі матриці (8.37) розташовуються дисперсії досліджуваної гідрометеорологічної величини σ_j. Порядковий номер дисперсії на

діагоналі відповідає номеру гідрометеорологічної станції, якщо йдеться про гідрометеорологічні поля, або номеру предиктора, якщо досліджуються статистичні особливості предикторів при побудові моделі прогнозу. Інші елементи матриці (8.37) – відповідні коваріації.

Матриця коваріації має такі властивості:

- її елементи є дійсними числами;
- вона є симетричною;
- матриця коваріацій є додатно визначеною.

З останньої властивості випливає, що $|K_x| > 0$. (Прямими дужками позначається визначник). До речі, матриця коваріацій, поряд з вектором математичних сподівань m_x, відіграє роль параметра щільності імовірностей багатовимірного нормального розподілу

$$f(X) = \frac{1}{(2\pi)^{n/2}|K_x|^{1/2}} \times exp\left\{-\frac{1}{2}(X - m_x)K_x^{-1}(X - m_x)\right\} \qquad (8.40)$$

Маючи матрицю коваріацій, можна легко сформувати діагональну матрицю σ середніх квадратичних відхилень. Вона має вид:

$$\sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix} \qquad (8.41)$$

Обернена матриця від діагональної матриці визначається таким чином

$$\sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{bmatrix} \qquad (8.42)$$

Якщо помножити ліворуч та праворуч матрицю коваріацій K_x на матрицю (8.42) отримаємо матрицю кореляцій R_x

$$R_x = \sigma^{-1} K_x \sigma^{-1} \tag{8.43}$$

Матриця кореляцій порядку m має наступний вигляд

$$R_x = \begin{bmatrix} 1 & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2j} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{i1} & r_{12} & \dots & r_{ij} & \dots & r_{1n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nj} & \dots & 1 \end{bmatrix}, \tag{8.44}$$

де r_{ij} – коефіцієнт кореляції, який характеризує тісноту лінійного зв'язку між двома змінними (i та j).

Коефіцієнт кореляції пов'язаний із коваріацією K_{ij} наступним співвідношенням

$$r_{ij} = \frac{K_{ij}}{\sigma_i \sigma_j} \tag{8.45}$$

При наявності рядів спостережень довжиною m, вибіркове значення коефіцієнту кореляції між двома змінними x_j та x_k розраховується таким чином

$$r_{jk} = \frac{\sum_{i=1}^{m} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{(m-1)\sigma_j \sigma_k}, \tag{8.46}$$

де x_{ij} – значення j-того ряду спостережень довжиною m;

x_{ik} – значення k-того ряду спостережень довжиною m;

$\bar{x}_j, \bar{x}_k$ – середні арифметичні значення для ряду j та ряду k;

σ_j, σ_k – середні квадратичні відхилення для ряду j та ряду k;

m – кількість спостережень.

При $k = j$ $r_{jj} = 1$, оскільки коваріація у такому випаду є дисперсією вихідного ряду j.

Абсолютне значення коефіцієнту кореляції змінюється від 0 до 1 ($0 \le |r_{ij}| \le 1$). Знак "–" означає існування оберненої залежності між двома змінними. Якщо $r_{ij} = 0$, то лінійного зв'язку між змінними не існує (величини не корельовані), якщо $r_{ij} = 1$, то існує функціональний зв'язок.

Матриця кореляцій має властивості аналогічні властивостям матриці коваріації, тобто вона дійсна симетрична і додатно визначена.

Контрольні запитання.

1. Дати визначення матриці.
2. Дати визначення квадратної матриці.
3. Дати визначення симетричної квадратної матриці.
4. Назвати властивості матриць коваріацій та кореляцій.

ЛЕКЦІЯ 9
«РІВНЯННЯ ЛІНІЙНОЇ ПАРНОЇ РЕГРЕСІЇ. ДИСПЕРСІЙНИЙ АНАЛІЗ»

9.1 Рівняння лінійної парної регресії

Умовним законом розподілу величини Y, яка входить до системи (X,Y), називається її закон розподілу, обчислений за умови, що інша випадкова величина X прийняла визначене значення x. Умовний закон розподілу, який може бути заданий інтегральною функцією розподілу чи щільністю розподілу, записується в такий спосіб:

$$F(y/x) \text{ чи } f(y/x).$$

Умовний закон розподілу характеризується статистичними параметрами, такими як, умовне математичне сподівання $m_{y/x}$, умовна дисперсія або середнє квадратичне відхилення $\sigma_{y/x}$.

Для системи нормально розподілених величин (X,Y) умовне математичне сподівання $m_{y/x}$ записується у виді

$$m_{y/x} = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x), \tag{9.1}$$

а умовне середнє квадратичне відхилення –

$$\sigma_{y/x} = \sigma_y\sqrt{1 - r_{xy}^2}\ . \tag{9.2}$$

де σ_x, σ_y – середні квадратичні відхилення випадкових величин X та Y;

r_{xy} – коефіцієнт кореляції між випадковими величинами X та Y;

m_y, m_x – безумовні математичні сподівання випадкової величини X та випадкової величини Y.

З формул (9.1) і (9.2) випливає, що умовний закон розподілу величини Y характеризується тільки зміною математичного сподівання при зміні x, а умовна дисперсія від x не залежить. Величина $m_{y/x}$ називається умовним математичним сподіванням величини Y при заданому x. Залежність (9.1) можна зобразити на площині xOy, відкладаючи умовне математичне сподівання по осі ординат. У відповідності з (9.2), отримаємо пряму лінію, яка називається лінією регресії Y по X (рис. 9.1). Рівняння регресії X по Y може бути представленим у вигляді

$$m_{x/y} = m_x + r_{xy}\frac{\sigma_x}{\sigma_y}\left(y - m_y\right). \qquad (9.3)$$

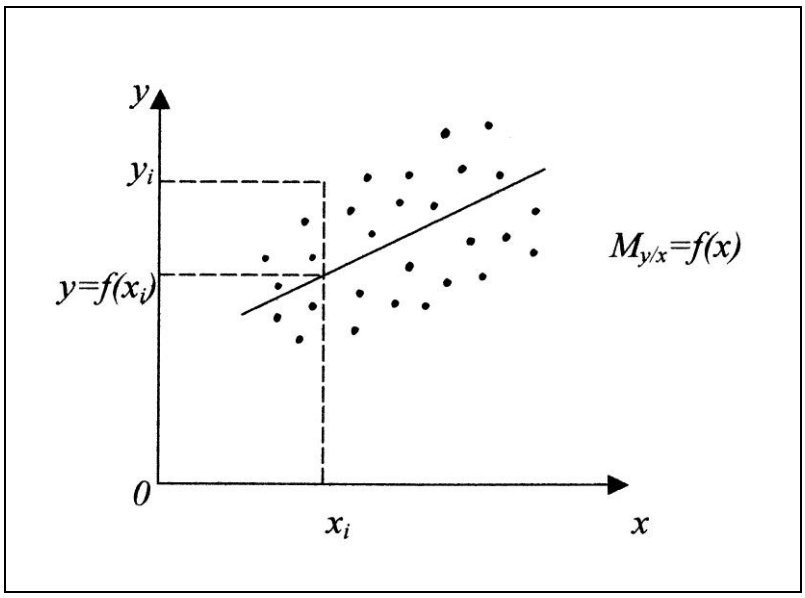

Рисунок 9.1 – Графічний вигляд рівняння лінійної регресії $m_{y/x} = f(x)$

Рівняння (9.1) і (9.3) можуть бути записані у вигляді рівнянь прямих ліній

$$m_{y/x} = \alpha_1 x + \beta_1 \ , \tag{9.4}$$

$$m_{x/y} = \alpha_2 y + \beta_2 \ , \tag{9.5}$$

де $\alpha_1, \beta_1, \alpha_2, \beta_2$ – коефіцієнти рівнянь регресії, які описують прямі

$$\alpha_1 = r_{xy} \frac{\sigma_y}{\sigma_x}; \quad \beta_1 = m_y - \alpha_1 m_x \ ; \tag{9.6}$$

$$\alpha_2 = r_{xy} \frac{\sigma_x}{\sigma_y}; \quad \beta_2 = m_x - \alpha_2 m_y \ . \tag{9.7}$$

Величини m_x, m_y – це безумовні математичні сподівання випадкових величин X і Y, які є центрами розподілу цих величин без урахування залежності між X і Y.

Для вибіркових даних рівняння умовного математичного сподівання представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = a x_i + b, \tag{9.8}$$

де x_i – дискретні значення випадкової величини X;

y – дискретні значення випадкової величини Y;

$\tilde{y}_i$ – значення випадкової величини Y, розраховані за рівнянням регресії (9.8); a, b – шукані параметри рівняння. функція регресії $\tilde{y}(x)$ є функція, яка мінімізує середню квадратичну похибку передбачення випадкової величини Y на основі значень X. Відповідним чином можна інтерпретувати і $\tilde{x}(y)$.

Оцінки параметрів, які входять до рівняння регресії, обчислюються за методом найменших квадратів. Це метод обробки емпіричного числового матеріалу, вимога якого полягає в тому, щоб сума квадратів відхилень даних спостережень y_i від лінії регресії була найменшою, тобто

$$\Delta = \sum_{i=1}^{n} \left[y_i - \tilde{y}(x_i) \right]^2 = min, \tag{9.9}$$

де y_i – спостережені значення випадкової величини Y;

$\tilde{y}(x_i)$ – значення випадкової величини Y, розраховані по рівняннях регресії для заданих x_i;

n – число спільно спостережених значень y_i і x_i.

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

$$\frac{\partial \Delta(a,b)}{\partial a} = \frac{\partial\left[\left(\sum_{i=1}^{n} y_i - ax_i - b\right)^2\right]}{\partial a} = 0 \ ; \qquad (9.10)$$

$$\frac{\partial \Delta(a,b)}{\partial b} = \frac{\partial\left[\left(\sum_{i=1}^{n} y_i - ax_i - b\right)^2\right]}{\partial b} = 0 \qquad (9.11)$$

Результатом розрахунків є такі формули для визначення параметрів a та b

$$a = r_{x,y}\frac{\sigma_y}{\sigma_x}. \qquad (9.12)$$

$$b = \bar{y} - a\bar{x} \ , \qquad (9.13)$$

де σ_x, σ_y – середні квадратичні відхилення випадкових величин X та Y;

r_{xy} – коефіцієнт кореляції;

$\bar{x}, \bar{y}$ – середні арифметичні значення випадкових величини X Y.

Прикладом застосування регресійного аналізу у практиці гідроекологічних досліджень може слугувати рівняння регресії, які показують зв'язок між мінералізацією води та питомою електропровідністю, густиною та мінералізацією води (за прожареним залишкам розчинених у воді речовин), які вимірювалися під час виконання працівниками кафедри гідроекології та водних дослідженьнаукових

досліджень з гідроекологічного обстеження стану Куяльницького лиману та морської води з Одеської затоки в 2016-2021 рр. (*2021*).

За результатами вимірювань в Куяльницькому лимані, Одеській затоці та водотоках, що впадають до лиману, були установлені лінійні зв'язки між значеннями питомої електропровідності, густини та мінералізації води (за прожареним залишком розчинених у воді речовин). Приклади цих залежностей показані нижче (рис.9.2), рис.(9.3). Отримані зв'язки апроксимовані лінійними рівняннями і мають високі коефіцієнти кореляції (r), тому можуть бути використані для експрес-визначення (не більш ніж за 5 хв.) мінералізації води в цих водних об'єктах у майбутньому, у тому числі, під час проведення їх діагностичного моніторингу.

