Лобода Н.С.

МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ПРИ ВИРІШЕННІ ГІДРОЕКОЛОГІЧНИХ ЗАДАЧ

Конспект лекцій

Одеса
«ОДЕКУ»
2017

ЗМІСТ

ЗМІСТОВНИЙ МОДУЛЬ 2

1. Загальні уявлення про теорію випадкових процесів

1.1Поняття про випадкову функцію

На практиці доводиться оперувати величинами, які змінюються у процесі проведення випробувань. Такі величини одержали назву випадкових функцій. Вивченням подібних випадкових явищ, в яких випадковість набуває форми процесу, займається спеціалізованийрозділ теорії ймовірностей – теорія випадкових функцій або стохастичних процесів.

Випадкова функція– це функція, значення якої встановлюються за допомогою випробувань і можуть бути різними в залежності від ходу випробувань.***Випадковою функцією називається функція, яка в разі випробувань може набрати того чи іншого конкретного вигляду, наперед невідомо, якого саме. Конкретний вигляд, якогонабуває випадкова функція в результаті випробувань, називається реалізацією випадкової функції.***

Зустрічаються випадкові функції, які залежать не від одного аргументу, а від декількох. У гідрологічних розрахунках найчастіше розглядаються функції тільки одного аргументу. 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 є випадковими величинами. У загальному випадку випадкова функція складається з системи реалізацій*** $x_i(t)$ $(i=\overline{1,n})$***, які відбивають результати окремих експериментів. Випадкові функції при*** $t=t_j$ ***приїх перетині 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.

Сімейство реалізацій функції, яка відповідає ймовірнісному процесу з безперервною зміною функції в залежності від аргументу,наведене на рис.1.

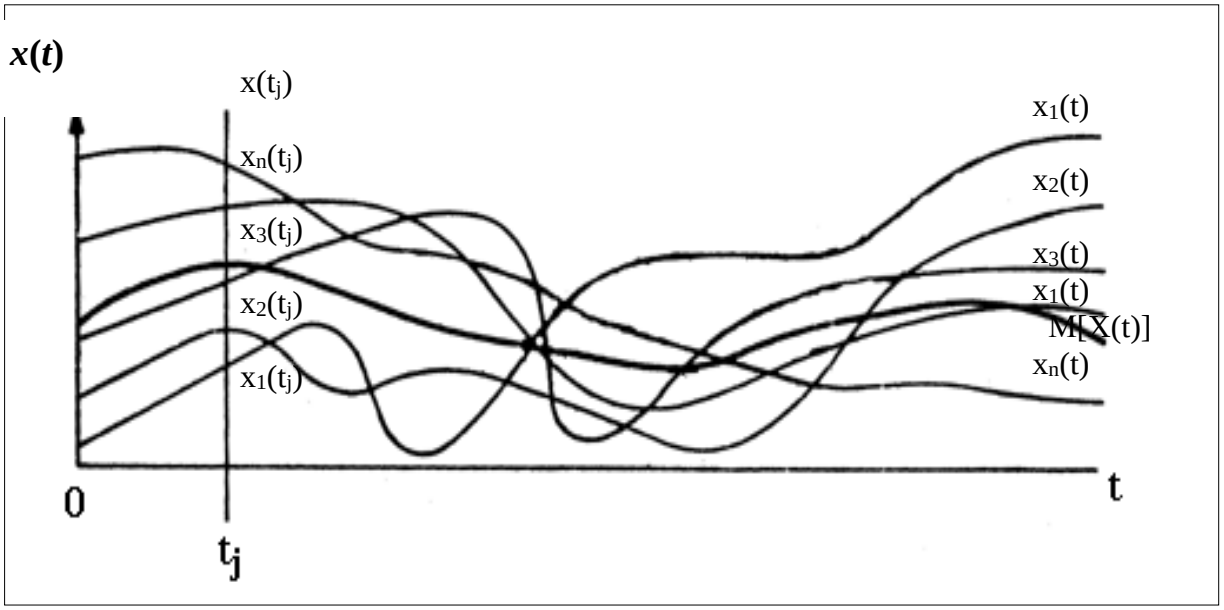


Рисунок 1.1 - Сімейство реалізацій випадкової функції

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

1.2 Закон розподілу випадкової функції

Закон розподілу однієї випадкової величини є функція одного аргументу, закон розподілу системи двох величин є функцією двох аргументів і т.п. Закон розподілу випадкової функції може представляти собою функцію безлічі аргументів, яку чисто формально можна записати в символічній формі.

Якщо зафіксувати деяке значення аргументу t_j*, то при заданому значенні аргументу* t_j *(рис.1.1) 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 суміщує 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.

Розглянемо випадкову величину $X(t)$ - перетин випадкової функції в момент часу t. Ця випадкова величина має закон розподілу, який у загальному випадку залежить від t і позначається як $f(x,t)$. Функція $f(x,t)$ називається одновимірним законом розподілу випадкової функції $X(t)$. Функція $f(x,t)$ не є повною вичерпною характеристикою випадкової функції

$X(t)$. Дійсно, ця функція характеризує тільки закон розподілу $X(t)$ для заданого, хоча й довільного t; вона не відповідає на питання про взаємозв'язок випадкових величин $X(t)$ при різних t. З цього погляду більш повною характеристикою випадкової функції $X(t)$ є так званий двовимірний закон розподілу

$$f(x_1, x_2; t_1, t_2), \qquad (1.1)$$

який описує систему двох випадкових величин $X(t_1)$, $X(t_2)$, тобто є законом розподілу для двох довільно взятих перетинів випадкової функції $X(t)$. Однак і ця характеристика не є вичерпною, ще більш повним описом випадкової функції був би тривимірний закон:

$$f(x_1, x_2, x_3; t_1, t_2, t_3). \qquad (1.2)$$

Теоретично можна необмежено збільшувати число аргументів, але оперувати ними вкрай незручно. У зв'язку з цим на практиці розглядають не закони розподілу, а найпростіші характеристики випадкових функцій, аналогічні числовим характеристикам випадкових величин.

1.3 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, ***коваріаційна*** $K_{xx}(t_i, t_j)$ та ***кореляційна*** $r_{xx}(t_i, t_j)$ ***функції***, що 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:

$$m_X(t) = M[X(t)]. \qquad (1.3)$$

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.1.1 більш тонкими лініями показані реалізації випадкової функції, а жирною лінією – її математичне сподівання.

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

$$D_X(t) = D[X(t)]. \qquad (1.4)$$

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

$$\sigma_x(t) = \sigma[X(t)] = \sqrt{D[X(t)]}. \qquad (1.5)$$

Досить часто користуються центрованою характеристикою випадкової функції

$$\delta[X(t)] = X(t) - M[X(t)]. \qquad (1.6)$$

Очевидно, що

$$M\delta[X(t)] = 0. \qquad (1.7)$$

Нормована випадкова функція представляється у вигляді

$$X_0(t) = \delta[X(t)] / \sigma[X(t)], \qquad (1.8)$$

для якої справедливі рівняння

$$M[X_0(t)] = 0; \quad D[X_0(t)] = 1. \qquad (1.9)$$

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. 1.2 показані реалізації

випадкових функцій $X(t),Y(t),Z(t)$, що мають однакові математичні сподівання, але різні дисперсії.

Математичне сподівання і дисперсія визначають тільки смугу можливих реалізацій випадкової функції, але не поводження реалізацій усередині такої смуги.

На рис. 1.Знаведені сімейства реалізацій (а) випадкових функцій, скедастичні криві (б), що показують зміну дисперсій, і кореляційні еліпси (в), які характеризують взаємний зв'язок випадкових ординат по перетинах t_1 і t_2. Зв'язок між випадковими величинами $X(t_i)$ і $X(t_j)$ можна виразити за допомогою коваріаційногомоменту чи коефіцієнта кореляції.Ступінь залежності між перетинами випадкової функції у різних аргументах t характеризується коваріаційною або кореляційною функціями.

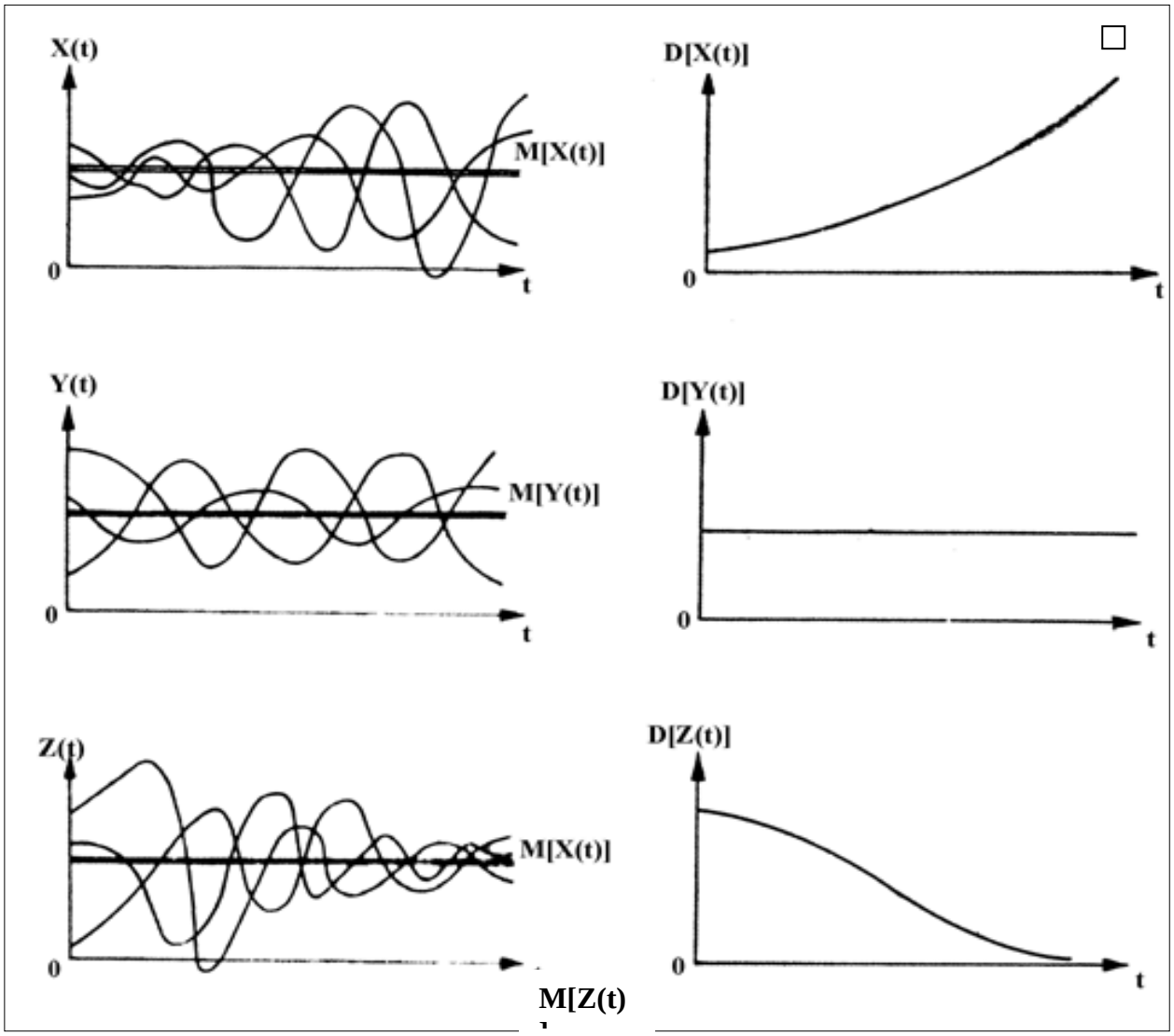


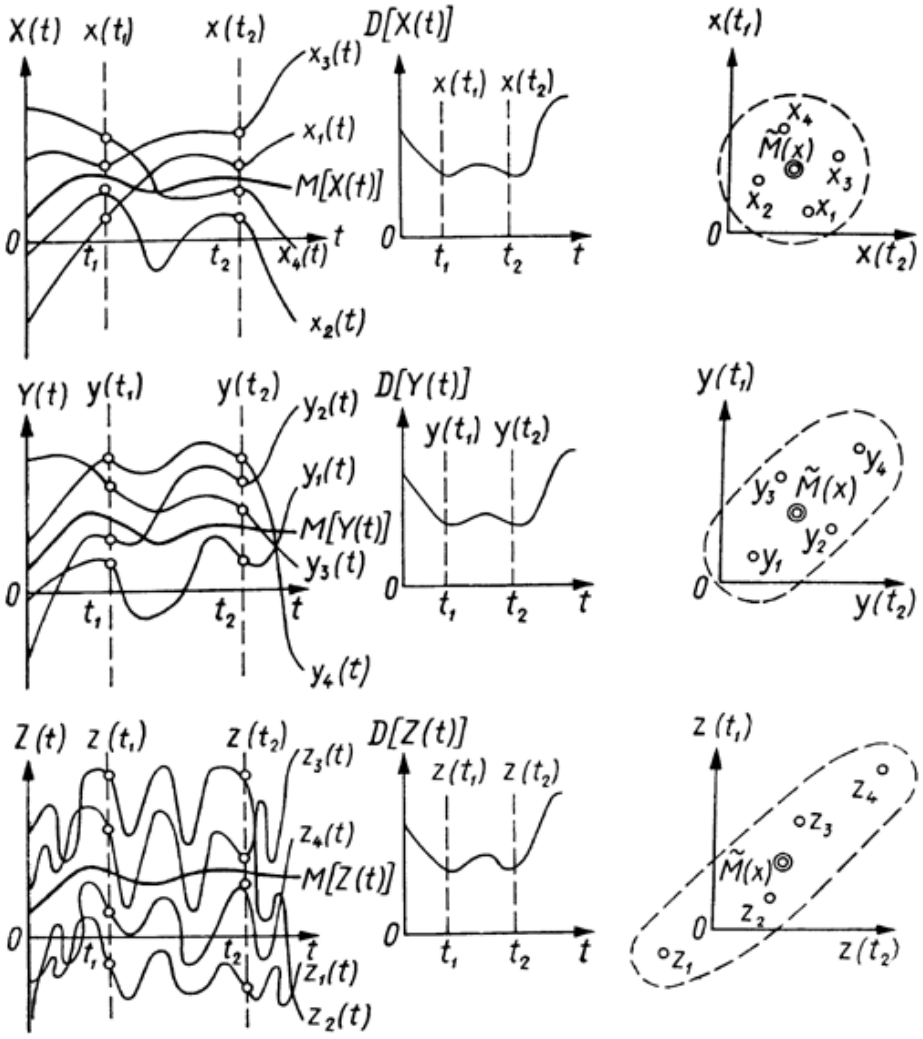
Рисунок 1.2- Зміна дисперсії у часі для різних випадкових процесів (Н.Н. Кузьменко, Ю.В. Поліщук, Л.А. Шаповалова, 1958)

Коваріаційним моментом випадкової функції 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 коваріаційномумоментуміж випадковими величинами у двох перетинах*

$$K_x = K\left[X(t_i), X(t_j)\right] = M\left\{\delta\left[X(t_i)\right] \cdot \delta\left[X(t_j)\right]\right\}. \qquad (1.9)$$

Розглянемо два випадкових процеси $X_1(t)$ і $X_2(t)$ з однаковими математичними сподіваннями та дисперсіями (рис. 1.4, 1.5).

Коваріаційна функція випадкової функції $X_1(t)$повільно убуваєв міру збільшення проміжку t, t'. Навпаки, коваріаційна функція випадкової функції $X_2(t)$убуває швидко. Ці особливості випадкових процесів характеризуються саме коваріаційними моментами.



а) б) в)

Рисунок 1.3 - Випадкові функції, їх дисперсії та еліпси розсіювання[56]

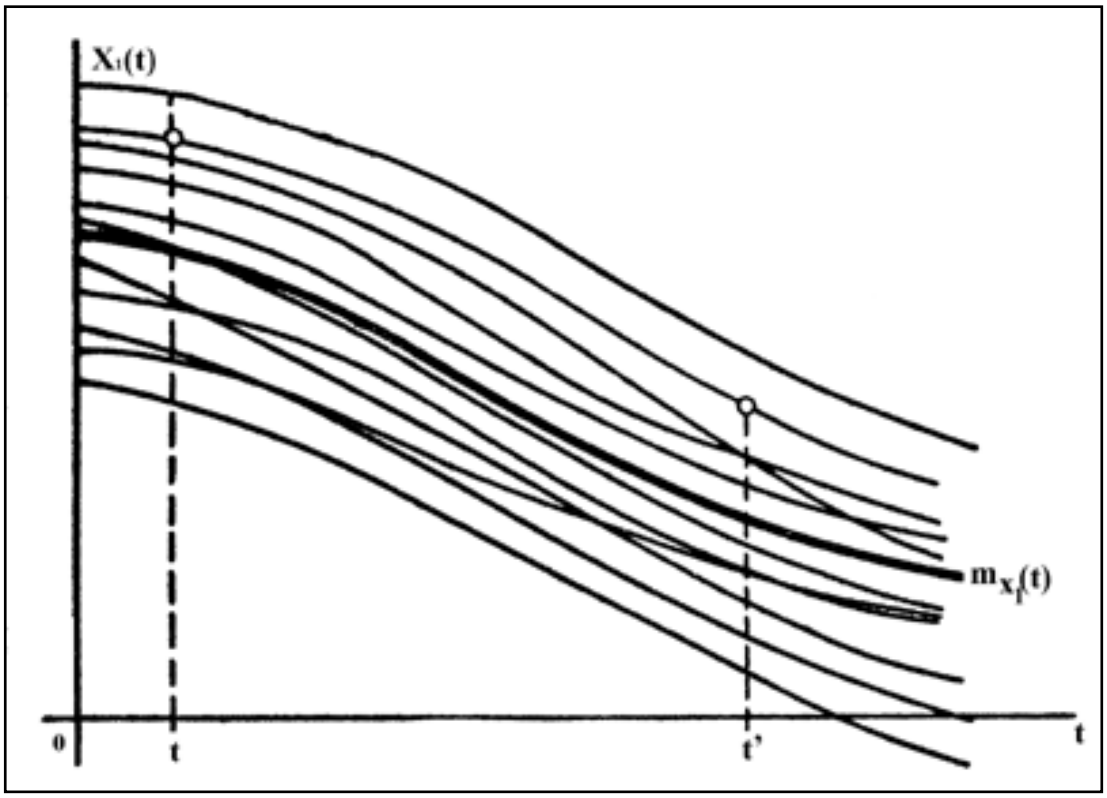


Рисунок 1.4 - Реалізації випадкової функції $X_1(t)$

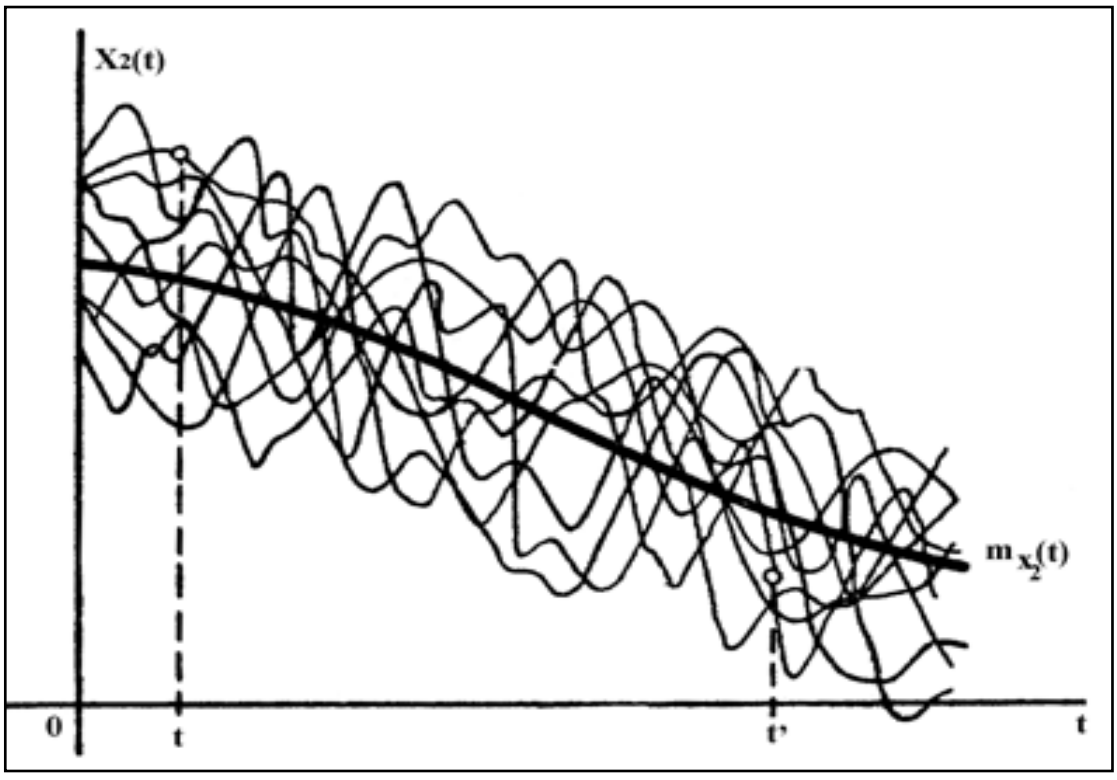


Рисунок 1.5 - Реалізації випадкової функції $X_2(t)$

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

$$r_x = r\left[X(t_i), X(t_j)\right] = \frac{K\left[X(t_i), X(t_j)\right]}{\sigma(t_i)\cdot\sigma(t_j)}. \qquad (1.10)$$

Кореляційну функцію випадкового процесу часто називають ***автокореляційною***, що означає кореляцію процесу із самим собою.
При $t_i = t_j = t$ одержимо

$$K\left[X(t), X(t)\right] = M\left[\delta^2(t)\right] = D_X(t)\ ; \qquad (1.11)$$

$$r_x = r\left[X(t\ ), X(t\ )\right] = \frac{K\left[X(t\ ), X(t\ )\right]}{\sigma(t)\cdot\sigma(t)} = \frac{D_X(t)}{D_X(t)} = 1, \qquad (1.12)$$

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, при якій ці величини розглядаються, отже, коваріаційна функція симетрична відносно своїх аргументів, тобто не змінюється при зміні аргументів місцями

$$K_x(t,t') = K_x(t',t). \qquad (1.13)$$

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 $K\left[X(t_i), X(t_j)\right] = 0$ при будь-якому значенні іншого аргументу.

Коваріаційна (кореляційна) функція симетрична щодо своїх аргументів. Вона зображується в тривимірній системі прямокутних координат t_i, t_j і $K\left[X(t_i), X(t_j)\right]$ у вигляді поверхні, симетричної щодо вертикальної площини Q, яка проходить через бісектрису кута $t_i O t_j$ (рис.1.6).

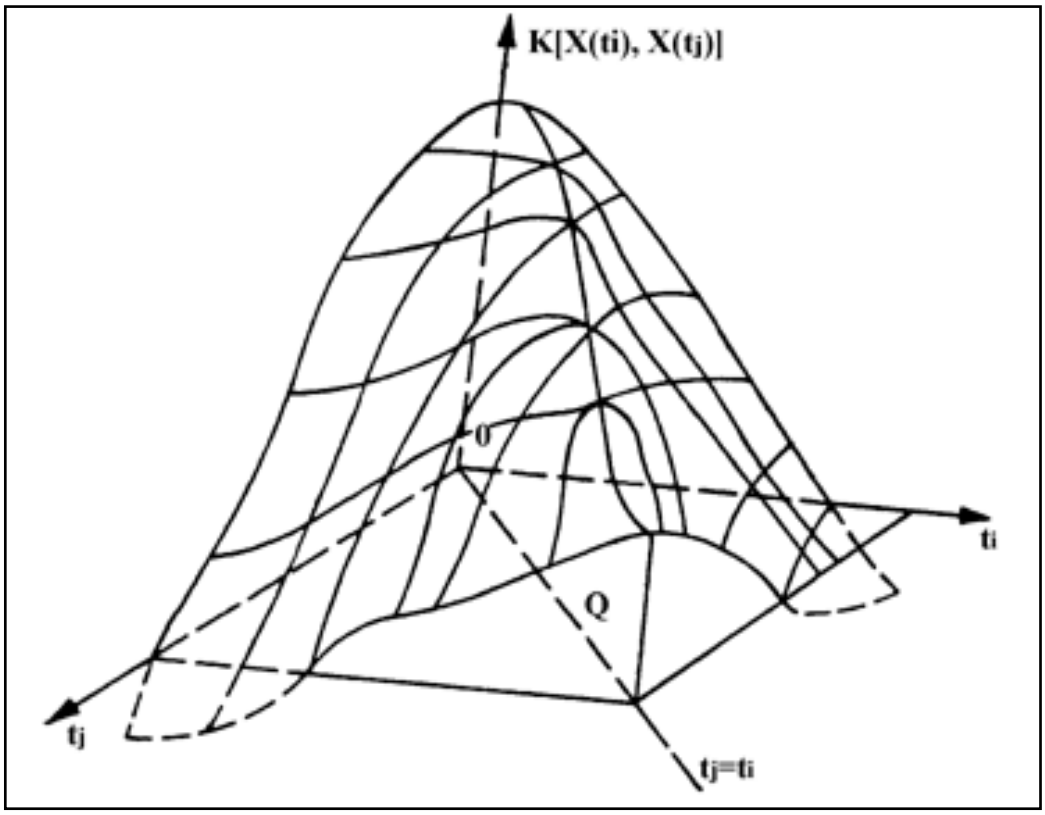


Рис. 1.6 - Коваріаційна функція у вигляді поверхні

Нехай над випадковою функцією $X(t)$ зроблено n незалежних випробувань (спостережень), у результаті отримано n 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

$$\bar{x}(t_i) = \frac{\sum_{j=1}^{m} x(t_i)}{m}; \qquad (1.14)$$

оцінкою дисперсії -

$$\hat{\sigma}_x^2(t_i) = S_x^2(t_i) = \frac{1}{m}\sum_{j=1}^{m}\{\delta[X(t_i)]\} = \frac{\sum_{j=1}^{m}\left[x(t_i) - \bar{x}(t_i)\right]^2}{m-1}. \qquad (1.15)$$

Для коваріаційнихі кореляційних функцій розрахунок виконується за такими формулами:

$$\hat{K}_x(t_i,t_k)=\frac{\sum_{j=1}^{m}\left[x(t_i)-\bar{x}(t_i)\right]\cdot\left[x(t_k)-\bar{x}(t_k)\right]}{(m-1)}; \qquad (1.16)$$

$$\hat{r}_x(t_i,t_k)=\frac{\sum_{j=1}^{m}\left[x(t_i)-\bar{x}(t_i)\right]\cdot\left[x(t_k)-\bar{x}(t_k)\right]}{(m-1)S_x(t_i)S_x(t_k)}. \qquad (1.17)$$

Після того, як характеристики обчислені, можна побудувати залежності середніх арифметичних $\bar{x}$ і дисперсії S_x^2 від часу. Функції двох аргументів $\hat{r}_x(t_i,t_k)$ і $\hat{K}_x(t_i,t_k)$також відтворюються в прямокутній сітці точок. За необхідністю всі ці функції апроксимуються аналітичними виразами.