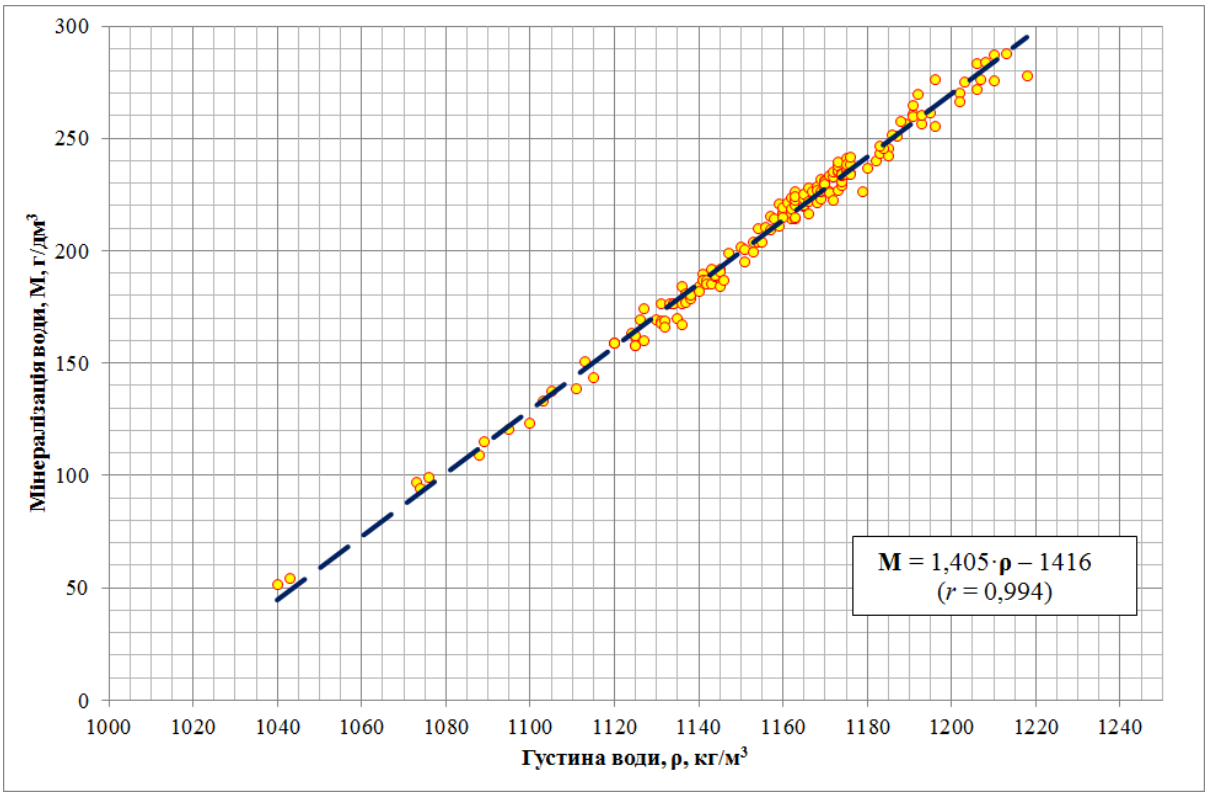

Рисунок 9.2 – Зв'язок між густиною та мінералізацією води (за прожареним залишком) в Куяльницькому лимані за період 2016-2021 рр. *(Гриб, О.М. 2021)*

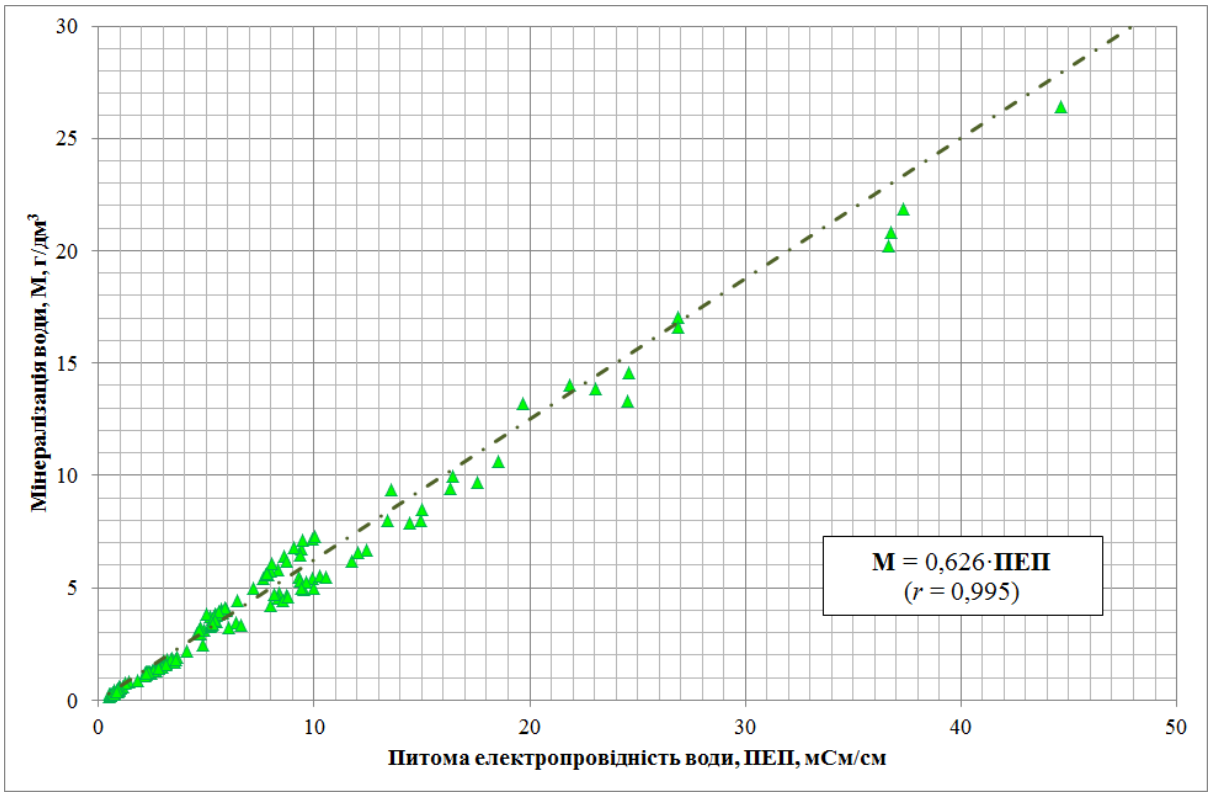

Рисунок 9.3 – Зв’язок між питомою електропровідністю та мінералізацією води (за прожареним залишком) в річках, балках і скидних лотках лиману за період 2016-2021 рр. *(Гриб О.М 2021)*

За отриманими залежностями можна установлювати мінералізацію для водних об’єктів, на яких не виконуються гідрохімічні спостереження. Щоб установити мінералізацію достатньо провести вимірювання питомої електропровідності води (ПЕП, мСм/см), використовуючи pH-метр-мілівольтметр з електродом ЭСКЛ-08М.1 (№ 0516), pH-150 М.

9.2 Дисперсійний аналіз побудованих рівнянь регресії. Регресійна та залишкова складові дисперсії

Загальну дисперсію σ_y^2 випадкової величини Y, яка характеризує розсіяння точок відносно генеральної середньої m_y можна представити у вигляді суми залишкової $\sigma_{зал}^2$ та регресійної σ_p^2 складових

$$\sigma_y^2 = \sigma_p^2 + \sigma_{зал}^2 \qquad (9.14)$$

де σ_p^2 – частина повної дисперсії випадкової величини Y, яка являє собою тільки ту частину розсіювання змінної Y, яка обумовлена впливом змінної X,

$\sigma_{зал}^2$ – частина розсіювання випадкової величини Y відносно її безумовного математичного сподівання m_y, яка виникає за рахунок випадковостей та неврахованих факторів.

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

Складова загальної дисперсії $\sigma_{зал}^2$ має назву „непоясненої дисперсії" або „залишків".

Як міру впливу змінної X на змінну Y використовують співвідношення виду

$$\eta_{y/x}^2 = \frac{\sigma_p^2}{\sigma_y^2}, \qquad (9.15)$$

Величина $\eta_{y/x}^2$ називається кореляційним відношенням або коефіцієнтом детермінації і змінюється у межах $0 \le \eta_{y/x}^2 \le 1$. Чим ближче значення $\eta_{y/x}^2$ до одиниці, тим більший внесок у загальну дисперсію випадкової величини Y вносить розсіювання Y, обумовлене впливом X. Таким чином, чим більше $\eta_{y/x}^2$, тим більш тісний зв'язок між Y та X. За умови лінійного стохастичного зв'язку між випадковими величинами X та Y (лінійна регресія) справедливий вираз

$$\eta_{y/x}^2 = r_{yx}^2, \qquad (9.16)$$

де r_{yx} – коефіцієнт кореляції між випадковими величинами Y та X.

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

$$\sigma_y^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}. \qquad (9.17)$$

В основі формули (9.17) лежить відхилення спостереженої величини y_i від середнього арифметичного значення (рис. 9.4).

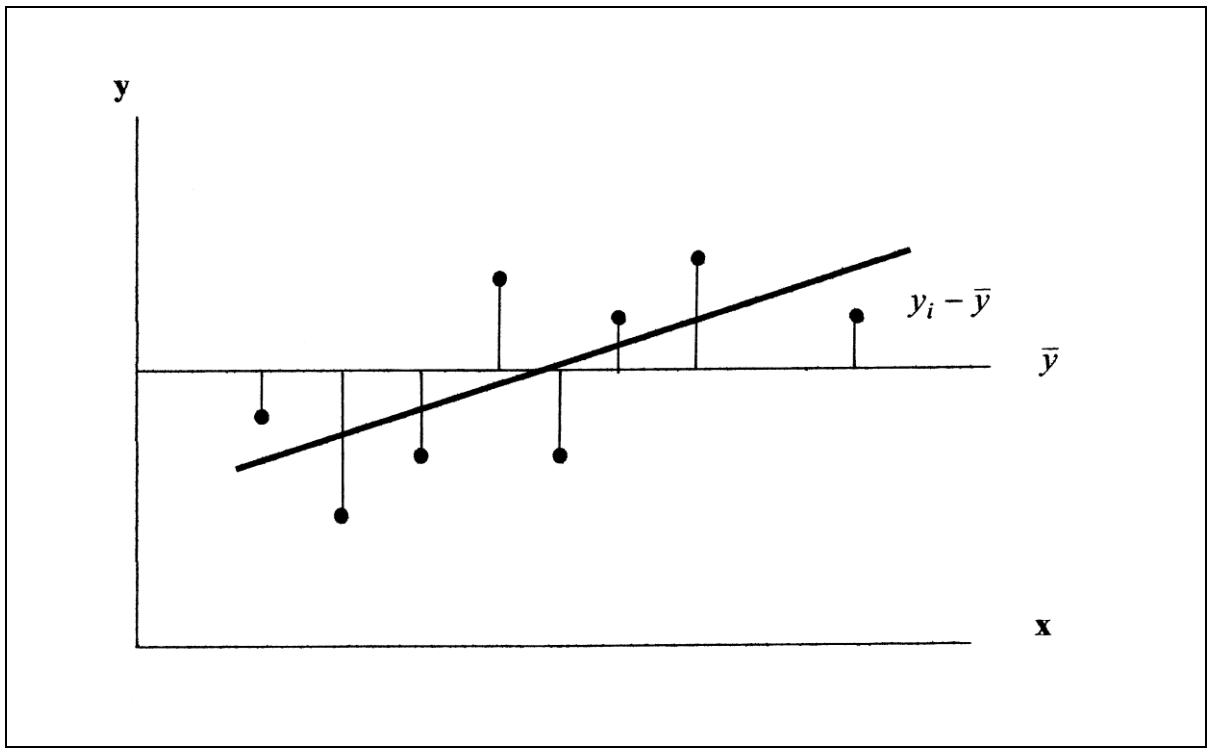

Рисунок 9.4 – Ілюстрація розсіювання спостережених значень y_i від $\bar{y}$

——— – лінія регресії

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

$$y_i - \bar{y} = (y_i - \tilde{y}_i) + (\tilde{y}_i - \bar{y}), \text{ або } y_i - \bar{y} = (\tilde{y}_i - \bar{y}) + (y_i - \tilde{y}) \qquad (9.18)$$

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

Обидві частини рівняння зведемо у квадрат

$$(y_i - \bar{y})^2 = \left[(\tilde{y}_i - \bar{y}) + (y_i - \tilde{y})\right]^2. \qquad (9.19)$$

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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i) + \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \qquad (9.20)$$

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

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-1} \qquad (9.21)$$

Рівняння (9.21) представляє собою вираз (9.10) записаний для вибірок, тобто

$$\hat{\sigma}_y^2 = \hat{\sigma}_p^2 + \hat{\sigma}_{зал}^2. \qquad (9.22)$$

Відхилення лінії регресії $(\tilde{y}_i)$ від $\bar{y}$ графічно представлене на рис. 9.5.