1.4Стаціонарність випадкових процесів

Випадкові процеси, що протікають у часі приблизно однорідно, називають стаціонарними. Вони мають вигляд безперервних коливань відносно деякого середнього значення. Реалізації таких процесів знаходяться ніби-то у стані статистичної рівноваги. З часом середня амплітуда і характер коливань істотно не змінюються.

У строгому розумінні слова, випадковий процес називають стаціонарним, якщо всі його закони розподілу не змінюються при додаванні до всіх значень аргументу одного й того ж числа, тобто якщо усі вони залежать тільки від взаємного розташування значень аргументу одного й того ж числа або якщо усі вони залежать тільки від взаємного розташування значень аргументу, але не від самих цих значень.

Таким чином, випадковий процес $X(t)$ є стаціонарним, якщо при будь-якому n і будь-якому t_0 будуть виконуватися рівності

$$f_n(x_1,x_2,...,x_n;t_1,t_2,...,t_n)=f_n(x_1,x_2,...,x_n;t_1+t_0,t_2+t_0,...,t_n+t_0). \qquad (1.18)$$

Отже, щільності розподілу інваріантні стосовно зсуву початку відліку аргументу t.

Зокрема, для одновимірної щільності розподілу $f_1(x,t)$ стаціонарного випадкового процесу, беручи $t_0=-t$, отримуємо

$$f_1(x,t)=f_1(x;t-t)=f_1(x;0)=f_1(x). \qquad (1.19)$$

Таким чином, одновимірна щільність розподілу стаціонарного процесу не залежить від t, вона є однією і тією ж для усіх перетинів випадкового

процесу. Двовимірна щільність розподілу при $t_0 = -t_1$ представляється у вигляді

$$f_2(x_1, x_2; t_1, t_2) = f_2(x_1, x_2; 0; t_2 - t_1) = f_2(x_1, x_2; t_2 - t_1) = f_2(x_1, x_2; \tau), \quad (1.20)$$

тобто двовимірна щільність розподілу залежить не від двох аргументів t_1, t_2, а тільки від одного аргументу – їхньої різниці $\tau = t_2 - t_1$. Звідси для стаціонарного випадкового процесу, згідно з (1.19), одержуємо

$$m_x(t) = \int_{-\infty}^{\infty} x f_1(x) dx = m_x = const, \quad (1.21)$$

тобто математичне сподівання стаціонарного випадкового процесу не залежить від аргументу t і є постійною величиною.

Згідно з (1.20) вираз для коваріаційногомоменту буде такий

$$K_x(t_1, t_2) = \int_{-\infty}^{\infty}\int (x_1 - m_x)(x_2 - m_x) \cdot f_2(x_1, x_2; \tau) dx_1 dx_2 = K_x(\tau). \quad (1.22)$$

Таким чином, коваріаційна функція стаціонарного випадкового процесу є функцією тільки одного аргументу $\tau = t_2 - t_1$. Умови (1.21) і (1.22) виконуються для будь-якого стаціонарного процесу. Однак вони не є достатніми для визнання стаціонарності, тобто їхнє виконання не гарантує виконання умов (1.18). Таке розуміння стаціонарності випадкових функцій називають стаціонарністю у вузькому розумінні.

У прикладній теорії стаціонарних випадкових функцій користуються поняттям“стаціонарність” у широкому розумінні(Д.І. Казакевич, 1971). За звичай припускають, що стаціонарний випадковий процес повинен задовольняти три умови:

$$M[X(t_s)] = const, D[X(t_s)] = const; K_X[X(t_s), (t_k)] = K(t_i - t_j) = K_\tau. \quad (1.23)$$

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.

Однак не всі ці умови рівноцінні. Не можна вважати сталість математичного сподівання істотною вимогою до стаціонарної випадкової функції. При переході до центрованої випадкової функції відразу ж виконується рівність $M\{\delta[X(t)]\} = 0$, тобто перша умова з (1.23). Отже у випадку змінного математичного сподівання $M[X(t)]$ досить перейти до центрування для того, щоб виконати одну з вимог до стаціонарного випадкового процесу.

Виконання другої умови з (1.23) 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 (1.22). Покладемо $t_s = t;\ t_k = t + \tau$. Якщо $\tau = 0$, то згідно (1.13)

$$D_x(t) = K_X\left[X(t), X(t+0)\right] = K_X(\tau = 0) = const\ . \qquad (1.24)$$

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

$$K_x(\tau) = K_x(-\tau); \qquad (1.25)$$

$$r_x(\tau) = r_x(-\tau) \qquad (1.26)$$

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.

Найчастіше зустрічаються стаціонарні випадкові процеси, коваріаційні функції яких апроксимуються функціями таких типів (рис. 1.7)

$$K_X(\tau) = \sigma^2 e^{-a|\tau|}, \quad a > 0; \qquad (1.27)$$

$$K_X(\tau) = \sigma^2 e^{-a\tau^2}, \quad a > 0; \qquad (1.28)$$

$$K_X(\tau) = \sigma^2 e^{-a|\tau|} \cdot cos\,\beta\tau, \quad a > 0; \qquad (1.29)$$

$$K_X(\tau) = \sigma^2 e^{-a\tau^2} \cdot cos\ \beta\tau, \quad a > 0; \qquad (1.30)$$

$$K_X(\tau) = \sigma^2 e^{-a|\tau|}\left(cos\ \beta\tau + \frac{\alpha}{\beta} sin\ \beta|\tau|\right), \quad a > 0, \beta > 0; \qquad (1.31)$$

$$K_X(\tau) = \begin{cases} \sigma\left(1 - \dfrac{\tau}{\tau_0}\right) & \text{ïðè } |\tau| \le \tau_0 \\ 0 & \text{ïðè } |\tau| > \tau_0 \end{cases}. \qquad (1.32)$$

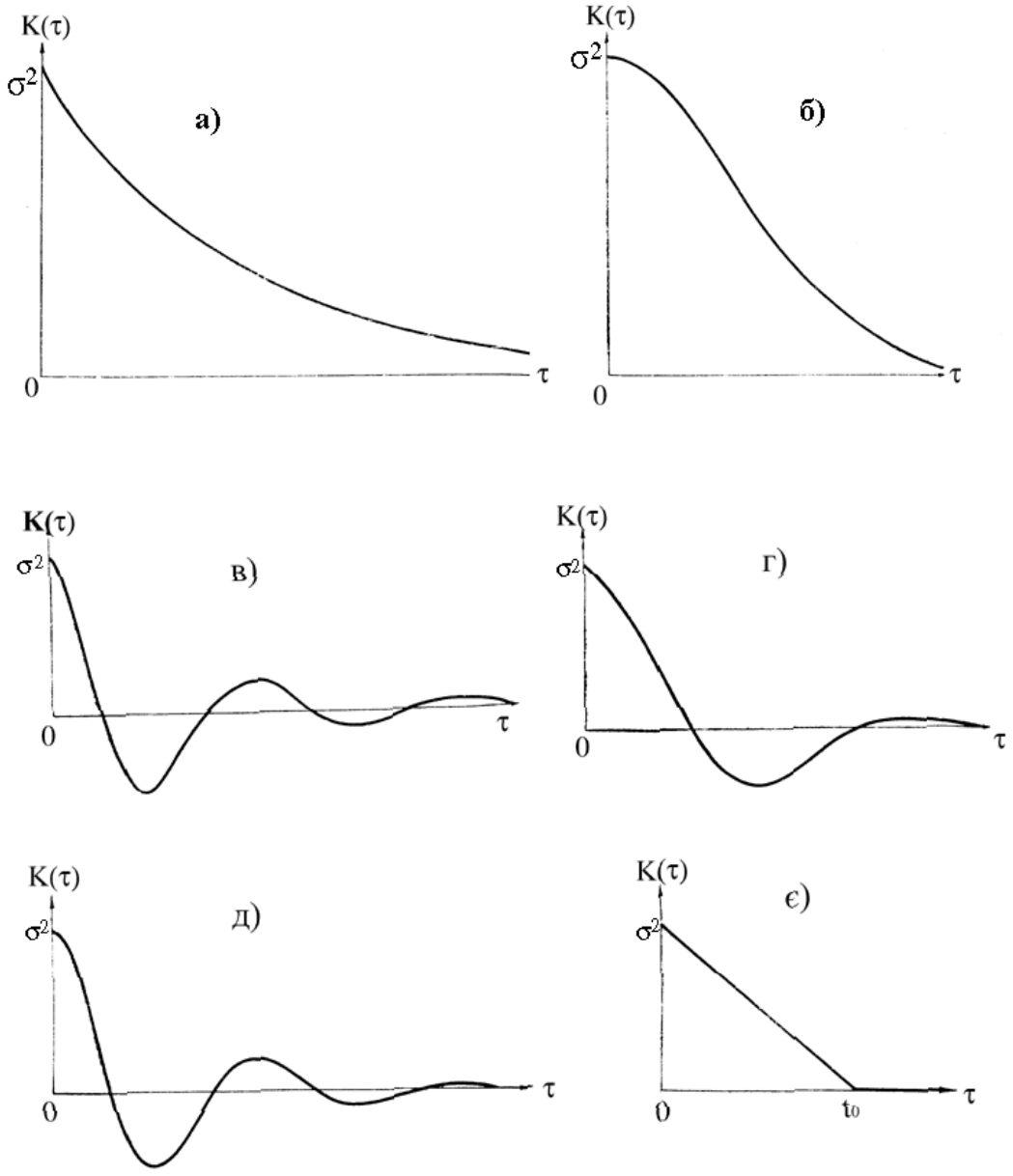


Рисунок 1.7 - Приклади коваріаційних функцій стаціонарних процесів (Д.І. Казакевич, 1971)

На рис. 1.7 наведені графіки коваріаційних функцій тільки для $\tau > 0$, у силу парності цих функцій при $\tau < 0$ їм будуть відповідати криві, симетричні щодо осі ординат. З рис. 1.8 видно, що значення коваріаційних функцій убувають із зростанням τ, тобто кореляційний зв'язок між різними перетинами випадкових функцій убуває зі збільшенням інтервалу часу між ними.

Одержання від'ємних значень $K_X(\tau)$ означає прояв від'ємних зв'язків між перетинами випадкової функції. Для всіх наведених випадків коваріаційна, а отже, і кореляційна функція прагне до нуля при прагненні τ

до нескінченності. Ця властивість виконується практично для всіх гідрометеорологічних рядів.

Коли в структурі випадкової функції як постійний доданок, є деяка випадкова величина, при $\tau \to \infty$ коваріаційна функція буде прагнути не до 0, а до дисперсії D цієї величини (рис.1.8). Як правило, вигляд автокореляційної функції, наведений на рис.1.8, характерний для випадкових процесів, які знаходяться під впливом монотонно зростаючих антропогенних перетворень (А.В. Рождественський, О.І. Чеботарьов, 1974).

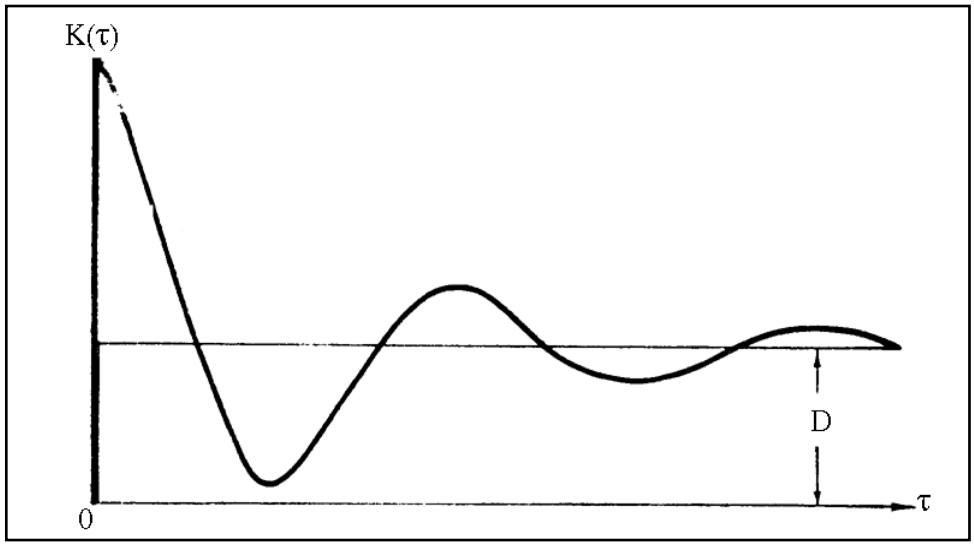


Рисунок 1.8 - Кореляційна функція при наявності постійного впливу на випадковий процес

1.5Ергодичність стаціонарних випадкових процесів

Переважній більшості стаціонарних випадкових функцій властива ергодичність. Це відноситься до випадкових функцій, що характеризуються стаціонарністю у вузькому і широкому розумінні. *Властивість ергодичності полягає у тому, що кожна окрема реалізація стаціонарної випадкової функції є повноправним представникомусієї сукупності можливих реалізацій. Одна реалізація достатньої довжини може замінити при статистичній обробці всі реалізації в цілому. Статистичніхарактеристики кожної реалізації є одними й тими ж для усіх реалізацій.*

По відношенню до математичного сподівання достатньою умовою ергодичності є прямування коваріаційної та кореляційної функцій до нуля (Д.І. Казакевич, 1971).