Величина $(\tilde{y}_i - y_i)$ характеризує розсіювання точок, які відповідають розсіюванню даних спостережень від лінії регресії (рис. 9.6).

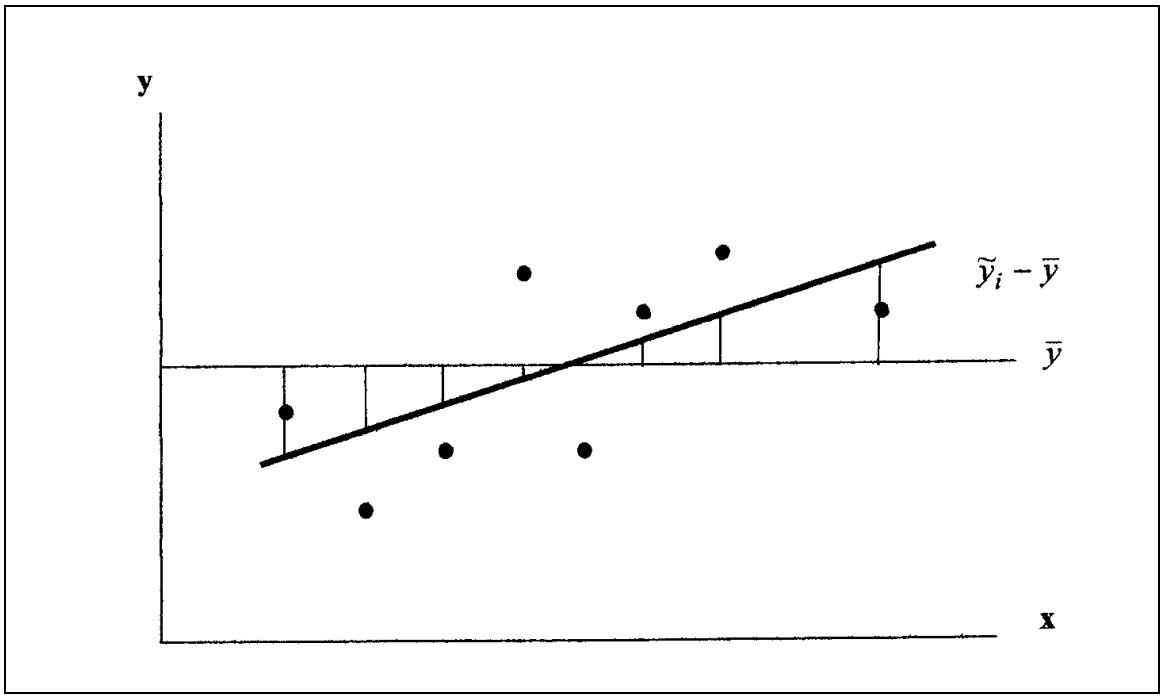

Рисунок 9.5 – Ілюстрація відхилення лінії регресії від $\bar{y}$

——— – лінія регресії

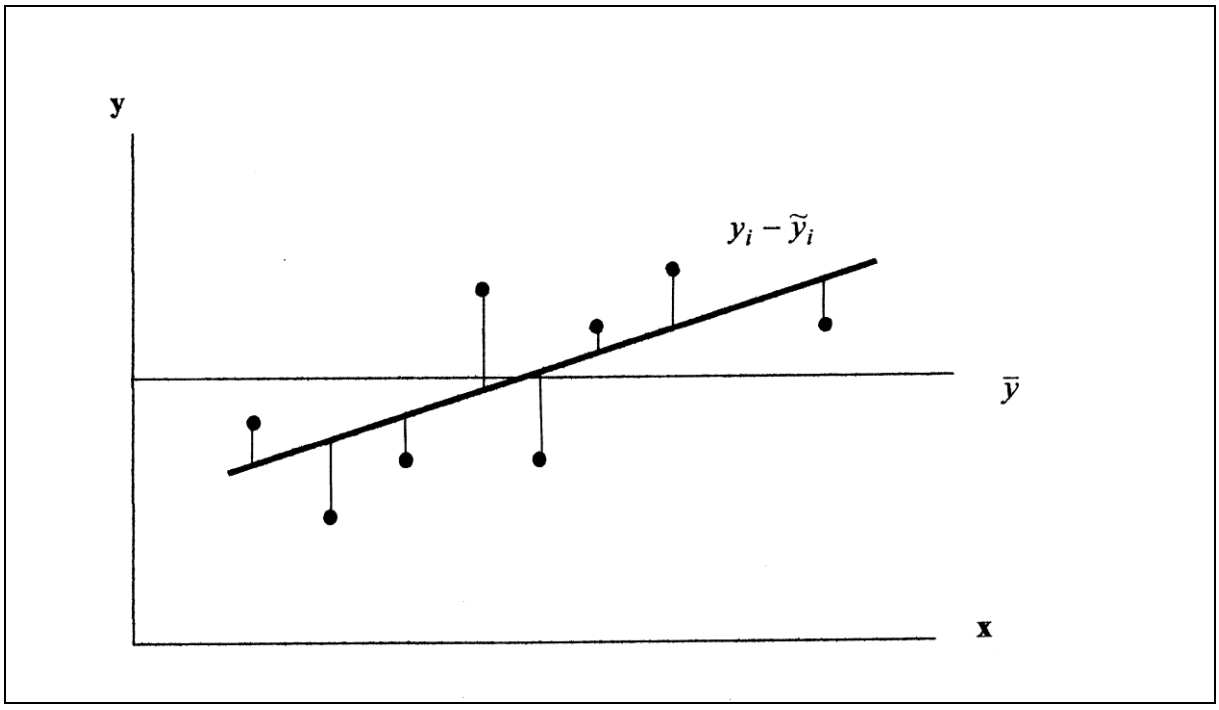

Рисунок 9.6 – Ілюстрація відхилення спостережених даних від лінії регресії

——— – лінія регресії

9.3 Перевірка гіпотези про відповідність регресійної моделі

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

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 / n-1}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n-1} = \frac{\hat{\sigma}_y^2}{\hat{\sigma}_{зал}^2}. \qquad (9.23)$$

Гіпотеза H_0 про те, що залишкова дисперсія незначуще відрізняється від загальної дисперсії, не відхиляється, коли для критерію Фішера F виконується умова

$$F < F_{кр}(\alpha, \nu_1, \nu_2), \qquad (9.24)$$

де $\nu_1 = n-1;\ \nu_2 = n-2$.

Контрольні запитання

1. Записати вираз рівняння лінійної парної регресії та формули для визначення коефіцієнтів цього рівняння.
2. Записати вираз для розрахунку парного коефіцієнта кореляції.
3. Сутність методу найменших квадратів.
4. Граничні значення коефіцієнту кореляції.
5. Як визначається коефіцієнт детермінації?

ЛЕКЦІЯ 10
«РЕГРЕСІЙНІ МОДЕЛІ»

10.1 Рівняння множинної лінійної регресії.
10.2 Визначення коефіцієнтів рівняння за матрицею кореляцій.
10.3 Коефіцієнт множинної кореляції.
10.4 Способи добору оптимальних предикторів.
10.5 Частинні коефіцієнти кореляції.

10.1 Рівняння множинної лінійної регресії. Визначення коефіцієнтів рівняння за матрицею кореляцій. Коефіцієнт множинної кореляції.

Математична модель множинної лінійної регресії представляється рівнянням виду

$$\widetilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_{2)} + b_3(x_{3i} - \bar{x}_{3)} \ldots + b_k(x_{ki} - \bar{x}_{k)} \quad (10.1)$$

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

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

$b_1, b_2, b_3, \ldots b_k$ – коефіцієнти рівняння множинної лінійної регресії;

k – число предикторів.

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

$$b_j = \frac{\sigma}{\sigma_j}\frac{D_{0j}}{D_{00}}, \quad (10.2)$$

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

$$D = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{vmatrix}, \qquad (10.3)$$

а мінори цього визначника записуються таким чином

$$D_{00} = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}, \qquad (10.4)$$

$$D_{01} = \begin{vmatrix} r_{10} & r_{12} \\ r_{20} & r1 \end{vmatrix}, \qquad (10.5)$$

$$D_{02} = \begin{vmatrix} r_{10} & 1 \\ r_{20} & r_{21} \end{vmatrix} \qquad (10.6)$$

У записах виду (10.3-10.6) r_{0j} – коефіцієнт кореляції між предиктантом (0) та предиктором (j).

Рівняння лінійної множинної регресії для двох предикторів буде мати вигляд

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2), \qquad (10.7)$$

де

$$b_1 = \frac{\sigma}{\sigma_1}\frac{D_{01}}{D_{00}} \tag{10.8}$$

$$b_2 = \frac{\sigma}{\sigma_2}\frac{D_{02}}{D_{00}} \tag{10.9}$$

Рівняння (10.7) може бути записане і таким чином

$$\tilde{y}_i = b_1 x_{1i} + b_2 x_{2i} + b_0, \tag{10.10}$$

де

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2, \tag{10.11}$$

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 визначене за наступною залежністю

$$\sigma_{\tilde{y}} = \sigma\sqrt{1 - R^2}, \tag{10.12}$$

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

$$R = \sqrt{1 - \frac{D}{D_{00}}}, \tag{10.13}$$

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) незначуще відрізняється від дисперсії предиктанта. Якщо така гіпотеза приймається, то прогноз (розрахунок) по моделі не відрізняється від випадкового.

10.2 Способи добору оптимальних предикторів при побудові рівняння множинної лінійної регресії

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, у зв'язку з чим матриця кореляції може бути погано обумовленою.

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

Вибір оптимальних предикторів виконується за наступною схемою:

1. З числа предикторов вибирається той, який має найтісніший 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 предиктора вже враховано, вибирається найбільший за абсолютною величиною (номер 2) і знов розраховується матриця частинних коефіцієнтів кореляції, але вже з урахуванням впливу перших двох предикторів.

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

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

Вибір оптимальних предикторів називають операцією "просіювання". Просіювання може відбуватися за допомогою частинних коефіцієнтів кореляції.

10.3 Визначення частинних коефіцієнтів кореляції

Приведемо визначення частинного коефіцієнта кореляції й розглянемо алгоритм його оцінки. Припустимо, що на випадкову величину Y впливають дві випадкові величини X_1 и X_2. *Частинним коефіцієнтом кореляції між випадковими величинами* Y *и* X_1 ($r_{yx_1 \bullet x_2}$) *називають коефіцієнт кореляції між ними при умові, що вплив другої випадкової величини* X_2 *на* Y *вже є врахованим.* Таким же чином визначається частинний коефіцієнт кореляції $r_{yx_2 \bullet x_1}$ (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д., 1999*).