Розглянемо дві стаціонарні випадкові функції $X_1(t)$ і $X_2(t)$, наведені на рис.1.9. Для функції $X_2(t)$ кожна із реалізацій має своє математичне сподівання, яке не співпадає із математичним сподіванням процесу у цілому,

свою дисперсію. Кожна з реалізацій випадкової функції $X_1(t)$ має одні й тіжсамі характерні ознаки: середнє значення, навколо якого коливаються реалізації; розмах коливань.

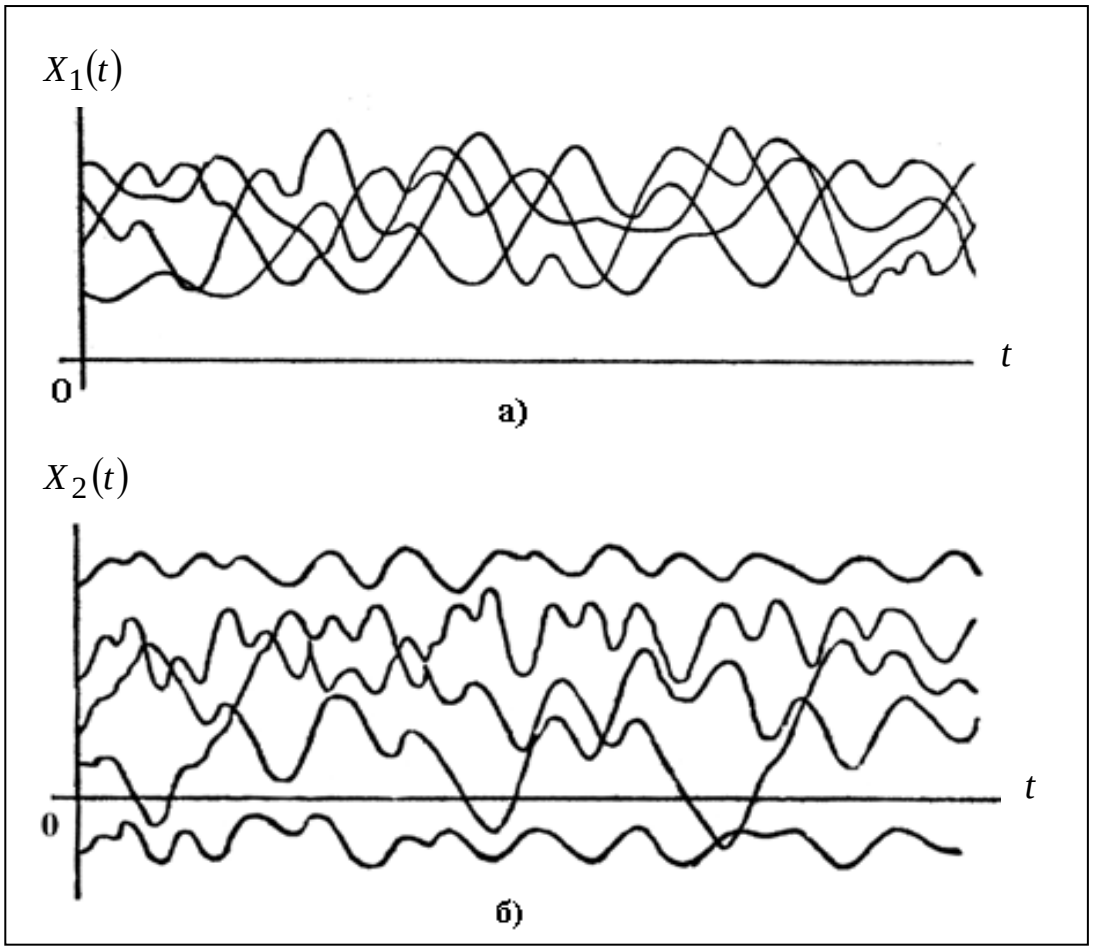


Рисунок 1.9 – Ергодичний $X_1(t)$ та неергодичнийстаціонарні $X_2(t)$ випадкові процеси

Якщо вибрати довільно одну з реалізацій функції $X_1(t)$, то при досить довгій її тривалості T на її основі можна одержати повне уявлення про випадкову функцію в цілому.

Осереднюючи значення цієї реалізації за часом, одержують оцінку математичного сподівання

$$\hat{m}_x \approx \frac{1}{T}\int_0^T x(t)dt, \qquad (1.33)$$

коваріаційноїфункції

$$\hat{K}_x(\tau) \approx \frac{1}{T-\tau}\int_0^{T-\tau}[x(t)-\hat{m}_x]\cdot[x(t+\tau)-\hat{m}_x]dt\ . \qquad (1.34)$$

Якщо розбити інтервал запису випадкової функції на n рівних частин довжиною Δt і позначити середини отриманих ділянок $t_1, t_2, ..., t_n$ (рис.1.10), то інтеграли (1.33) і (1.34) можна представити у вигляді:

$$\hat{m}_x = \frac{1}{n}\sum_{i=1}^{n} x(t_i) = \bar{x}; \quad (1.35)$$

$$\hat{K}_x(\tau) = \frac{1}{n-\tau}\sum_{i=1}^{n-\tau}\left[x(t_i) - \bar{x}\right]\cdot\left[x(t_i+\tau) - \bar{x}\right], \quad (1.36)$$

де $\tau = m\Delta t$, причому $m = 0,1,2,...$

У подальшому викладенні для простоти запису знак ^ , який означає, що мова йде про оцінку характеристики за вибірковими даними, прибирається, а t_i представляється як i, звідки отримуються вже відомі формули для визначення середньої арифметичної величини та коваріаційної функції ряду випадкових величин:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad (1.37)$$

$$K_x(\tau) = \frac{1}{n-\tau}\sum_{i=1}^{n-\tau}\left(x_i - \bar{x}\right)\left(x_{i+\tau} - \bar{x}\right). \quad (1.38)$$

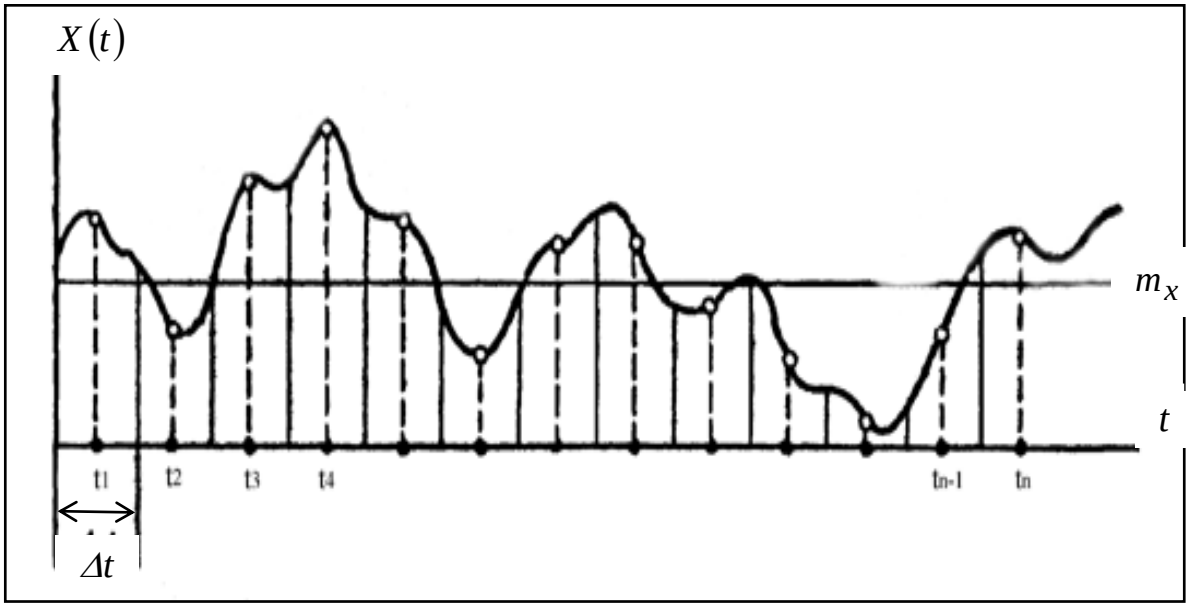


Рисунок 1.10- Запис випадкової функції, яка розбивається на n частин

Загальний вигляд функції $K_x(\tau)$ відтворюється по окремих точках. Для того, щоб математичне сподівання і коваріаційнафункція були визначені із задовільною точністю, потрібно, щоб число n було досить великим (порядку сотні, а в деяких випадках, навіть і більше). Вибір довжини елементарної ділянки Δt визначається характером зміни випадкової функції. Якщо випадкова величина змінюється порівняно повільно, ділянки Δt можна вибирати більшими, ніж у випадку, коли функція робить різкі і часті коливання. Чим більш високочастотний склад мають коливання, що утворюють випадкову функцію, тим частіше потрібно розміщувати опорні точки при обробці.

Значення величини m задаються послідовно, аж до таких m, при яких коваріаційнафункція практично дорівнює нулю чи починає робити невеликі нерегулярні коливання біля нуля.

Властивість ергодичності має велике практичне значення, оскільки при її виконанні для визначення статистичних характеристик випадкових функцій нема необхідності у великій кількості реалізацій. Це дуже важливо для гідрології суші, де ряди стоку являють собою лише одну реалізацію.

У гідрологічних розрахунках річковий стік розглядається як ергодичний процес з річним періодом. Така гіпотеза зручна й необхідна, оскільки по кожному гідрологічному створу є лише одна реалізація процесу стоку, яка представлена гідрологічними спостереженнями.

Кореляційний аналіз є одним з традиційних методів в дослідженнях зв'язків між послідовними членами рядів річного стоку. Проте значне випадкове розсіювання вибіркових оцінок коефіцієнтів кореляції при невеликій тривалості вихідних рядів і невеликих числових значеннях $\hat{r}_x(\tau)$ спричиняє природне утруднення в інтерпретації результатів розрахунку.

Ординати автокореляційної функції за вибірковими даними обчислюються з урахуванням їх зміщеності за такою формулою:

$$\hat{r}_x(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\hat{\sigma}_x^2 (n - \tau - 1)}, \qquad (1.39)$$

де n - число членів вихідного 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;

$\bar{x}, \hat{\sigma}_x$ - відповідно середнє і стандарт вихідного ряду.

При розрахунках ординат автокореляційної функції необхідно враховувати зміни середнього арифметичного і стандарту, які відбуваються за рахунок зсуву τ і втрат крайових членів ряду. При цьому (1.39) набуває вигляду

$$\hat{r}_x(\tau) = \frac{\sum_{i=1}^{n-\tau} (x_i - \overline{x_i})(x_{i+\tau} - \overline{x}_{i+\tau})}{\sigma_i \sigma_{i+\tau}(n-\tau-1)}, \qquad (1.40)$$

де $x_i, \overline{x}_{i+\tau}, \sigma_i$ та $\sigma_{i+\tau}$ - середні і стандарт відповідних відрізків вихідного ряду.

ЛЕКЦІЯ 2

«Кореляційний аналіз»

Питання:

1. Внутрирядна та міжрядна кореляції.

2. Автокореляційна функція.

3. Взаємна кореляційна функція.

4. Матриці кореляцій та коваріацій

2.1 Внутрирядна кореляція, автокореляційна функція

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.

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

Ординати автокореляційної функції за вибірковими даними обчислюються з урахуванням зміщення за такою формулою:

$$r_x(\tau) = \frac{\sum_{i=1}^{n-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\hat{\sigma}_x^2(n-\tau-1)}, \qquad (2.1)$$

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;

$\bar{x}, \hat{\sigma}_x$ - відповідно середнє і стандарт початкового ряду.

Рекомендується ураховувати зміну середнього і стандарту вибірки з ростом запізнювання автокореляційної функції. Тоді формула (2.1) набирає вигляду

$$r_x(\tau) = \frac{\sum_{i=1}^{n-\tau}(x_i - \bar{x}_i)(x_{i+\tau} - \bar{x}_{i+\tau})}{\sigma_i \sigma_{i+\tau}(n+\tau+1)}, \qquad (2.2)$$

де x_i, $\bar{x}_{i+1}$, σ_i, $\sigma_{i+\tau}$ - середні і стандарт відповідних відрізків вихідного ряду.

Навіть при відносно тривалих рядах стоку (50-60 років) недостовірні як окремі ординати автокореляційної функції, так і її контури. Тому прирозрахунках 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 $\tau = 1$ коефіцієнт кореляції між стоком суміжних років розраховується за формулою

$$r(\tau = 1) = r_{(1)} = \frac{\sum_{i=1}^{n-\tau}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n-\tau}(x_i - \bar{x})^2}. \qquad (2.3)$$

Середня квадратична похибка еміпиричного коефіцієнта автокореляції визначається за рівнянням

$$\sigma_{r(1)} = \frac{1 - r(1)^2}{\sqrt{n-1}}. \qquad (2.4)$$

Розрахункове значення коефіцієнта кореляції є значущим, якщо виконується умова

$$r(1) \geq 2\sigma_{r(1)}. \qquad (2.5)$$

За рахунок внутрішньорядних зв'язків у рядах річного стоку зменшується об'єм незалежної інформації. Для усунення впливу внутрішньорядних зв'язків між суміжними членами рядів річного стоку на точність розрахунків статистичних параметрів, вводяться поправки у розрахунки коефіцієнтів варіації та асиметрії:

$$Cv = (a_1 + a_2/n) + (a_3 + a_4/n)\tilde{C}v + (a_5 + a_6/n)\tilde{C}v^2; \qquad (2.6)$$

$$Cs = (b_1 + b_2/n) + (b_3 + b_4/n)\tilde{C}s + (b_5 + b_6/n)\tilde{C}s^2, \qquad (2.7)$$

де $a_1,....,a_6; b_1,.....,b_6$ - коефіцієнти, що визначаються по табл. 2.1 і 2.2;

$\tilde{C}v$, $\tilde{C}s$ - оцінки коефіцієнтів варіації і асиметрії, установлені без урахування внутрішньорядних зв'язків:

$$\tilde{C}v = \sqrt{\frac{\sum_{i=1}^{n}(k_i - 1)^2}{n-1}}; \qquad (2.8)$$

$$\tilde{C}s = \frac{n}{(n-1)(n-2)} \frac{\sum(k_i - 1)^3}{\tilde{C}^3 v}; \qquad (2.9)$$

де $k_i = x_i / \bar{x}$.

Таблиця 2.1 – Коефіцієнти *a* у формулі (2.6)

C_s / C_v	r(1)	a_1	a_2	a_3	a_5	a_6	a_7
1	2	3	4	5	6	7	8
2.00	0.00	0.00	0.19	0.99	-0.88	0.01	1.54
	0.30	0.00	0.22	0.99	-0.41	0.01	1.51
	0.50	0.00	0.18	0.98	0.41	0.02	1.47
3.00	0.00	0.00	0.69	0.98	-4.34	0.01	6.78
	0.30	0.00	1.15	1.02	-7.53	-0.04	12.38
	0.50	0.00	1.75	1.00	-11.79	-0.05	21.13
4.00	0.00	0.00	1.36	1.02	-9.68	-0.05	15.55
	0.30	-0.02	2.61	1.13	-19.85	-0.22	34.15
	0.50	-0.02	3.47	1.18	-29.71	-0.41	58.08

Таблиця 2.2 – Коефіцієнти *b* у формулі (2.6)

r(1)	b_1	b_2	b_3	b_4	b_5	b_6
0.00	0.03	2.00	0.92	-5.09	0.03	8.10
0.30	0.03	1.77	0.93	-3.45	0.03	8.03
0.50	0.03	1.63	0.92	-0.97	0.03	7.94

Середні квадратичні похибки вибіркових параметрів обчислюються за наступною формулою:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{\frac{1 + r(1)}{1 - r(1)}}; \qquad (2.10)$$

$$\sigma_{Cv} = \frac{Cv}{n + 4C^2 v} \sqrt{\frac{n(1 + C^2 v)}{2}} \left(1 + \frac{3Cv r(1)^2}{1 - r(1)} \right). \qquad (2.11)$$

2.2 Матриці кореляцій та коваріацій

Для вирішення чисельних задач гідроекології необхідні знання статистичної структури гідрометеорологічних полів (наприклад, поля стоку, опадів, швидкості вітру, тощо).Сукупність *m* гідрометеорологічних полів, які відносяться до визначених термінів спостереження, можна зобразити матрицею порядку *n* x *m* наступного вигляду

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{ni} & \dots & x_{nm} \end{bmatrix}. \qquad (2.12)$$

Матриця (2.12) утримує великий об’єм інформації про n об’єктів. Рядки матриці являють собою часові ряди відповідної гідрометеорологічної величини. Матричне зображення гідрометеорологічних та гідроекологічних 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.*

Для побудови матриці коваріацій необхідно спочатку знайти вектор середніх значень

$$\overline{X} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_i \\ \dots \\ \overline{x}_n \end{bmatrix}, \qquad (2.13)$$

де

$$\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \quad (i = \overline{1,m}). \qquad (2.14)$$

Якщо мова йде про поля гідрометеорологічних величин, то вектор (2.13) є осередненим полем гідрометеорологічних величин. На основі

матриці (2.12) та вектора (2.13) *визначають відповідну матрицю центрованих елементів,* для чого від елементів кожного рядка матриці (2.12) віднімають відповідне середнє арифметичне значення і отримують наступний вираз

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{i1} & \Delta x_{i2} & \dots & \Delta x_{ij} & \dots & \Delta x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{n1} & \Delta x_{n2} & \dots & \Delta x_{ni} & \dots & \Delta x_{nm} \end{bmatrix}, \qquad (2.15)$$

де

$$\Delta x_{ij} = x_{ij} - \bar{x}_i \ . \qquad (2.16)$$

Операція проведена над матрицею (2.12) називається операцією центрування. Матриця коваріацій K_x визначається за рівнянням:

$$K_x = \frac{1}{m} \Delta X \Delta X^{/}, \qquad (2.17)$$

де $\Delta X^{/}$ - транспонована матриця центрованих величин.

Доказ цього твердження можна показати на простій матриці

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \end{bmatrix}. \qquad (2.18)$$

Для матриці (2.18) рівність (2.17) в координатній формі буде мати вигляд

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

$$= \begin{bmatrix} \frac{1}{m} \sum_{j=1}^{m} \Delta x^2_{1j} & \frac{1}{m} \sum_{j=1}^{m} \Delta x_{1j} \Delta x_{2j} \\ \frac{1}{m} \sum_{j=1}^{m} \Delta x_{2j} \Delta x_{1j} & \frac{1}{m} \sum_{j=1}^{m} \Delta x^2_{2j} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & K_{12} \\ K_{21} & \sigma_2^2 \end{bmatrix}. \qquad (2.19)$$

Після обчислювань отримаємо:

$$K_x = \begin{bmatrix} \sigma_1^2 & K_{12} & \dots & K_{1j} & \dots & K_{1n} \\ K_{21} & \sigma_2^2 & \dots & K_{2j} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{i1} & K_{i2} & \dots & K_{ij} & \dots & K_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nj} & \dots & \sigma_n^2 \end{bmatrix}. \qquad (2.20)$$

Елементи матриці (2.20) розраховуються за формулами:

$$\sigma_i^2 = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}^2 , \qquad (2.21)$$

$$K_{ij} = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}\Delta x_{js} . \qquad (2.22)$$

Таким чином, із формул (2.21)-(2.22) випливає, що на головній діагоналі матриці (2.20) 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 (2.20) – відповідні коваріації.

Властивості матриці коваріації:

- *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).

Матриця коваріацій поряд з вектором математичних сподівань m_x відіграє роль параметра *щільності ймовірностей багатовимірного нормального розподілу*

$$f(x) = \frac{1}{(2\pi)^{n/2}\left|K\right|^{1/2}} \times \exp\left\{-\frac{1}{2}(X - m_x)K_x^{-1}(X - m_x)\right\} . \qquad (2.23)$$

Із матриці коваріацій можна сформувати діагональну матрицю σ середніх квадратичних відхилень. Вона має наступний вид

$$\sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix}. \qquad (2.24)$$

Обернена матриця буде мати наступний вид

$$\sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{bmatrix}. \qquad (2.25)$$

Якщо помножити ліворуч та праворуч матрицю коваріацій K_x на матрицю (2.25), отримаємо матрицю кореляцій R_x

$$R_x = \begin{bmatrix} 1 & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2j} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{i1} & r_{i2} & \dots & r_{ij} & \dots & r_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nj} & \dots & 1 \end{bmatrix}, \qquad (2.26)$$

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

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

$$r_{ij} = \frac{K_{ij}}{\sigma_i \sigma_j}. \qquad (2.27)$$

За наявності рядів спостережень довжиною m вибіркове значення коефіцієнта кореляції між двома змінними x_j та x_k розраховується наступним чином

$$r_{jk} = \frac{\sum_{i=1}^{m}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)}{(m-1)\sigma_j\sigma_k}, \qquad (2.28)$$

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 випадку є дисперсією вихідного ряду *j*.

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

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

ЛЕКЦІЯ 3

«Структурна функція»

Питання.

1. *Часові та просторові структурні функції.*
2. *Зв'язок між структурною функцією та кореляційною функцією.*

В сучасній стохастичній гідрології масштабування та пов'язане з ним визначення фрактальних розмірностей більшості гідрометеорологічних величин може бути отримано в результаті дослідження часової або просторової варіації величини, яка вивчається, в часі або просторі з кроком s. Під варіацією мається на увазі зміна досліджуваної характеристики на деякому часовому інтервалі або заданій відстані, яку частіше за все представляють у вигляді статистичної функції наступного вигляду

$$\sqrt{\mathrm{M(s)}} = \mathrm{F_2(s)} \sim \mathrm{s^H} \qquad (3.1)$$

Функція M(s) *має назву структурної функції і являє сосбою просторову або часову дисперсію випадкової функції.* Із (3.1) витікає, що функція $F_2(s)$ являє собою квадратний корінь з відомої в статистичній гідрології структурної функції $M(s)$, яка застосовується при дослідженні гідрометеорологічних величин, чия стаціонарність носить локальний характер й зберігається на порівняно невеликих інтервалах зміни аргументу. Структурну функцію визначають як математичне сподівання квадрата різниці перерізів випадкової функції. Якщо реалізація ергодичного випадкового процесу задана в дискретних точках вимірювання, то структурна функція має вигляд

$$M(s) = \frac{1}{(n-s)} \sum_{i=1}^{n-s} \left(\Delta x_i - \Delta x_{i+s}\right)^2 = 2\left[\sigma_x^2 - B(s)\right] \qquad (3.2)$$

де n - число вимірювань;

σ_x^2 -дисперсія;

$B(s)$-автоковаріаційна функція, яка розраховується за формулою:

$$B(s) = \frac{1}{(n-s)} \sum_{i=1}^{n-s} (\Delta x_i \cdot \Delta x_{i+s}). \qquad (3.3)$$

Нормована автоковаріаційна функція має вигляд

$$R(\tau) = \frac{B(\tau)}{\sigma_x^2} \qquad (3.4)$$

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

Нормована структурна функція записується у вигляді

$$m(s) = \frac{M(s)}{\sigma_x^2} = 2[1 - R(s)] \qquad (3.5)$$

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

$$m(s) = \frac{\frac{1}{2} M(s)}{\sigma_x^2} = 1 - R(s). \qquad (3.6)$$

Тоді при $R(s) = 1$ нормована структурна функція $m(s)$ буде наближуватись до нуля, й при $R(s) = 0$ - до одиниці. Коли ж s змінюється від 0 до $-\infty$, то значення $m(s)$ будуть змінюватися від нуля до 2.

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.

Лекція 4

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 забруднюючих 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 забруднюючими речовинами.

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 забруднюючих 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 забруднюючих 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 забруднюючі речовини обумовлюють більші коливання 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 забруднюючих речовин порушуються природні біохімічні процеси і концентрація забруднюючих речовин за санітарними чи іншими показниками перевищує встановлені норми. Дослідження проводяться у трьох створах (фоновому, головному і замикаючому). Фоновий створ повинен бути розташований вище скиду забруднених вод. Концентрації забруднюючих речовин у цьому створі не повинні

перевищувати гранично допустимі. Додаткові створи встановлюються між головним (контрольним) і замикаючим створами. Кількість таких створів залежить від завдань спостережень, місцевих умов та складу забруднюючих 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 забруднюючих 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).

Самоочисна здатність води на ділянці обчислюється за таким рівнянням

$$СЗ = \frac{C_B - C_H}{C_B} 100\% \quad , \qquad (4.1)$$

де C_B - концентрація забруднюючої речовини у верхньому створі,мг/дм^3; C_H - концентрація забруднюючої речовини у нижньому створі, мг/дм^3.

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

$$Sm = Q(C_B - C_H) \ , \qquad (4.2)$$

де $C_в$ – концентрація забруднюючої речовини у верхньому створі, мг/дм^3;$C_н$ - концентрація забруднюючої речовини у нижньому створі, мг/дм^3; Q – витрата води, м^3/с.