Будемо вважати, що нам відома матриця кореляцій

$$R_x = \begin{vmatrix} 1 & r_{x_2} \\ r_{x_2 x_1} & 1 \end{vmatrix} \tag{10.14}$$

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

$$R_{yx} = \begin{vmatrix} r_{yx_1} \\ r_{yx_2} \end{vmatrix} \tag{10.15}$$

На їх основі сформуємо розширену матрицю кореляцій

$$\tilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1 x_2} \\ r_{yx_2} & r_{x_1 x_2} & 1 \end{vmatrix} \tag{10.16}$$

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 (10.16) розрахуємо мінори $|R_x|, D_{yx_i}, D_{\overline{yx}_i}$ $(i=1,2)$. Мінори D_{yx_i} складаються таким чином: стовпець, на першому місці котрого розташовується парна кореляція r_{yx_i}, переставляється на перше місце, а на його місці становиться перший стовпець і, після цього, викреслюють перші рядок і стовпець. Очевидно:

$$D_{yx_1} = r_{yx_1} - r_{yx_2} r_{x_1x_2}, \tag{10.17}$$

$$D_{yx_2} = r_{yx_2} - r_{yx_1} r_{x_1x_2} \tag{10.18}$$

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

$$D_{\overline{yx}_1} = \begin{vmatrix} 1 & r_{yx_2} \\ r_{yx_2} & 1 \end{vmatrix} = 1 - r_{yx_2}^2, \tag{10.19}$$

$$D_{\overline{yx}_2} = \begin{vmatrix} 1 & r_{yx_1} \\ r_{yx_1} & 1 \end{vmatrix} = 1 - r_{yx_1}^2 \tag{10.20}$$

Частинні коефіцієнти кореляції визначаються таким чином:

$$r_{yx_1 \bullet x_2} = \frac{D_{yx_1}}{\sqrt{|R_x| D_{\overline{yx}_1}}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{x_1x_2}^2\right)\left(1 - r_{yx_2}^2\right)}}, \tag{10.21}$$

$$r_{yx_2 \bullet x_1} = \frac{D_{yx_2}}{\sqrt{|R_x| D_{\overline{yx}_2}}} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{x_1x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} \tag{10.22}$$

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 (10.21) і (10.22).

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

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

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

$$\tilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \dots & r_{yx_k} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} & \dots & r_{x_2x_n} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} & \dots & r_{x_3x_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 & \dots & r_{x_kx_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & r_{x_nx_3} & \dots & r_{x_nx_k} & \dots & 1 \end{vmatrix} \qquad (10.23)$$

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

$$r_{yx_k \bullet x_1, x_{2,\ldots}, x_{k-1}, x_{k+1}, \ldots x_m} = \frac{D_{yx_k}}{\sqrt{|R_x| D_{\overline{yx_k}}}}, \qquad (10.24)$$

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.

1. Розглянемо приклад побудови рівнянь множинної регресії для розрахунків стоку наносів лівобережних приток Дністра у його верхній течії (Поділля) (*Мельник С. В, Лобода Н.С., 2019*).

для періоду 1953-1982 гг.

$$M_R^{\partial} = -426\ln(K_{підз}) + 0{,}207 M_{\max} M_{CEP} + 1{,}07 H_{CEP} + 1393, R = 0{,}95 \qquad (10.25)$$

для періоду 1983-2010 гг.

$$M_R^{в} = -39{,}4\ln(K_{підз}) + 0{,}515L + 0{,}252 H_{CEP} + 83{,}5, R = 0{,}96 \qquad (10.26)$$

де M_R^{∂} – модуль стоку наносів у додатний (позитивний) період коливань річного стоку;

$M_R^{в}$ – модуль стоку наносів у від'ємний період коливань річного стоку;

$K_{підз}$ – частка підземного стоку (%);

H_{CEP} $H_{вс}$ – середня висота водозбору (м);

M_{CEP} – середній багаторічний річний стік води (дм^3/с$*$км^2),

$M_{\max}$ – середній багаторічний максимальний стік води у рік (дм^3/с$*$км^2);

L – відстань від створу досліджень до вище розташованого водосховища або ставка (км).

Для додатної фази коливань стоку води (1953-1982 рр.) до числа предикторів входять показники стоку води (M_{max}, M_{CEP}), а у від'ємну фазу коливань ці показники перестають впливати на формування стоку наносів. Такий результат обумовлений існуванням поверхневого змиву ґрунту підчас паводків чи водопіль. Якщо роки маловодні, то вплив паводків і водопіль значно менший, отже характеристики стоку води не увійшли до числа оптимальних предикторів, проте у рівнянні з'явилася характеристика *L*, яка відображає вплив вище розташованих штучних водойм як відстійників для наносів та забруднювальних речовин.

Контрольні запитання

1. Записати вираз рівняння лінійної множинної регресії та вирази для визначення коефіцієнтів цього рівняння.
2. Записати вираз для розрахунку коефіцієнта множинної кореляції.
3. Граничні значення коефіцієнту множинної кореляції.
4. Назвати способи добору оптимальних предикторів при побудові рівнянь множинної регресії

ЛЕКЦІЯ 11
«МОДЕЛЬ 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 змінних, але описують кожну з них із своєю вагою (факторним навантаженням). Визначені фактори не повністю описують вихідні змінні. Неописана частина інформації називається залишками. Основна перевага методу факторного аналізу полягає в тому, що велика кількість вихідних змінних, які корелюють між собою, описується набагато меншим числом статистичних властивостей або факторів. Фактори не корелюють між собою.

Задача факторного аналізу полягає у представлені даних спостережень у вигляді лінійних комбінацій факторів:

$$x_j = \sum_{p=1}^{k} l_{jp} f_p + \nu_j , \ (j = 1, m) \qquad (11.1)$$

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;

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

Якщо у рівності (11.1) розгорнути суму, то прийдемо до такої системи рівнянь (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.,1999*):

$$\begin{cases} x_1 = l_{11}f_1 + l_{12}f_2 + ... + l_{1k}f_k + v_1 \\ x_2 = l_{21}f_1 + l_{22}f_2 + ... + l_{2k}f_k + v_2 \\ ..... \\ x_m = l_{m1}f_1 + l_{m2}f_2 + ... + l_{mk}f_k + v_m \end{cases} \quad (11.2)$$

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

Матрична форма рівняння (11.1) має вигляд

$$X = LF + V, \quad (11.3)$$

де X – матриця центрованих вихідних величин;

L – матриця факторних навантажень;

V – матриця незалежних залишків.

Матриця коваріацій знаходиться як

$$K = M\left[X \cdot X^{'}\right] \quad (11.4)$$

Ураховуючи (11.3), можна прийти до такого матричного виразу

$$K = LL' + D, \quad (11.5)$$

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 центрованих 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 ваг та дисперсій залишків може відбуватися на основі методу найменших 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 факторних вагів l_{jp} та дисперсій залишків d_j, використовуючи частинні похідні $\frac{\partial L_П}{\partial l_{jp}}$ та $\frac{\partial L_П}{\partial d_j}$

$$\frac{\partial L_П}{\partial l_{jp}} = 0; \quad \frac{\partial L_П}{\partial d_j} = 0, \tag{11.6}$$

де $L_П$ – функція правдоподібності.

Результатом пошуку є наступні рівнянні, які показують як елементи коваріаційної матриці виражаються через факторні навантаженнями й дисперсії залишків

$$K_{jj} = \sum_{p=1}^{k} l_{jp}^2 + d_j, \text{ при } \quad j = i; \tag{11.7}$$

$$K_{ji} = \sum_{p=1}^{k} l_{jp} l_{ip} \quad \text{при} \quad j \neq i. \tag{11.8}$$

Оскільки у матриці коваріацій на діагоналі розташовані дисперсії вихідних змінних, то можна зробити висновок, що квадрати факторних навантажень l_{ip}^2 є частками дисперсій змінних, які описуються відповідними факторами. Чим більша дисперсія залишків d_j, тим нижча якість опису змінних через фактори.

Оскільки вибіркова коваріаційна матриця може бути розрахованою, то пошук факторних навантажень та дисперсій залишків відбувається шляхом ітераційного процесу (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.,1999*).

Сума квадратів факторних навантажень по всіх виділених факторах може бути розрахованою таким чином

$$h_j^2 = \sum_{p=1}^{k} l_{jp}^2 \qquad (11.9)$$

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

$$S_p = \frac{\sum_{j=1}^{m} l_{pj}^2}{m} 100\%, \qquad (11.10)$$

де m – кількість розглядуваних змінних.

Загальний внесок всіх виділених факторів в сумарну дисперсію досліджуваних змінних дорівнює

$$S = \sum_{p=1}^{k} S_p \ . \qquad (11.11)$$

Величина S має назву міри факторизації. **Як правило, при практичному застосуванні до розгляду приймається така кількість факторів, яка описує понад 70% сумарної дисперсії змінних.**

Про ступінь зв'язку між змінними можна судити по певних графічних побудуваннях. Координатні осі розглядаються у таких випадках як осі факторних навантажень. Кожна змінна може бути представленою як вектор, проведений з початку координат у точку з координатами, які відповідають навантаженню кожного розглянутого фактору на цю змінну. Довжина такого вектору для двох факторів розраховується за виразом

$$H_j = \sqrt{l_{j1}^2 + l_{j2}^2} \ , \qquad (11.12)$$

де l_{j1} і l_{j2} – вагові коефіцієнти (факторні навантаження) першого та другого факторів на j-ту змінну.

Величина H ототожнюється з h (див.11.9). Вона визначає повноту відображення *j-ї* змінної першими двома факторами, а косинус кута між *i*-тим та *k*-тим вектором є коефіцієнтом кореляції між ними

$$r_{i,k} = \cos\varphi_{i,k}\,. \tag{11.13}$$

Кут між двома векторами є мірою кореляції між досліджуваними змінними.

Таким чином, про ступінь зв'язку між рядами змінних можна судити по угрупуваннях точок, які утворюються на площині. Як міра схожості в даному випадку використовується міра відстані: чим ближче розташовані точки на графіку й менше косинус кута між ними, тим ближче значення коефіцієнта кореляції до одиниці.

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

$$\lambda_j = arcsin\frac{l_{j2}}{\sqrt{l_{j1}^2 + l_{j2}^2}}, \tag{11.14}$$

де

$$H_j = \sqrt{l_{1j}^2 + l_{2j}^2 + l_{3j}^2}\,, \tag{11.15}$$

де θ та λ – полярні (сферичні) координати (в градусах або радіанах), які визначають положення пересікання j -тим вектором одиничної сфери.

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

Таким чином, ***графічна інтерпретація факторів дозволяє спростити аналіз інформації, яка міститься в кореляційній або коваріаційній матриці***. Компактне трактування кореляційної матриці досягається при розгляді кожного угрупування. При наближенні кута між

векторами, спрямованими з початку координат до центрів угрупувань, до 90°, виділені угрупування розглядаються як такі, що характеризуються тісними кореляційними зв’язками. Чим меньший кут між угрупуваннями, тим тісніший зв’язок між змінними.

На рисунку 11.1 показаний розподіл пестицидів у річці Волтурна (Італія) (*Maria Triassi, Antonio Nardone, Maria Cristina Giovinetti, Elvira De Rosa, Silvia Canzanella, Pasquale Sarnacchiaro, Paolo Montuori, 2019*). Видно, що уособлюється група Malathion та Methdathon (пестициди сучасного рівня), які мають тісний кореляційний зв'язок між собою. Пестициди давнього застосування (хлорпирифос, пиримифос-метил, фенитротион, толклофос-метил) з ними не пов’язані, оскільки утворюють угрупування під кутом близьким до 90 градусів, що означає наближення коефіцієнту кореляції між угрупуваннями до нуля (відсутність кореляційного зв’язку).