Швидкість самоочищення визначається за зниженням концентрації забруднюючої речовини за одиницю часу

$$Sr = \frac{dC}{dt} = \frac{C_B - C_H}{\tau} \ . \qquad (4.3)$$

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

$$K_C = \frac{2.3}{\tau} \lg \frac{C_B}{C_H} = \frac{1}{\tau} \ln \frac{C_B}{C_H} \quad , \qquad (4.4)$$

де $C_в$ – концентрація забруднюючої речовини у верхньому створі, мг/дм^3; $C_н$ - концентрація забруднюючої 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 середньо-арифметичні 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 забруднюючої речовини у нижньому створі, із завчасністю яка дорівнює τ.

$$C_H = C_B 10^{-\frac{\tau K_C}{2.3}} = C_B e^{-\tau K_C} \quad , \qquad (4.5)$$

де $C_в$ – концентрація забруднюючої речовини у верхньому створі; $C_н$ - концентрація забруднюючої 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:

$$r_{xy}(\tau) = \frac{K_{xy}(\tau)}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n-\tau} (x_i - \overline{x})(y_{i+\tau} - \overline{y})}{\sigma_x \sigma_y (n - \tau_{зсув} - 1)}, \qquad (4.6)$$

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 (4.6) можна представити у вигляді:

$$r_{xy}(\tau) = \frac{K_{Q_B Q_H}(\tau)}{\sigma_{Q_B}\sigma_{Q_H}} = \frac{\sum_{i=1}^{n-\tau}(Q_{Bi} - \overline{Q_B})(Q_{Hi+\tau} - \overline{Q}_H)}{\sigma_{Q_B}\sigma_{Q_H}(n - \tau_{зсув} - 1)}, \qquad (4.7)$$

де Q_{Bi} – витрати води у верхньому створі в час і; Q_{Hi} – витрати води у нижньому створі в часі і+τ; $\sigma_{Q_B}, \sigma_{Q_H}$ - 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.4.1; 4.2).

Під рядами випадкових величин Х та У слід розуміти ряд стоку (витрат води) у верхньому створі Q_B та нижньому - Q_H.

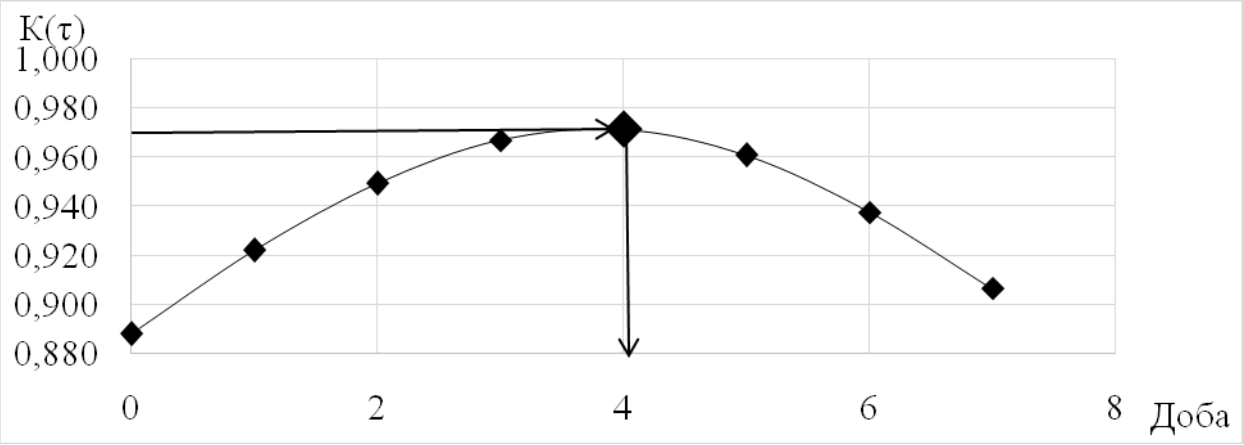


Рис 4.1. Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Псел – м. Суми відносно ряду стоку р. Псел – с. Крупець

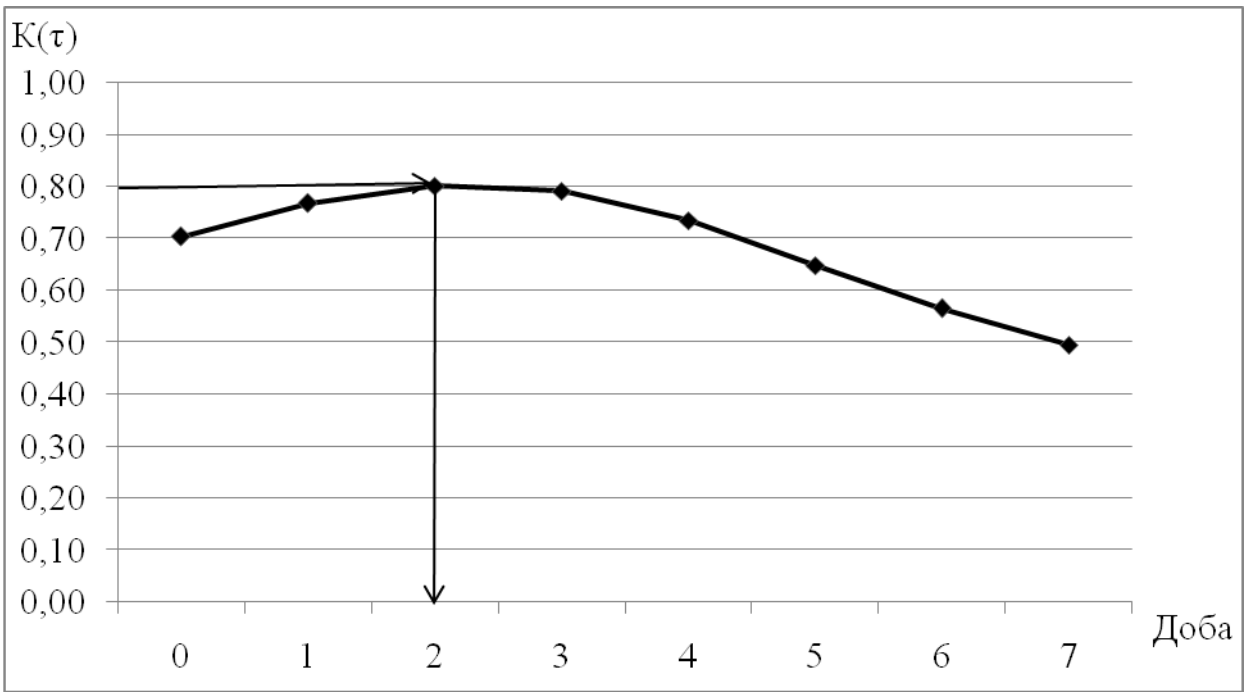


Рис. 4.2 Хід взаємної кореляційної функції при зсуві у часі ряду стоку р. Ворскла – с. Чернеччина відносно ряду стоку р. Ворскла – с. Козинка

ЛЕКЦІЯ 5

«Фрактальність гідроекологічних об’єктів»

Питання.

1. *Поняття про фрактали.*
2. *Геометрична фрактальність.*
3. *Фрактальність у математичних функціях.*
4. *Автокореляційні та спектральні функції, показники ступеня у експоненціальній залежності як показники фрактальності.*
5. *Варіаційна функція.*
6. *Визначення фрактальної розмірності за структурною функцією.*

5.1. Поняття про фрактали

Термін “фрактал” був введений Мандельбротом, співробітником дослідницького центру імені Томаса Дж.Уотсона корпорації IBM в Йорктаун-Хейтсі (шт.Нью-Йорк) і походить від латинського слова ***fractus***, що означає ламати, розбивати. 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.) фрак тали 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 (само подібністю) 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

$$d = \frac{\ln N}{\ln k}, \tag{5.1}$$

де d - фрактальна розмірність.

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.

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

$$L^{1/d} \approx A^{0.5}, \tag{5.2}$$

де L - довжина річки уздовож продольної осі;

A - площа водозбору;

d - фрактальна розмірність.

Зміна величини d може бути підставою до районування річкових систем, а сам масштабний множник може використовуватися при побудові моделей гідрологічних процесів як чинник.

Серед російських та українських вчених, які вивчають мультифрактальні властивості гідрометеорологічних об'єктів слід віднести А.Г. Бернадського (1990), досліджуючого турбулентні структури, С.С. Іванова (1994), який розглядав фрактальні властивості глобального рельєфу. Сучасний погляд на генезис фрактальних розмірностей турбулентних пульсацій в атмосфері, представлений ученими В.Д. Русовим та О.В. Глушковим.

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

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

$$E(f) \sim f^{-\beta}, \tag{5.3}$$

де f - частота.

Як відомо, автокореляційна функція стоку може бути представленою функцією виду

$$r_x(\tau) = \frac{\sum_{i=1}^{n-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\hat{\sigma}_x^2(n-\tau-1)}, \tag{5.4}$$

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;

$\bar{x}, \hat{\sigma}_x$ - відповідно середнє і стандарт початкового ряду.

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

$$\gamma = 1 - \beta. \tag{5.5}$$

При цьому γ та β можуть відігравати роль масштабуючих множників (рис.5.1).

Якщо $\beta = 0$, то вважається, що дані некорельовані и розглядуваний процес є процесом "білого шуму".

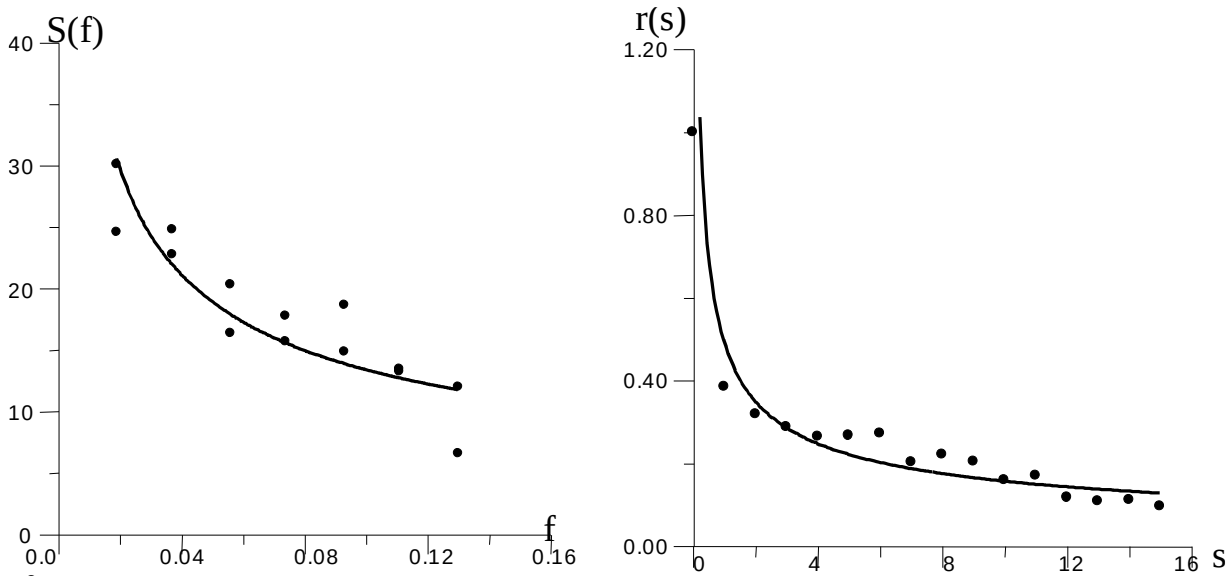


Рисунок 5.1 – Спектральна та автокореляційна функції для ряду річного стока

р.Дерекойка – г.Ялта (γ = 0,498, β = 0,497)

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

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

$$F_q(s) = s^{h(q)} \tag{5.6}$$

де $h(q)$ - масштабіруючий ступінь Рени або фрактальна розмірність, яку по відношенню до часових рядів называють також узагальненим показником Хурста.

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

$$F_2(s) = s^H. \tag{5.7}$$

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

5.2 Визначення фрактальних розмірностей за просторовою структурною функцією

Використання структурної функції постановлено в основу методу варіацій Марка та Аронсона, розробленого для вивчення самоподібних та самоаффінних (наближено само подібних) об'єктів. Суть методу полягає в дослідженні масштабної поведінки варіації деякої функції F, заданої на безлічі точок на площині. Просторова варіація V визначається як структурна функція

$$V = \langle (Zi - Zj)^2 \rangle \qquad (5.8)$$

де - Z_i, Z_j значення функції в точках і та j, а трикутні дужки означають усереднювання по всьому ансамблю точок, що відповідає рівнянню структурної функції *M(s)* (11.2). Якщо варіація масштабована з показником самоподібності Н, тобто, можна записати $V \sim L^{2H}$, то остаточно виходять результати, які зазвичай представляються у вигляді залежності $V^{1/2}$ від відстані L, в подвійному логарифмічному масштабі.

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

$$M(\Delta L) = \frac{\sum (A_i - A_j)^2}{m - 1} \qquad (5.9)$$

де $M(\Delta L)$ - значення просторової структурної функції на відрізку ΔL, віднесене до його центру; A_i, A_j - досліджувані характеристики в точках

i та j (в даному прикладі $A_i = \overline{q}_i$); m - число пар розглядуванихзначень, які потрапили у відрізок ΔL.

В межах України була задана координатна сітка із сторонами 75 км. Відстані між центрами тяжіння виділених водозборів визначалися за допомогою їх умовних координат, що розраховувалися за теоремою Піфагора:

$$L_{i,j} = \sqrt{\left(x_j - x_i\right)^2 + \left(y_j - y_i\right)^2} \tag{5.10}$$

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

З метою обчислення просторової структурної функції були задані сегменти, що не перекриваються (градації) ΔL. На основі матриці відстаней вибиралися пари водозборів, які потрапляють в задану градацію ΔL. Для кожної з пар водозборів визначалася різниця $(A_i - A_j)$ й для кожної градації – відповідне значення структурної функції, розрахованої за (11.4). Результати розрахунків звичайно представляються у вигляді графіків або (рис.5.2).

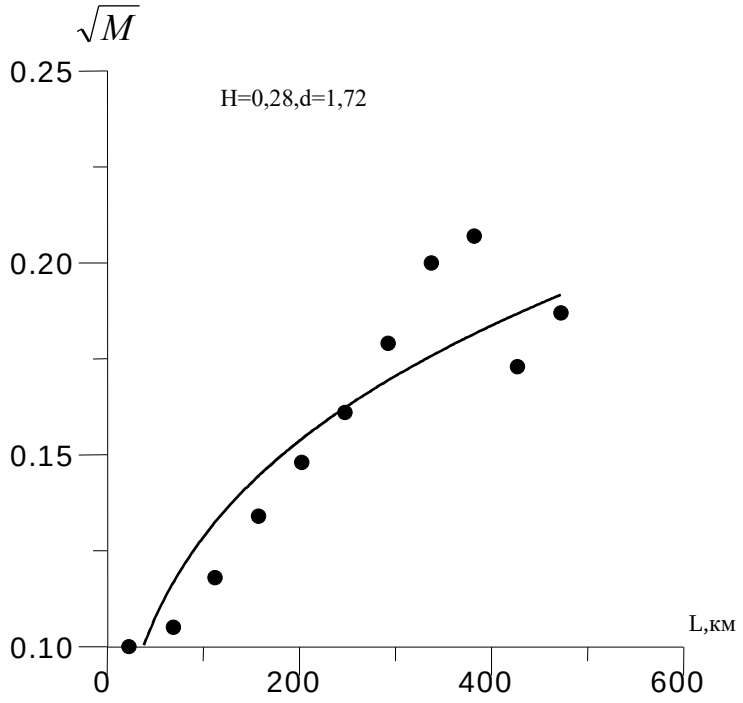


Рисунок 5.2. Залежність значень флуктуаційної функції другого порядку від відстаней між центрами тяжіння водозборів

З рисунка видно, що при $L \le 400км$ залежність варіації від відстані носить степеневий характер з показником ступеня, який визначався на основі подвійного логарифмування осей. При $L > 400км$, набуваємо шукане значення фрактальної розмірності $H = 0{,}28$. При $L > 400км$ структурна функція $M(s)$ й відповідна їй узагальнена функція $F_2(s) = \sqrt{M(s)}$ досягають стану насичення, що розглядається як ознака відсутності кореляційних зв'язків. Аналогічним чином були установлені властивості інваріантності у просторовому розподілі коефіцієнта варіації Cv річного стоку, для якого фрактальна розмірність складає $H = 0.37$. На відміну від середньоарифметичного значення просторова скорельованність значень Cv простежується тільки на відстані $L \le 200км$.

5.3 Визначення фрактальних розмірностей для часових рядів

Для визначення фрактальних розмірностей часових рядів використовуються не початкові x_i або центровані щодо середнього арифметичного ряди ($\varphi_i = x_i - \bar{x}$), а їх послідовно інтегровані значення z_n

$$z_n = \sum_{i=1}^{i=n} \varphi_i \qquad n = 1,2,...,N. \tag{5.11}$$

Слід зазначити, що z_n є функцією послідовного приросту відхилень досліджуваної величини від середнього й має назву “профіля стока”. Цей профіль має ту ж саму фізичну суть, що й різницева інтегральна крива (рис.5.3), за якою робляться висновки щодо зміни фаз водності у коливаннях стоку. Якщо переважають позитивні відхилення функція z_n буде рости, а коли переважають від’ємні відхилення, то функція z_n буде убувати.

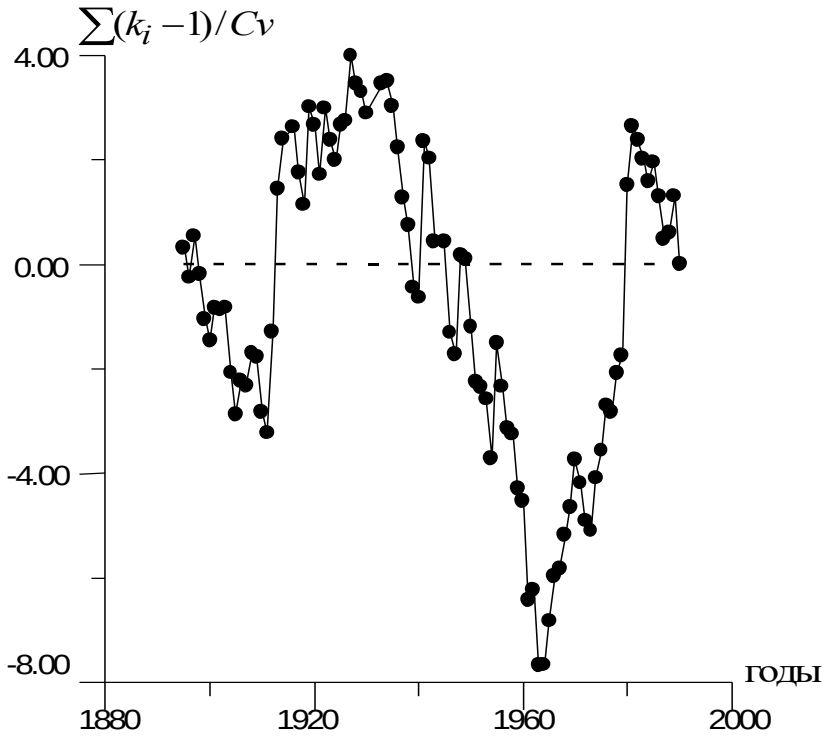


Рисунок 5.3 – Різницева інтегральна крива річного стоку р.Дністер – м. Галич

Спектральна щільність величини z_n, за теоремою Вінера – Клінчина, може бути представленою у вигляді

$$\widetilde{E}(f) \approx f^{-2-\beta}, \tag{5.12}$$

де $2+\beta$ також може розглядатися як показник масштабування.

Другий статистичний момент визначається за формулою

$$\sqrt{M(s)} = F_2(s) = \left\{ \frac{1}{2N_s} \sum_{v=1}^{Ns} \left[z_{vs} - z_{v-1,s} \right]^2 \right\}^{\frac{1}{2}} \tag{5.13}$$

де $F_q(s)$ - варіаційна або флуктуаційна функція, яка являє собою осереднену характеристуку варіації (мінливості) досліджуваної величини на часовому інтервалі s;

v - порядковый номер розглядуваного интервалу довжиною s.

$N_S = \text{int}(\frac{N}{s})$ - число інтервалів довжиною s, які не перекриваються, визначається як ціле від ділення довжини ряду N на довжину интервала s;

$z_{vs}, z_{v-1,s}$ - послідовно інтегровані значення φ_i в кінцевих точках $v-1$ та v-того интервалів довжиною s;

Підсумовування виконується від початку ряду до його кінця та у зворотній бік.

Флуктуаційна функція, що використовується у фрактальному аналізі, по суті, виконує розбиття множини N на підмножини N_s. На основі цього розкладання робляться висновки про існування ознак масштабної самоподібності. З метою встановлення таких ознак розглядаються залежності узагальненої функції $F_2(s)$ від s, подібно до аналізу рис.5.2.

Якщо $h(q)$ залежить від q, то це розглядається як властивість мультіфрактальності. У цьому випадку використовується момент порядку q:

$$F_q(s) = \left\{ \frac{1}{2Ns} \sum_{v=1}^{2Ns} \left| z_{vs} - z_{v-1,s} \right|^q \right\}^{\frac{1}{q}} \sim s^{h(q)}, \qquad (5.14)$$

де змінна q може бути будь-яким дійсним числом, що відрізняється від нуля.

6.2 Матриці кореляцій та коваріацій

Для вирішення чисельних задач гідроекології необхідні знання статистичної структури гідрометеорологічних полів (наприклад, поля стоку, опадів, швидкості вітру, тощо).Сукупність*m* гідрометеорологічних полів, які відносяться до визначених термінів спостереження, можна зобразити матрицею порядку *n*x*m*наступного вигляду

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{ni} & \dots & x_{nm} \end{bmatrix}. \qquad (6.1)$$

Матриця (6.1) утримує великий об’єм інформації про n об’єктів. Рядки матриці являють собою часові ряди відповідної гідрометеорологічної величини. Матричне зображення гідрометеорологічних та гідроекологічних 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.*

Для побудови матриці коваріацій необхідно спочатку знайти вектор середніх значень

$$\overline{X} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \dots \\ \overline{x}_i \\ \dots \\ \overline{x}_n \end{bmatrix}, \qquad (6.2)$$

де

$$\overline{x}_i = \frac{1}{m}\sum_{j=1}^{m} x_{ij} \ (i = \overline{1,m}). \qquad (6.3)$$

Якщо мова йде про поля гідрометеорологічних величин, то вектор (6.2) є осередненим полем гідрометеорологічних величин. На основі матриці (6.1) та вектора (6.2) *визначають відповідну матрицю центрованих елементів,* для чого від елементів кожного рядка матриці (6.1) віднімають відповідне середнє арифметичне значення і отримують наступний вираз

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{i1} & \Delta x_{i2} & \dots & \Delta x_{ij} & \dots & \Delta x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \Delta x_{n1} & \Delta x_{n2} & \dots & \Delta x_{ni} & \dots & \Delta x_{nm} \end{bmatrix}, \qquad (6.4)$$

де

$$\Delta x_{ij} = x_{ij} - \overline{x}_i \ . \qquad (6.5)$$

Операція проведена над матрицею (6.1) називається операцією центрування. Матриця коваріаційK_xвизначається за рівнянням:

$$K_x = \frac{1}{m} \Delta X \Delta X^{/}, \qquad (6.6)$$

де $\Delta X^{/}$ - транспонована матриця центрованих величин.
Доказ цього твердження можна показати на простій матриці

$$\Delta X = \begin{bmatrix} \Delta x_{11} & \Delta x_{12} & \dots & \Delta x_{1j} & \dots & \Delta x_{1m} \\ \Delta x_{21} & \Delta x_{22} & \dots & \Delta x_{2j} & \dots & \Delta x_{2m} \end{bmatrix}. \qquad (6.7)$$

Для матриці (6.7) рівність (6.6) в координатній формі буде мати вигляд

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

$$= \begin{bmatrix} \frac{1}{m}\sum_{j=1}^{m} \Delta x^2_{1j} & \frac{1}{m}\sum_{j=1}^{m} \Delta x_{1j}\Delta x_{2j} \\ \frac{1}{m}\sum_{j=1}^{m} \Delta x_{2j}\Delta x_{1j} & \frac{1}{m}\sum_{j=1}^{m} \Delta x^2_{2j} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & K_{12} \\ K_{21} & \sigma_2^2 \end{bmatrix}. \qquad (6.8)$$

Після обчислювань отримаємо:

$$K_x = \begin{bmatrix} \sigma_1^2 & K_{12} & \dots & K_{1j} & \dots & K_{1n} \\ K_{21} & \sigma_2^2 & \dots & K_{2j} & \dots & K_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{i1} & K_{i2} & \dots & K_{ij} & \dots & K_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ K_{n1} & K_{n2} & \dots & K_{nj} & \dots & \sigma_n^2 \end{bmatrix}. \qquad (6.9)$$

Елементи матриці (6.9) розраховуються за формулами:

$$\sigma_i^2 = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}^2 , \qquad (6.10)$$

$$K_{ij} = \frac{1}{m}\sum_{s=1}^{m} \Delta x_{is}\Delta x_{js} . \qquad (6.11)$$

Таким чином, із формул (6.10)-(6.11) випливає, що на головній діагоналі матриці (6.9) 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 (6.9) – відповідні коваріації.

Властивості матриці коваріації:

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

З останньої властивості випливає, що $|K_x| > 0$. (Прямими дужками позначаєтьсявизначник).

Матриця коваріацій поряд з вектором математичних сподівань m_xвідіграє роль параметра *щільності ймовірностей багатовимірного нормального розподілу*

$$f(x) = \frac{1}{(2\pi)^{n/2}\left|K\right|^{1/2}} \times \exp\left\{-\frac{1}{2}(X - m_x)K_x^{-1}(X - m_x)\right\} . \qquad (6.12)$$

Із матриці коваріацій можна сформувати діагональну матрицю σ середніх квадратичних відхилень. Вона має наступний вид

$$\sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{bmatrix}. \qquad (6.13)$$

Обернена матриця буде мати наступний вид

$$\sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} \end{bmatrix}. \qquad (6.14)$$

Якщо помножити ліворуч та праворуч матрицю коваріацій K_x на матрицю (6.14), отримаємо матрицю кореляцій R_x

$$R_x = \begin{bmatrix} 1 & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2j} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{i1} & r_{i2} & \dots & r_{ij} & \dots & r_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & r_{nj} & \dots & 1 \end{bmatrix}, \qquad (6.15)$$

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

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

$$r_{ij} = \frac{K_{ij}}{\sigma_i \sigma_j}. \qquad (6.16)$$

За наявності рядів спостережень довжиною m вибіркове значення коефіцієнта кореляції між двома змінними x_j та x_k розраховується наступним чином

$$r_{jk} = \frac{\sum_{i=1}^{m}\left(x_{ij} - \overline{x}_j\right)\left(x_{ik} - \overline{x}_k\right)}{(m-1)\sigma_j\sigma_k}, \qquad (6.17)$$

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

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

$\overline{x}_j$ $\overline{x}_k$ - 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 випадку є дисперсією вихідного ряду *j*.