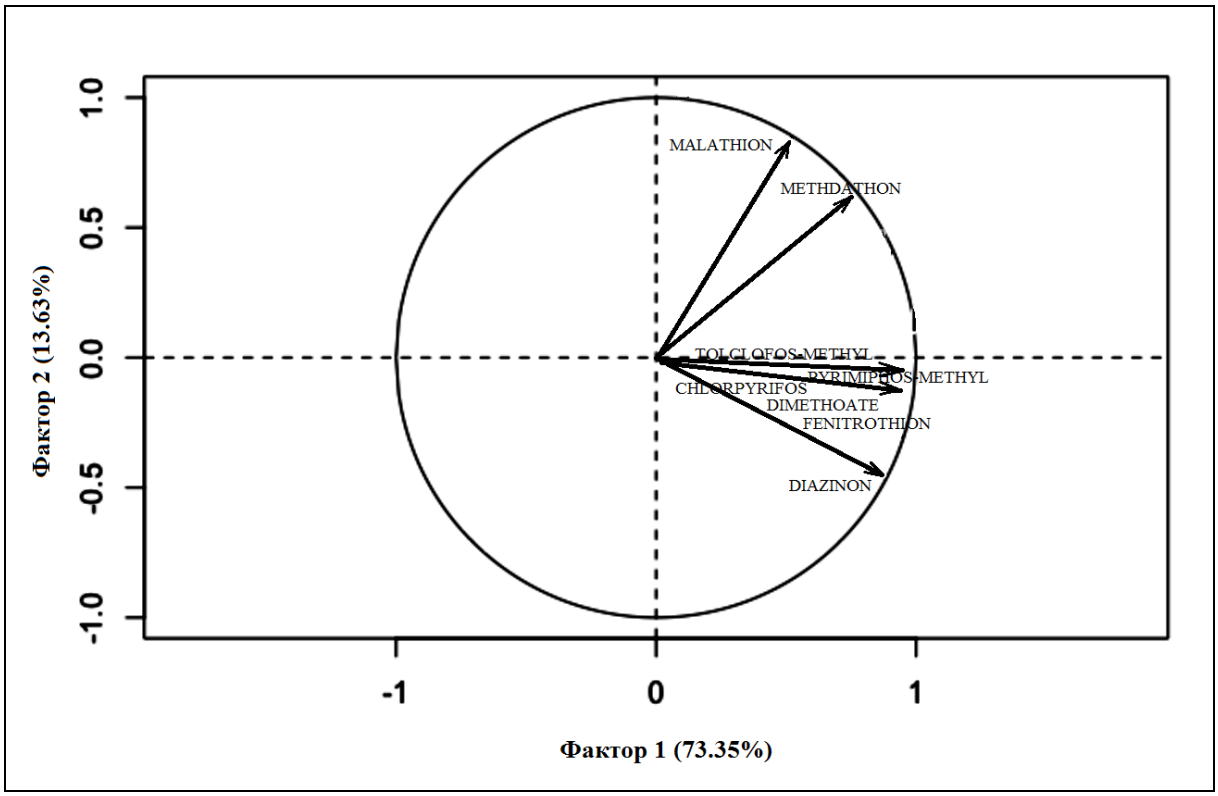

Рисунок 11.1 − Розподіл хімічних елементів в координатах двох перших факторів

Контрольні запитання

1. Які задачі можуть бути вирішені за допомогою факторного аналізу?
2. Як визначається кількість факторів, достатніх для опису вихідної інформації.
3. Класифікація чинників на основі геометричної інтерпретації факторів.

ЛЕКЦІЯ 12
«ДИСКРИМІНАНТНИЙ АНАЛІЗ ЯК МЕТОД ПРОГНОЗУ АБО ВИРІШАЛЬНЕ ПРАВИЛО»

12.1 Схема побудови розв'язувального правила

Дискримінантний аналіз є статистичним методом багатовимірного аналізу, який дозволяє визначати статистичну значущість різниці між двома або більше групами об'єктів по декількох змінних одночасно.

У гідроекології часто виникає задача щодо віднесення того чи іншого об'єкта або спостереження до одного з відомих класів. Такими класами можуть бути класи якості вод, до яких має бути віднесений той чи інший об'єкт. В практиці гідроекологічного прогнозування часто виникає потреба складати прогноз здійснення або нездійснення того чи іншого явища. Наприклад, можна прогнозувати ступінь ризику порушення екологічного стану водного об'єкту в результаті надходження комунальних стічних вод у поверхневий водоток, або прогнозувати реакцію водного об'єкту на забруднення сполуками азоту за допомогою коефіцієнта вразливості (*Loboda N. & Daus M., 2021*)

Прогноз за принципом: відбудеться явище чи ні, називають альтернативним. Поставлені перед дискримінантним аналізом задачі називаються задачами прийняття альтернативних рішень, або задачами класифікації. Питання про віднесення об'єкту до тієї чи іншої сукупності (групи) за принципом «так або ні» вирішується у такому випадку шляхом порівнювання ознак, характерних для кожної із розглядуваних сукупностей, із ознаками самого розглядуваного об'єкту (*Лобода Н.С., 2010*).

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 розділити весь простір образів на два підпростори, у першому з яких явище відбувається, а в другому – ні. По-друге, треба побудувати правило, за допомогою якого можна віднести образ, який підлягає розпізнаванню, до того чи іншого підпросторів.

Нехай ми маємо множину V векторів-предікторів (образів), що складають простір зображень Rv. Припустимо, що цей простір розділяється 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.

Наперед за все 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.12.1). На площині (x_1, x_2) показана крива $x_1 = f (x_2)$, яка відділяє підпростір R_{V1} від підпростору R_{V2}. У тривимірному просторі $R_V = R_V(x_1, x_2,x_3)$ (рис.12.2) підпростори R_{V1} і R_{V2} розділяє деяка поверхня, рівняння якої має вид $x_3=\varphi(x_1, x_2)$.

Більш складні умови виникають, коли розглядаються образи із багатовимірного простору $R_V = R_V(x_1,x_2,\ldots,x_n)$. Поверхня, що розділяє цей простір на підпростори R_{V1} і R_{V2} називається розділяючою гіперповерхнею, а її рівняння має вид:

$$F(x_1,x_2,\ldots,x_n) = 0. \qquad (12.1)$$

Задачею дискримінантного аналізу є отримання правила, 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 за розв’язувальним правилом $X \in R_{V1}$, то явище прогнозується («так»). Якщо $X \in R_{V2}$, то явище не прогнозується («ні»). **Етап, який складається з побудови розділяючої гіперповерхні та розв'язувального правила, носить назву етапу навчання.**

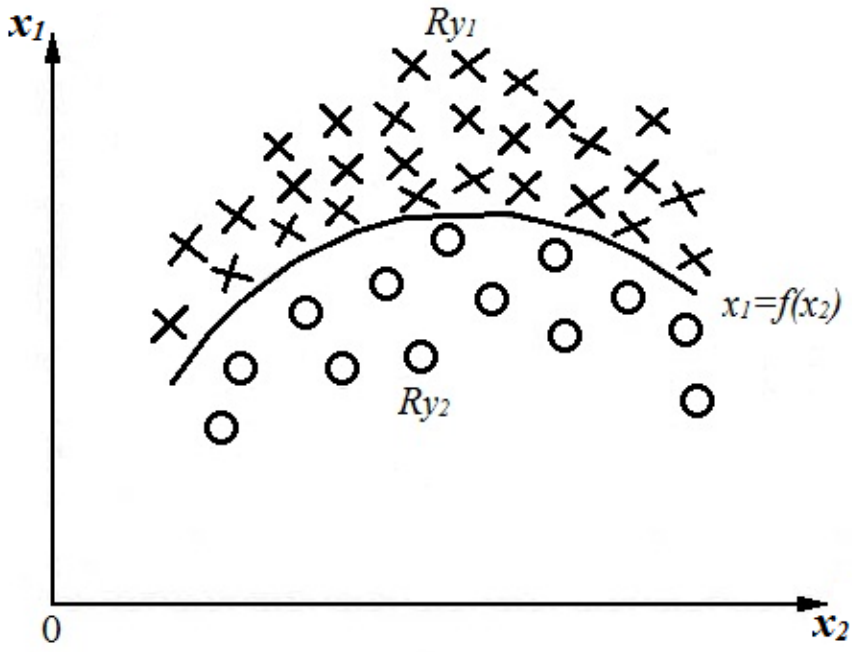

Рисунок 12.1 – Образи і розділяюча функція в двовимірному просторі

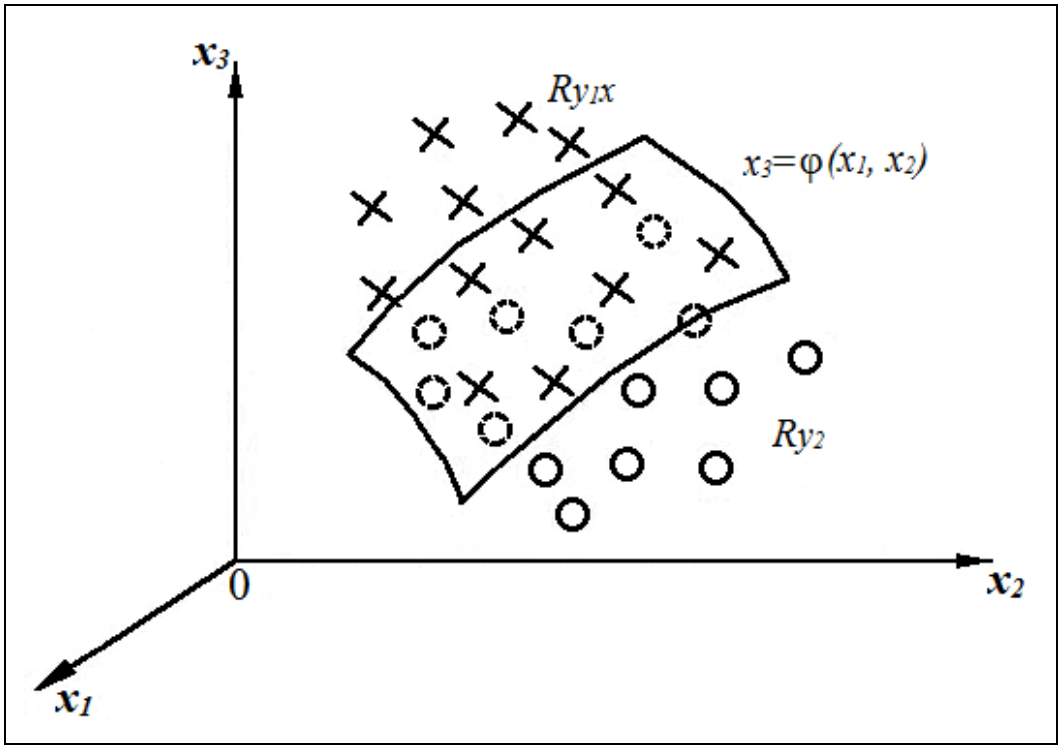

Рисунок 12.2 – Образи та розділяюча поверхня в трьохвимірному просторі

Прийняття рішення про належність вектора X до підпросторів R_{V1} чи R_{V2} називають етапом розпізнавання. Множина V векторів-предікторів, 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 представлене у вигляді деякої математичної функції, яку назвемо дискримінантною.***

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

Застосування дискримінантного аналізу, з однієї сторони, передбачає принципову роздільність класів, а з іншої сторони, допускає іхню часткову "перекриваність". Таким чином, рішення про віднесення явища або об'єкту до того чи іншого класу приймається з тією чи іншою долею похибки. ***"Перекриваність" 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 їх кількісна ознака лежить у області перекриття.

На рисунку 12.3 показана область перекриття функцій розподілу щільності ймовірностей величини X. Коли класифікація виконується по одній із ознак, то водний об'єкт за екологічним станом буде віднесений до першого класу, якщо ознака x буде меншою за x_0 ($x < x_0$), і до другого класу, якщо $x > x_0$. Неправильні рішення позначимо через δ_A та δ_B. У якості x_0 можна вибрати середину інтервалу перекриття $x_0^{'}$ або абсцису точки перетину кривих щільності розподілу $x_0^{''}$. Але більш доцільно вибрати таку точку x_0^{*}, для якої буде виконуватись $\alpha_A = \alpha_B$. За цієї умови максимальна похибка (найбільша з двох) буде приймати найменше 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 відмічалося, при використанні двох ознак розділ сукупностей відбувається відносно прямої, трьох – відносно поверхні, і т.д.

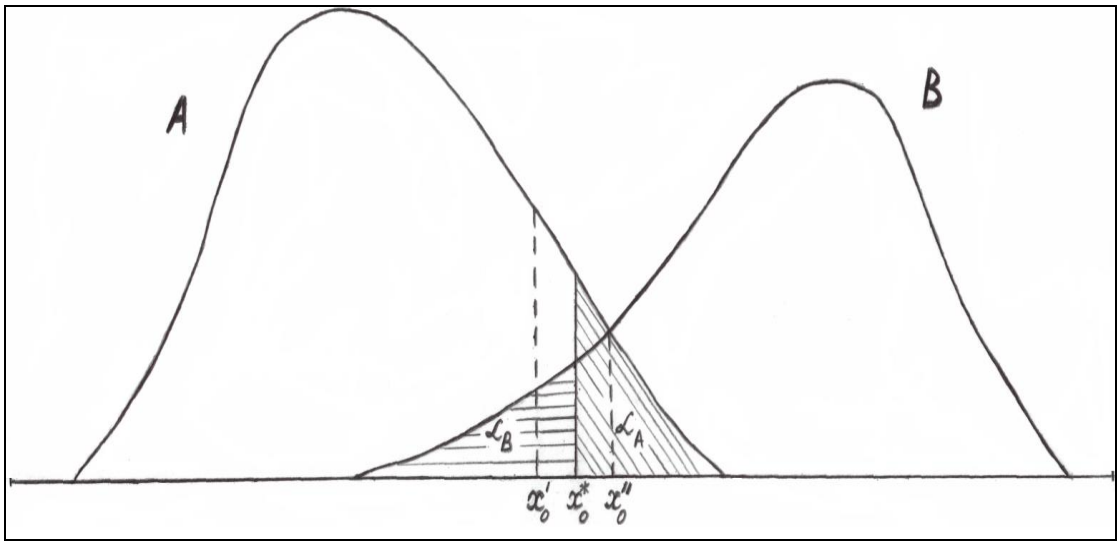

Рисунок 12.3 – Вирішальні правила при розмежуванні одновимірних сукупностей

Розглянемо основні ідеї теорії розпізнавання образів (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.,1999*). 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 обґрунтовано 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:

$$\mathrm{P}\left(x_1, x_2, ..., x_n / V_1\right) \tag{12.2}$$

та

$$\mathrm{P}\left(x_1, x_2, ..., x_n / V_2\right) \tag{12.3}$$

Позначимо ймовірність помилки 1-го роду через P_a, а 2-го роду через P_b. Знаючи ймовірності (12.2) та (12.3), 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

$$\lambda\left(x_{1,}x_{2,...,}x_{n}\right)=\frac{P\left(x_{1,}x_{2,...,}x_{n}/V_{1}\right)}{P\left(x_{1,}x_{2,...,}x_{n}/V2\right)} \tag{12.4}$$

а величину

$$\frac{\delta_{b}P\left(V_{2}\right)}{\delta_{a}P\left(V_{1}\right)}=\theta \tag{12.5}$$

називають порогом.

Розв'язувальне правило записується таким чином

вектор $X\in V_1$, якщо $\lambda(x_1,x_2,\ldots,x_n) > \theta$ (12.6)

вектор $X \in V_2$, якщо $\lambda(x_1,x_2,\ldots,x_n) < \theta$. (12.7)

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

$$\begin{gathered}\text{вектор } X \in V_1, \text{ якщо функція подібності } \lambda(x_1,x_2,\ldots,x_n) > 1,\\ \text{вектор } X \in V_2, \text{ якщо функція подібності } \lambda(x_1,x_2,\ldots,x_n) < 1.\end{gathered} \tag{12.8}$$

Ураховуючи (12.5), можна зазначити, що розв'язувальне правило базується на нерівності

$$\frac{P\left(x_{1,}x_{2,...,}x_{n}/V_{1}\right)}{P\left(x_{1,}x_{2,...,}x_{n}/V_{2}\right)}>\frac{\delta_{b}P\left(V_{2}\right)}{\delta_{a}P\left(V_{1}\right)}, \tag{12.9}$$

яка може бути представленою у логарифмичному виді

$$\ln\frac{P\left(x_{1},x_{2},....,x_{n}/V_{1}\right)}{P\left(x_{1},x_{2},...,x_{n}/V_{2}\right)}>\ln\frac{\delta_{b}P\left(V_{2}\right)}{\delta_{a}P\left(V_{1}\right)} \tag{12.10}$$

Функцію виду

$$F(x_{1,}x_{2,...,}x_n)=\ln P(x_{1,}x_{2,...,}x_n/V_1)-\ln P(x_{1,}x_{2,...,}x_n/V_2)+\frac{\delta_a P(V_1)}{\delta_b P(V_2)} \qquad (12.11)$$

називають дискримінантною функцією.

При перенесенні порогу у ліву частину (12.11) розв'язувальне правило приймає вид:

$$X \in V_{1,} якщо F(x_1, x_2, ..., x_n) > 0; \qquad (12.12)$$

$$X \in V_{2,} якщо F(x_1, x_2, ..., x_n) < 0; \qquad (12.13)$$

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

Методи, що основані на теорії статистичних рішень, мають такі обмеження: для їх реалізації необхідно знати щільності умовних розподілів образів у класах V_1 і V_2. Ці закони розподілів є багатовимірними, і на основі множин векторів-предікторів класів V_1 і V_2, отримання їх аналітичною виду є дуже складною задачею. 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 приймають, що класи векторів-предикторів підпорядковуються умовним нормальним законам розподілу. У 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 предікторів має одномодальний розподіл. Ця умова у більшості випадків виконується (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.,1999*).

12.2. Побудова розв'язувального правила. Квадратична, лінійна та спрощена дискримінантна функція

Будемо вважати, шо вектори-предиктори

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix} \tag{12.14}$$

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

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} \tag{12.15}$$

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

$$P(x_1, x_2, ..., x_n / V_1) = \frac{1}{(2\pi)^{n/2} |K_1|^{1/2}} \exp\left[-\frac{1}{2} (X - \mu_1) K_1^{-1} (X - \mu_1) \right] \tag{12.16}$$

$$P(x_1, x_2, ..., x_n / V_2) = \frac{1}{(2\pi)^{n/2} |K_2|^{1/2}} \exp\left[-\frac{1}{2} (X - \mu_2) K_2^{-1} (X - \mu_2) \right] \tag{12.17}$$

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

Після підстановки (12.16) і (12.17) до дискримінантної функції виду (12.11) прийдемо до такого рівняння:

$$\begin{aligned} F(x) = &-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |K_1| - \frac{1}{2} (X - \mu_1)' K_1^{-1} (X - \mu_1) + \\ &+ \frac{n}{2} \ln 2\pi + \frac{1}{2} \ln |K_2| + \frac{1}{2} (X - \mu_2)' K_2^{-1} (X - \mu_2) + \ln \frac{P(V_1)}{P(V_2)} \end{aligned} \tag{12.18}$$

Після скорочень дискримінантна функція набуде виду

$$F(x)=\frac{1}{2}\left[(X-\mu_2)K_2^{-1}(X-\mu_2)-(X-\mu_1)K_1^{-1}(X-\mu_1)+\ln\frac{|K_2|}{|K_1|}\right] + \ln\frac{P(V_1)}{P(V_2)} \tag{12.19}$$

Вважається, що ціни помилок першого і другого роду однакові $\delta_a=\delta_b$. *Дискримінантна функція F(x), яка визначається формулою (12.19), носить назву квадратичної дискримінантної функції*. Така назва пов'язується з тим, що перші два члени у квадратних дужках являють собою квадратичні форми, тобто многочлени степеня не більше другого.

Дискримінантна функція (12.20) містить у собі операції обернення коваріаційних матриць K_1 і K_2. Ця операція може привести до негативних наслідків, коли матриці коваріацій є погано обумовленими. У такому разі похибки, що містяться в коваріаціях, можуть привести до великих похибок розрахунків 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:

$$K_1 = K_2 = K, \tag{12.20}$$

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

$$K == \frac{K_1 + K_2}{2} \tag{12.21}$$

Якщо прийняти умову (12.21) у дискримінантній функції виду (12.19) і вважати, що $P(V_1) = P(V_2)$, то отримаємо:

$$F(x)=\frac{1}{2}\left[(X-\mu_2)'K^{-1}(X-\mu_2)-(X-\mu_1)'K^{-1}(X-\mu_1)\right]= =(\mu_1-\mu_2)'K^{-1}X+\frac{1}{2}\left[\mu_2'K^{-1}\mu_2-\mu_1'K^{-1}\mu_1\right] \tag{12.22}$$

Дискримінантна функція виду (12.22) 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 (дисперсій), і ними можна знехтувати.

Тоді

$$K = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \tag{12.23}$$

а її обернена матриця виглядає таким чином

$$K^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix}, \tag{12.24}$$

тобто операція обернення матриць коваріацій значно спрощується. При цьому спрощується й процедура розрахунків коефіцієнтів квадратичних форм у квадратичній дискримінантній функції (12.11).

Коли виконується умова (12.20) і, крім того, можна вважати коваріаційну матрицю діагональною, значно спрощується вид лінійної дискримінантної функції (12.22). У цьому випадку вона дорівнює:

$$F(x) = \sum_{i=1}^{n} \frac{\mu_{1i} - \mu_{2i}}{\sigma_i^2} x_i + \sum_{i=1}^{n} \frac{\mu_{2i}^2 - \mu_{1i}^2}{\sigma_i^2} \tag{12.25}$$

Рівняння (12.25) є спрощеною лінійною дискримінантною функцією.

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

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

$$\Delta = (\mu_1 - \mu_2)' K^{-1} (\mu_1 - \mu_2), \qquad (12.26)$$

а при використанні спрощеної лінійної дискримінантної функції (12.25)

$$\Delta = \sum_{i=1}^{n} (\mu_{1i} - \mu_{2i})^2 / \sigma_i^2 . \qquad (12.27)$$

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

Прогноз за дискримінантною функцією випускається таким чином:

якщо $F(x_1, x_2, ..., x_n) \geq 0$ *– прогнозується поява досліджуваного явища;*

якщо $F(x_1, x_2, ..., x_n) < 0$ *– досліджуване явище не прогнозується.*

Прогнозною датою появи досліджуваного явища $D^{/}$ є дата, коли:

$$D^{/} = D_{F \geq 0} \qquad (12.28)$$

де $D^{/}$ – прогнозна дата появи досліджуваного явища;

$D_{F \geq 0}$ – дата, на яку дискримінантна функція F за конкретних умов стає рівною або більшою за нуль.

Оцінка точності прогнозу виконується шляхом розрахунку забезпеченості допустимої похибки прогнозу

$$p = \frac{n - m}{n} \cdot 100\%, \qquad (12.29)$$

де p – забезпеченість прогнозу;

n – загальна кількість перевірочних прогнозів;

m – кількість прогнозів, що не виправдались.

Прогноз вважається «добрим», якщо забезпеченість ≥85% і «задовільним», якщо забезпеченість становить 84-60%.

Розглянемо приклад побудови дискримінантної функції при класифікації якості вод за ІЗВ (індексом забруднення води).

Гідрохімічний індекс забруднення ІЗВ відноситься до категорії показників, що найчастіше використовуються для оцінки якості водних об'єктів. Він визначається як середнє арифметичне значення перевищення концентрацій певних речовин (азот амонійний, азот нітритний, нафтопродукти, феноли, розчинений кисень, $БСК_5$) над ГДК:

$$ІЗВ = \frac{1}{6}\sum_{i=1}^{n}\frac{C_i}{ГДК_i}, \qquad (12.30)$$

де C_i – середня концентрація одного з шести показників якості води;

$ГДК_i$ – гранично допустима концентрація показників якості води у відповідності із галуззю водопостачання.

Індекс ІЗВ може бути модифікованим, якщо з шести перелічених компонент залишити $БСК_5$ та O_2, а інші показники обрати за найбільшим відношення до ГДК.

За таблицею 12.1 можна установити клас якості води в залежності від значення ІЗВ.

Задача полягає у побудові дискримінантної функції з використанням даних гідрохімічних спостережень за такими показниками: $БСК_5$, O_2, амоній, нітрити, нітрати, фосфати.

На першому етапі необхідно виділити дві навчальні вибірки, які розрізняються між собою ознаками. Виділяємо дві групи, які попадають у різні класи забруднення: I клас – «чисті» (0,3 < ІЗВ < 1,0), та II клас – «помірно забруднені» (1,0 < ІЗВ < 2,5).

На другому етапі необхідно побудувати спрощену лінійну дискримінантну функцію.

Таблиця 12.1 – Критерії оцінки якості вод за ІЗВ

Клас якості води	**Характеристика класу**	**Величина ІЗВ**
I	Дуже чиста	≤ 0,30
II	Чиста	0,31 – 1,00
III	Помірно забруднена	1,01 – 2,50
IV	Забруднена	2,51 – 4,00
V	Брудна	4,01 – 6,00
VI	Дуже брудна	6,01 – 10,0
VII	Надзвичайно брудна	> 10,0

Для побудови функції знаходимо середні значення середньорічних концентрацій забруднювальних речовин за весь період спостережень (1991-2011 рр.), для І класу та для ІІ класу. Дисперсія приймається як така, що характеризує дві групи у цілому).

В результаті розрахунків за (12.25) дискримінантна функція набуває виду

$$F(x)=6{,}72\frac{C_{(NO_3)}}{ГДК_{(NO_3)}}-3{,}37\frac{C_{(NO_2)}}{ГДК_{(NO_2)}}-2{,}86\frac{C_{(NH_4)}}{ГДК_{(NH_4)}}-1{,}55\frac{C_{(PO_4)}}{ГДК_{(PO_4)}}-$$
$$2{,}43-\frac{C_{(БСК5)}}{ГДК_{(БСК5)}}-5{,}28\frac{ГДК_{(O_2)}}{C_{(O_2)}}+39{,}0 \qquad (12.31)$$

На третьому етапі визначаємо число Махаланобіса, яке характеризує відстань між центрами виділених класів, і є характеристикою якості побудови дискримінантної функції

$$\Delta=\sum_{i=1}^{n}(\mu_{1i-}\mu_{2i})^2/\sigma_i^2=11{,}9$$

Оскільки Δ=11,9, що більше 11, то ймовірність похибки класифікації менше 5%. Це означає, що різниця між центрами класів є статистично значущою і класифікація є обґрунтованою.

На четвертому етапі перевіряємо як за даною функцією виконується класифікація, використовуючи дані незалежної вибірки, яка не увійшла у побудову дискримінантної функції. Наприклад, підставляючи дані гідрохімічних спостережень за 2019 рік у вираз (2,31) отримаємо

$$F(x) = -72.5$$

Оскільки $F(x)<0$, то якість води з ознаками ІЗВ належить до ІІ класу – «помірно забруднені» (1,0 < ІЗВ < 2,5), що відповідає дійсності.

Контрольні запитання

1. Який прогноз називається альтернативним.
2. У чому полягає задача розпізнавання образів.
3. З якою метою будується дискримінантна функція.
4. Назвати види дискримінантних функцій.
5. Критерій якості розрахункової або прогностичної методики, побудованої у вигляді дискримінантної функції.

ЛЕКЦІЯ 13
«МЕТОД ГОЛОВНИХ КОМПОНЕНТІВ (І)»

Метод головних компонентів є методом, який дозволяє замінити великий набір змінних на більш скорочений набір нових змінних, які містять у собі більшу частину інформації первинного набору (*Лобода Н.С., 2010*). Такий підхід має назву стискання інформації. Отримані нові змінні (компоненти) є некорельованими (ортонормованими). Перша компонента ураховує найбільшу частину інформації, яка міститься у вихідних даних. Кожна наступна компонента ураховує як можна більшу частину інформації, що не увійшла до попередніх компонент.

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 коваріацій або кореляцій визначається система ортогональних, лінійно незалежних функцій, які прийнято називати власними векторами. Вони відповідають системі незалежних випадкових величин, іменованих власними значеннями або власними числами матриці коваріацій або кореляцій. Пошук власних векторів та власних значень досягається шляхом розв’язання матричних рівнянь виду

$$R_X U_i - \lambda_i U_i = 0, \tag{13.1}$$

$$K_X W_i - \lambda_i W_i = 0, \tag{13.2}$$

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

K_X – матриця коваріацій розміром $m \times m$;

U_i – власний вектор матриці кореляцій;

W_i – власний вектор матриці коваріацій;

λ_i – відповідне власному вектору власне значення або власне число.

Власний вектор – це вектор U_i або W_i, який при даному лінійному перетворенні (13.1) або (13.2) не змінює свого напрямку, а тільки помножується на скаляр λ_i.

Власне значення матриці – це число λ_i, для якого існує ненульовий вектор U_i, який має властивість $R_X U_i = \lambda_i U_i$.

Власні вектори матриць кореляцій та коваріацій є ***ортонормованими*** й мають таку властивість

$$U_i^{'} U_j = \delta_{ij}, \tag{13.3}$$

де δ_{ij} – символ Кронекера, при цьому

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{13.4}$$

Сукупність ортонормованих власних векторів утворює ортогональну матрицю, яка має такі властивості

$$W^{'} W = E, \tag{13.5}$$

де E – одинична матриця.

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, необхідне розв'язання m систем лінійних рівнянь. Процедура розрахунку здійснюється, як правило, за допомогою ітераційних методів, серед яких найпоширенішим є метод Якобі (*Школьний Є.П., Лоєва І.Д,, Гончарова Л.Д.,1999*).

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

$$U' \cdot \varphi_i = Z_i; \tag{13.6}$$

$$W' \cdot \Delta X_i = Z_i, \tag{13.7}$$

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

W' – транспонована матриця W розміром $m \times m$;

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

ΔX_i – i-й випадковий вектор (поле) центрованих вихідних даних, що підлягає розкладанню;

Z_i – вектор головних компонент, який є результатом лінійного перетворення поля вихідних даних відповідним власним вектором.

Рівності (13.6) та (13.7) означають, що вхідне поле розкладене на m незалежних компонент в базисі власних векторів матриць кореляцій та коваріацій, відповідно.

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

$$z_{ip} = \sum_{k=1}^{m} U_{pk}\varphi_{ik}\,; \quad p = \overline{1,m}\,; \quad i = \overline{1,n} \tag{13.8}$$

або при використанні центрованих величин $\Delta X_i = x_i - \bar{x}$

$$z_{ip} = \sum_{k=1}^{m} W_{pk}\Delta X_{ik}\,; \quad p = \overline{1,m}\,; \quad i = \overline{1,n} \tag{13.9}$$

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) при використанні кореляційної матриці;

W_{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) при використанні коваріаційної матриці;

ΔX_{ik} – i-й випадковий вектор (поле) центрованих вихідних даних, що підлягає розкладанню;

φ_{ik} – i-й випадковий вектор (поле) центрованих та нормованих вихідних даних, що підлягає розкладанню.

Центровані та нормовані дані розраховуються наступним чином

$$\varphi_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}\,, \tag{13.10}$$

де x_{ik} – кожне значення i-того ряду спостережень при $i = \overline{1,n}$;

$\bar{x}_k$ – середнє арифметичне значення k-того ряду спостережень;

σ_k – середнє квадратичне відхилення k-того ряду спостережень.

Значення U_{pk} або W_{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 і має назву "базисної функції"***.

Оскільки базис власних векторів є ортогональним, то компоненти z_{ip} вектора Z_i є лінійно незалежними, а в статистичному значенні некорельованими.

Контрольні запитання

1. Що являє собою метод головних компонентів?
2. Що таке власний вектор?
3. Що таке власне значення матриці?
4. Чим утворюється ортогональна матриця?

ЛЕКЦІЯ 14
«МЕТОД ГОЛОВНИХ КОМПОНЕНТІВ (II)»

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

$$\lambda_P = \sigma^2_{Zp}, \tag{14.1}$$

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

$$\sum_{p=1}^{m} \sigma^2_{Zp} = \sum_{p=1}^{m} \lambda_p = t_R R_x = m, \tag{14.2}$$

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

Якщо розкладання за ПОФ виконувати в базисі власних векторів матриці коваріацій, то

$$\sum_{p=1}^{m} \sigma^2_{Zp} = \sum_{p=1}^{m} \lambda_p = t_R K_x = \sum_{p=1}^{m} \sigma^2_{Xp}. \tag{14.3}$$

Сума власних значень матриці кореляцій є, як це було показано раніше у (14.1), сумою дисперсій головних компонентів. Ця сума, отримана 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 (4.11), є першим власним значенням.

Сума дисперсій головних компонентів матриці коваріацій дорівнює сумі дисперсій вихідних рядів, тобто

$$\boxed{\sum_{p=1}^{m} \sigma^2_{Zp} = \sum_{j=1}^{m} \sigma^2_{Xj}}. \tag{14.4}$$

Таке подання дозволяє більш наочно зрозуміти суть методу головних компонентів, тому що ці некорельовані лінійні комбінації вихідних змінних в (13.8) та (13.9) 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 й відповідним їм компонентам. Оскільки великомасштабні процеси характеризуються великою дисперсією, то справедливе припущення, що саме вони відображені у перших компонентах.

Якщо розташувати власні числа матриці в убутному порядку, то перше власне число буде являти собою величину дисперсії, яка 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 компонентів.

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

$$S = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_s}, \qquad (14.5)$$

де чисельник дорівнює сумі дисперсій, яка припадає на p перших головних компонент, а знаменник дорівнює сумарній дисперсії поля. Це дозволяє говорити, що ***відношення (14.5) показує внесок*** p ***перших компонент у загальну дисперсію вихідних даних***.

Задаючись значенням S (наприклад, $S = 0{,}70\text{-}0{,}80$), можна встановити число перших компонент, які варто ураховувати, щоб скоротити обсяг вхідної інформації й зберегти при цьому її основний зміст. ***Скорочення вхідної інформації при збереженні її смислу має назву "стиску інформації".*** При використанні методу головних компонент

вхідна інформація не тільки заміщається малим числом статистичних функцій, але й зберігає в цих функціях їх фізичний зміст. Отримані в результаті розкладання функції є відображенням реальних фізичних процесів, які обумовлюють просторово-часовий розподіл досліджуваних гідрометеорологічних величин.

Коли розкладання за ПОФ виконується в базисі власних векторів вихідного центрованого поля ΔX_i, то розглядається вираз (13.7) $U' \cdot \varphi_i = Z_i$ буде мати вигляд

$$W' \Delta X_i = Z_i, \tag{14.6}$$

де W – матриця власних векторів матриці коваріацій.

У такому випадку

$$\sum_{p=1}^{m} \sigma_{Zp}^2 = \sum_{p=1}^{m} \lambda_p = t_R K_x = \sum_{p=1}^{m} \sigma_{Xp}^2. \tag{14.7}$$

Сума власних значень матриці є, як це було показано раніше, дисперсіями головних компонентів. 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 $\lambda_P = \sigma_{Zp}^2$ є першим власним значенням. Сума дисперсій головних компонентів матриці коваріацій дорівнює сумі дисперсій вихідних рядів, тобто

$$\sum_{p=1}^{m} \sigma_{Zp}^2 = \sum_{j=1}^{m} \sigma_{Xj}^2 \tag{14.8}$$

Таке подання дозволяє більш наочно зрозуміти суть методу головних компонентів, 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, те справедливе допущення, що саме вони відображені у перших компонентах.

Якщо розташувати власні числа матриці в убутному 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, те розділивши кожне власне число на m або $\sum_{j=1}^{m}\lambda_j$, можна одержати частку від сумарної дисперсії, що описується кожною з розглянутих компонентів. Частку істотної інформації із всієї сумарної інформації про поле оцінюється за допомогою співвідношення

$$S = \frac{\sum_{k=1}^{p}\lambda_k}{\sum_{s=1}^{m}\lambda_s}, \qquad (14.9)$$

де чисельник дорівнює сумі дисперсій, що приходиться 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 ураховувати, щоб скоротити обсяг вхідної інформації й зберегти при цьому її основний зміст. ***Скорочення вхідної інформації при збереженні її смислу має назву "стиску інформації".*** При використанні методу головних компонент вхідна інформація не тільки заміщається малим числом статистичних функцій, але й зберігає в цих функціях їх фізичний зміст. Отримані в результаті розкладання функції є відображенням реальних фізичних процесів, які обумовлюють просторово-часовий розподіл досліджуваних гідрометеорологічних та гідроекологічних величин.

Наприклад, застосування методу головних компонент до полів забруднення важкими металами донних відкладень озера Houguan (*Ke Rao , Tao Tang, Xiang Zhang, Mo Wang, Jianfeng Liu, Bi Wu , Ping Wang, Yongliang Ma, 2021*) показали, що перші три компоненти описують 70,62% вихідної інформації. Внесок першої компоненти становить 40,56%, внесок другої – 17,81, третьої – 12,25%. Перша компонента інтерпретується як

гідрохімічний процес, обумовлений природними чинниками, який призводить до появи у воді озера *Cu*, *Fe*, *Zn*, *Mn* і *Cr*. Друга компонента інтерпретується як гідрохімічний процес, обумовлений антропогенним впливом, який призводить до забруднення донних відкладів такими важкими металами як *As, Pb* і *Cd*. Третя компонента дозволяє виділити ртуть як окрему складову забруднення озера, не пов'язану із першими групами.

За необхідності виключити вплив малоінформативних фізичних процесів на формування полів даних використовують математичний апарат, роблячи зворотний перехід від компонентів до значень досліджуваної величини.

Будь-який елемент матриці вихідних центрованих та нормованих значень φ_{ij} (на i-тому розглянутому об'єкті в j-тий момент часу) може бути розрахований, якщо проблему власних векторів вирішено

$$\varphi_{ij} = \sum_{k=1}^{m} U_{ki} z_{kj}\text{, при } i = \overline{1,m};\ j = \overline{1,n}, \qquad (14.10)$$

де φ_{ij} – складові j-того випадкового вектора (поля) центрованих та нормованих вихідних даних, яке підлягало розкладанню за природними ортогональними функціями;

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.

Складові вектора-рядка матриці компонентів Z $[z_{k1}\ \ z_{k2}\ \ \cdot\ \cdot\ z_{kp}\ \cdot\ \cdot\ \cdot\ z_{kn}]$ ***можуть бути представлені як функції часу і мають назву амплітудних функцій. Амплітудні функції є загальними для всіх об'єктів, вони не залежать від координат і є функціями часу***

$$z_{kj} = f(t) = z_k(t). \qquad (14.11)$$

Розв'язання задачі фільтрації вихідної інформації полягає у розрахунках досліджуваної величини на основі тільки перших компонент розкладання поля за ПОФ. Як вже відмічалося, перші компоненти містять у собі основну частку всієї вихідної інформації. Отримане зворотним розрахунком поле вихідних даних є згладженим, оскільки відображає головні особливості просторово-часового розподілу досліджуваної величини.

При розгляді тільки перших компонентів розкладання, в яких утримується основна частина інформації вихідних полів, вираз (14.10) перетворюється до виду

$$\tilde{\varphi}_{ij} = \sum_{k=1}^{p} U_{ki} z_{kj}\text{, при } i = \overline{1,m};\ j = \overline{1,n} \qquad (14.12)$$

де $\tilde{\varphi}_{ij}$ – відфільтровані або згладжені значення центрованих та нормованих величин стоку;

p – число перших компонентів.

Матрична форма запису (14.12) має вигляд

$$\tilde{\varphi}_i = U \cdot \tilde{Z}_i, \qquad (14.13)$$

де $\tilde{\varphi}_i$ – відфільтрований вектор або поле центрованих та нормованих значень вихідної величини;

U – матриця власних векторів;

$\tilde{Z}_i$ – відфільтрований вектор компонент.

Перехід до величини $\tilde{x}_{ij}$ здійснюється за виразом

$$\tilde{x}_{ij} = \bar{x}_i + \sigma_i \sum_{k=1}^{p} U_{ki} \tilde{z}_{kj}\text{; при } i = \overline{1,m};\ j = \overline{1,n} \qquad (14.14)$$

або

$$\tilde{x}_{ij} = \bar{x}_i + \sum_{k=1}^{p} w_{ki} z_{kj}\text{; при } i = \overline{1,m};\ j = \overline{1,n} \qquad (14.15)$$

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

σ_i – середнє квадратичне відхилення вихідного ряду;

φ_{ki} – вагові коефіцієнти k-тої компоненти розкладання матриці коваріацій за ПОФ.

w_{ki} – вагові коефіцієнти k-тої компоненти розкладання матриці коваріацій за ПОФ.

В задачі фільтрації інформації вихідну інформацію треба звільнити від дрібномасштабних збурень. Цієї мети можна досягти, якщо всі складові k-тої компоненти від $p+1$ до m прирівняти до нуля, тоді k-а компонента представляється у вигляді

$$\tilde{Z}_j = \begin{pmatrix} Z_{1j} \\ Z_2 j \\ Z_3 j \\ . \\ . \\ Z_{pj} \\ 0 \\ . \\ . \\ 0 \end{pmatrix} \qquad (14.16)$$

Щоб отримати матрицю «фільтрованих» вихідних даних, представлених у центрованому виді, необхідно виконати скалярне множення матриць за виразом наступного виду

$$W\tilde{Z} = \Delta\tilde{X}, \qquad (14.17)$$

або

$$U\tilde{Z} = \tilde{\varphi}\ , \qquad (14.18)$$

де $\tilde{Z}$ – матриця, у якій виконано фільтрацію компонент згідно із (14.16);

$\Delta\widetilde{X}$ – матриця згладжених даних, представлених у вигляді центрованих величин;

$\widetilde{\varphi}$ – матриця згладжених даних, представлених у вигляді центрованих та нормованих величин.

Перехід до згладженої величини $\widetilde{x}_{ij}$ виконується за виразом

$$\widetilde{x}_{ij} = \bar{x}_i + \Delta\widetilde{x}_{ij}, \qquad (14.19)$$

де $\widetilde{x}_{ij}$ – згладжене j-те значення з ряду i;

$\bar{x}$ – середня арифметична величина ряду i;

$\Delta\widetilde{x}_{ij}$ – j-те центроване згладжене значення з ряду i отримане шляхом фільтрації.

Стиснення інформації у методі головних компонент дозволяє через складові базисних функцій проводити класифікацію гідрохімічних показників або показників якості води.

У гідроекологічних розрахунках часто як предиктори використовуються характеристики припливу прісних вод у вигляді стоку річок або даних про опади. За наявності декількох джерел надходження прісних вод можна використовувати амплітудні функції першої чи другої головних компонент, які відображають основні закономірностей коливань у часі досліджуваних характеристик.

Метод головних компонент часто використовується як метод стиснення даних у мультиспектральних і гіперспектральних зображеннях з метою статистичної максимізації кількості інформації з вихідного *n*-вимірного зображення на набагато меншу кількість нових компонентів. Цей метод перетворює три вихідні смуги зображення (синій, зелений і червоний) у статистично некорельовані смуги. Оскільки кореляція спектральних значень у видимому діапазоні спектра зазвичай дуже висока, це може призвести до зовсім іншого, барвистого зображення сцени. Метод головних компонент дозволяє переробити одноманітні за кольором рисунки на вражаючі варіації кольорів на нібито однорідному полі.

Контрольні запитання

1. Які задачі можуть бути вирішені на основі методу головних компонент.
2. Як визначається кількість компонент, які у достатній мірі описують вхідну інформацію.
3. У чому полягає розв’язання задачі “стискання” вхідної інформації.
4. У чому полягає розв’язання задачі “фільтрації” вхідної інформації.
5. Якими можуть бути власні числа матриць кореляцій чи коваріацій.

СПИСОК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ

Основна

1. Лобода Н.С., Гопченко Є.Д. Стохастичні моделі у гідрологічних розрахунках: навч. пос. Одеса: Екологія, 2006. 200 с.

2.Лобода Н.С. Методи статистичного аналізу у гідрологічних розрахунках і прогнозах: навч. пос. Одеса: Екологія, 2010. 184 с.

3. Лобода Н.С. Методи просторового узагальнення гідрологічної інформації: консп. лекцій Одеса: Екологія, 2008. 86 с.

Додаткова

4. Loboda, N. & Daus, M. (2021) Development of a method of assessment of ecological risk of surface water pollution by nitrogen compounds. *Eastern-European Journal of Enterprise Technologies.* Vol.5 №10 (113): Ecology, P.15-25. ISSN 1729-3774. https://doi.org/10.15587/1729-4061.2021.243058

5. Гидрохимическое картирование с применением вероятностно-статистических методов. / под общ. Ред. В.И. Пеляшенко. Киев: Вища школа, Головное издательство, 1979. 100 с.

6. Крицкий С.Н., Менкель М.Ф. Гидрологические основы управления водохозяйственными системами. Москва: Наука, 1982. 271 с.

7. Лобода Н.С. Расчеты и обобщения характеристик годового стока рек Украины в условиях антропогенного влияния: монография. Одесса: Экология, 2005. 208 с.

8. Осадчий В.І. Процеси формування хімічного складу поверхневих вод. Київ: Ніка-Центр, 2013. 240 с.

9. Сніжко С.І. Оцінка та прогнозування якості природних вод: підручник. Київ: Ніка-Центр, 2001. 204 с.

10. Хільчевський В.К., Осадчий В.І., Курило С.М. Основи гідрохімії. Київ: Ніка-Центр, 2012. 312с.

11. Loboda N.S., Pilipyuk.V.V Hydrochemical composition of Psyol and Vorskla river waters under conditions of antropogenic influence. *European Applied Sciences*. 2014. № 7. Pp. 57-60.

12. Loboda N.S., Pilipyuk V.V. Evaluation of ability for natural purification of the Psyol and Vorskla rivers. *International Journal of Research In Earth & Environmental Sciences*. 2015. №1. Pp. 28-32.

13. Наукові дослідження гідроекологічного режиму і стану Куяльницького лиману та морської води з Одеської затоки: науково-дослідні роботи з гідрологічного, гідрохімічного, гідробіологічного та медико-біологічного обстеження стану Куяльницького лиману та морської води з Одеської затоки. Звіт з НДР остаточний. ДР № 0116U007903, 2016 0121U113773, 2021. 128 с.

14. Школьний Є.П., Лоєва І.Д., Гончарова Л.Д. Обробка та аналіз гідрометеорологічної інформації: підручник, Київ: Міносвіти України, 1999. 538 с.

15. Мельник С. В, Лобода Н.С. Динамика наносов верхнего и среднего Днестра в условиях антропогенной нагрузки и изменения климата: Монография. Одесский государственный экологический университет. Одесса: ТЕС, 2019. 296 с.

16. Maria Triassi, Antonio Nardone, Maria Cristina Giovinetti, Elvira De Rosa, Silvia Canzanella, Pasquale Sarnacchiaro, Paolo Montuori, *Department of Public Health, University "Federico II", Via Sergio Pansini* no. 5, 80131 Naples, Italy Department Ecological risk and estimates of organophosphate pesticides loads into the Central Mediterranean Sea from Volturno River, the river of the "Land of Fires" area, southern Italy Science of the Total Environment 678. 2019, Pp.741-754.

17. Ke Rao, Tao Tang, Xiang Zhang, Mo Wang, Jianfeng Liu, Bi Wu, Ping Wang, Yongliang Ma Spatial-temporal dynamics, ecological risk assessment, source identification and interactions with internal nutrients release of heavy metals in surface sediments from a large Chinese shallow lake. *Chemosphere 282* (2021) 131041. Pp.1-9.

Навчальне електронне видання

ЛОБОДА Наталія Степанівна,
КАТИНСЬКА Ірина Вікторівна

МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ПРИ ВИРІШЕННІ ГІДРОЕКОЛОГІЧНИХ ЗАДАЧ

Конспект лекцій

Видавець і виготовлювач
Одеський державний екологічний університет
вул. Львівська, 15, м. Одеса, 65016
тел./факс: (0482) 32-67-35
E-mail: info@odeku.edu.ua
Свідоцтво суб'єкта видавничої справи
ДК № 5242 від 08.11.2016