Абсолютне значення коефіцієнта кореляції змінюється від 0 до 1 (0 ≤|r_{ij}|≤ 1). Знак «-» означає існування оберненої залежності між двома змінними. Якщо $r_{з_j} = 0$*, то лінійного зв'язку між змінними не існує (величини не корельовані), якщо* $r_{з_j} = 1$*, то існує функціональний зв'язок.*

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

ЛЕКЦІЯ 7
«Регресійні моделі»

Питання.

1. *Лінійна парна регресія.*
2. *Коефіцієнти рівняння лінійної парної регресії та коефіцієнт кореляції.*
3. *Перевірка гіпотез про статистичну значущість параметрів регресійного рівняння та коефіцієнту кореляції.*

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

$$m_{y/x} = m_y + r_{xy}\frac{\sigma_y}{\sigma_x}(x - m_x), \tag{7.1}$$

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

$$m_{x/y} = m_x + r_{xy}\frac{\sigma_x}{\sigma_y}(y - m_y), \tag{7.2}$$

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

r_{xy} - коефіцієнт кореляції;

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

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

Для вибірок (рядів спостережень) рівняння (5.1) представляється у вигляді

$$\tilde{y}_i = \tilde{y}(x_i) = \hat{m}_{y/x} = ax_i + b, \tag{7.3}$$

де x_i - дискретні значення випадкової величини X;

y_i - дискретні значення випадкової величини Y;

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

a,b - параметри рівняння.

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

$$\Delta = \sum_{i=1}^{n} \left[y_i - \tilde{y}(x_i) \right]^2 = min, \tag{7.4}$$

де n - довжина вибірки.

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

$$\frac{\partial \Delta(a,b)}{\partial a} = \frac{\partial \left[\left(\sum_{i=1}^{n} y_i - a x_i - b \right)^2 \right]}{\partial a} = 0 \ ; \tag{7.5}$$

$$\frac{\partial \Delta(a,b)}{\partial b} = \frac{\partial \left[\left(\sum_{i=1}^{n} y_i - a x_i - b \right)^2 \right]}{\partial b} = 0 \tag{7.6}$$

Вирішуючи рівняння (7.5) та (7.6) відносно a та b, одержуємо

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}; \tag{7.7}$$

$$b = \bar{y} - a\bar{x}, \tag{7.8}$$

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

Чисельник дробу, який знаходиться в правій частині рівняння (7.7) є оцінкою коваріації (коваріаційного моменту) $\hat{K}_{xy}$, розрахованого за дискретною вибіркою завдовжки n

$$\hat{K}_{xy} = \hat{cov}(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{7.9}$$

а знаменник – оцінкою дисперсії випадкової величини X

$$\hat{\sigma}_x^2 = S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \overline{x^2} - \bar{x}^2, \qquad (7.10)$$

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

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

$$\hat{r}_{xy} = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \qquad (7.11)$$

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

$$a = r\frac{S_y}{S_x}. \qquad (7.12)$$

Математична модель множинної лінійної регресії представляється рівнянням виду

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_{2)} + b_3(x_{3i} - \bar{x}_{3)} \ldots + b_k(x_{ki} - \bar{x}_{k)} \qquad (7.13)$$

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

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

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

k - число предікторов.

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

$$b_j = \frac{\sigma}{\sigma_j}\frac{D_{0j}}{D_{00}}, \tag{7.14}$$

де σ - оцінка середнього квадратичного відхилення досліджуваної характеристики y;

σ_j - оцінка середнього квадратичного відхилення j-того предіктора;

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

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

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

$$D = \begin{vmatrix} 1 & r_{01} & r_{02} \\ r_{10} & 1 & r_{12} \\ r_{20} & r_{21} & 1 \end{vmatrix}, \tag{7.15}$$

а мінори цього визначника записуються таким чином

$$D_{00} = \begin{vmatrix} 1 & r_{12} \\ r_{21} & 1 \end{vmatrix}, \tag{7.16}$$

$$D_{01} = \begin{vmatrix} r_{10} & r_{12} \\ r_{20} & r1 \end{vmatrix}, \tag{7.17}$$

$$D_{02} = \begin{vmatrix} r_{10} & 1 \\ r_{20} & r_{21} \end{vmatrix} \tag{7.18}$$

У записах виду (7.15-7.18) r_{0j} - коефіцієнт кореляції між предіктантом (0) та предіктором (j).

Рівняння лінійної множинної регресії для двох предікторів буде мати вигляд

$$\tilde{y}_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2), \tag{7.19}$$

де

$$b_1 = \frac{\sigma}{\sigma_1}\frac{D_{01}}{D_{00}} \tag{7.20}$$

$$b_2 = \frac{\sigma}{\sigma_2}\frac{D_{02}}{D_{00}} \tag{7.21}$$

Рівняння (7.19) може бути записане і таким чином

$$\tilde{y}_i = b_1 x_{1i} + b_2 x_{2i} + b_0, \tag{7.22}$$

де

$$b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2, \tag{7.23}$$

причому $\bar{y}$ - середнє арифметичне значення предіктанта;
$\bar{x}_1$ та $\bar{x}_2$ - середні арифметичні значення предікторів X_1 та X_2.

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

$$\sigma_{\tilde{y}} = \sigma\sqrt{1 - R^2}, \tag{7.24}$$

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

$$R = \sqrt{1 - \frac{D}{D_{00}}}, \tag{7.25}$$

причому D - визначник розширеної матриці коефіцієнтів кореляції.

Якщо парні коефіцієнти кореляції, які характеризують лінійний зв'язок між двома залежними випадковими величинами, змінюються від -1 до 1, то

повний коефіцієнт кореляції рівняння множинної регресії змінюється від 0 до 1.

Лінійна залежність відсутня при $r = 0$ і $R = 0$. У разі функціональної залежності $R = 1{,}0$. Чим більше коефіцієнт множинної кореляції, тим більшою мірою адекватності характеризується модель множинної регресії.

Оцінити міру адекватності можна і іншим шляхом, наприклад, шляхом перевірки статистичної гіпотези про те, що залишкова дисперсія (дисперсія вхідних даних, яка не описується рівнянням регресії) незначущо відрізняється від дисперсії предіктанта. Якщо така гіпотеза приймається, то прогноз (розрахунок) по моделі не відрізняється від випадкового.

ЛЕКЦІЯ 8
«Регресійні моделі»

Питання.

1. *Дисперсійний аналіз побудованих рівнянь регресії.*
2. *Регресійна та залишкова складові дисперсії.*
3. *Кореляційне відношення.*
4. *Перевірка гіпотези про адекватність регресійної моделі.*

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

$$\sigma_y^2 = \frac{\sum_{1}^{n}(y_i - \bar{y})^2}{n-1}. \qquad (8.1)$$

В основі формули (8.1) лежить відхилення спостереженої величини y_i від середнього арифметичного значення (рис. 8.1).

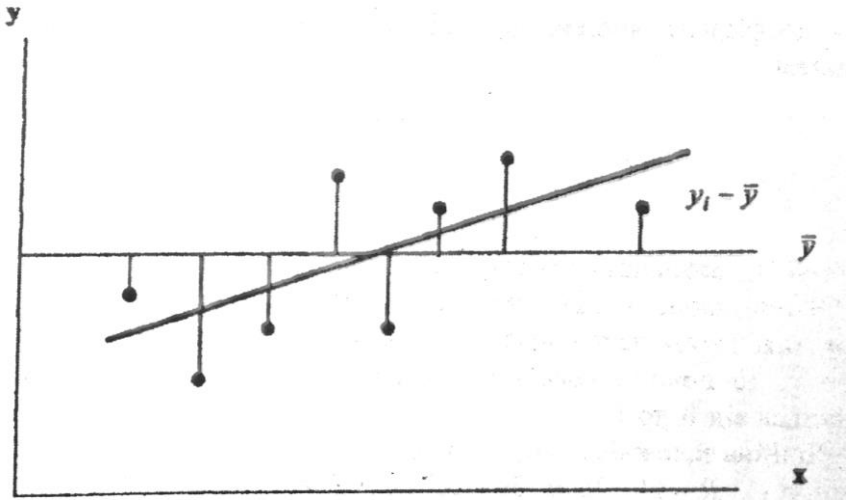


Рисунок 8.1 – Ілюстрація розсіювання спостережених значень y_i від $\bar{y}$

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

$$y_i - \bar{y} = (y_i - \tilde{y}_i) + (\tilde{y}_i - \bar{y}), \text{ або } y_i - \bar{y} = (\tilde{y}_i - \bar{y}) + (y_i - \tilde{y}) \qquad (8.2)$$

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

Обидві частини рівняння возведемо у квадрат

$$(y_i - \bar{y})^2 = [(y_i - \bar{y}) + (y_i - \tilde{y}_i)]^2. \quad (8.3)$$

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

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})(y_i - \tilde{y}_i) + \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \quad (8.4)$$

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

Отже, можна записати, що

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{n-1} + \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-1} \quad (8.5)$$

Рівняння (8.5) представляє собою суму дисперсій

$$\sigma_y^2 = \sigma_p^2 + \sigma_{зал}^2 \quad (8.6)$$

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

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

Відхилення лінії регресії $(\tilde{y}_i)$ від $\bar{y}$ графічно представлене на рис. 8.2.

Величина $(y_i - \tilde{y}_i)$ характеризує розсіювання точок, які відповідають даним спостережень від значень, розрахованих за рівнянням лінійної регресії (рис. 8.3).

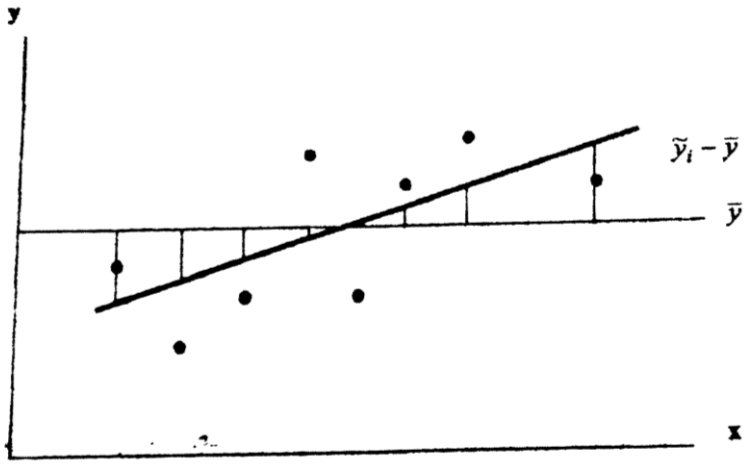


Рисунок 8.2 – Ілюстрація відхилення лінії регресії від $\bar{y}$

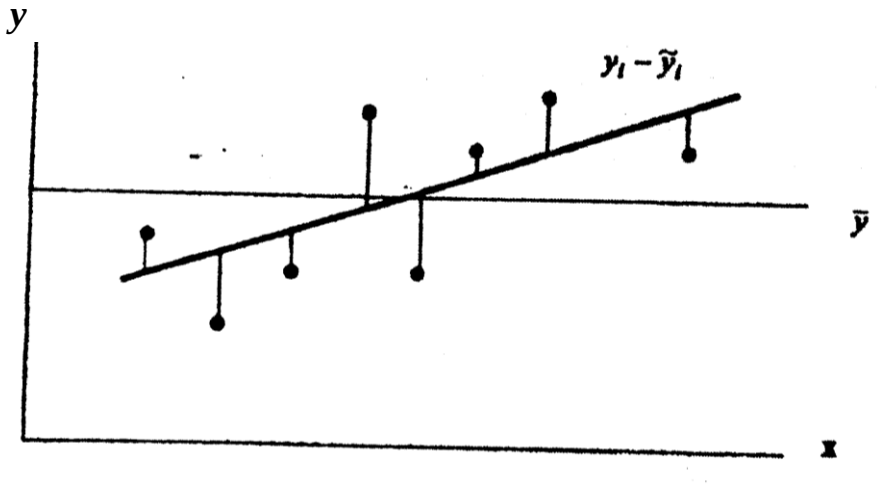


Рисунок 8.3 – Ілюстрація відхилення спостережених даних від лінії регресії

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

$$F = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 / n-1}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 / n-1} = \frac{\sigma_y^2}{\sigma_{зал}^2}. \qquad (8.7)$$

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

$$F < F_{кр}(\alpha, \nu_1, \nu_2), \qquad (8.8)$$

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

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

$$F = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1)}{\sum_{i=1}^{n-k} (y_i - \tilde{y}_i)^2 / (n-k-1)}, \qquad (8.9)$$

де n - об'єм вибірок;

k - кількість предікторів;

$y_i, \tilde{y}_i$ - фактичні і розрахункові величини.

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

ЛЕКЦІЯ 9

«Регресійні моделі»

Питання.

1. Рівняння множинної лінійної регресії.

2. Визначення коефіцієнтів рівняння за матрицею кореляцій.

3. Коефіцієнт множинної кореляції.

4. Лінійна множина регресія.

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

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

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

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

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

$$R_x = \begin{vmatrix} 1 & r_{x_2} \\ r_{x_2 x_1} & 1 \end{vmatrix} \tag{9.1}$$

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

$$R_{yx} = \begin{vmatrix} r_{yx_1} \\ r_{yx_2} \end{vmatrix} \tag{9.2}$$

На їх основі сформуємо розширену матрицю кореляцій

$$\tilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1 x_2} \\ r_{yx_2} & r_{x_1 x_2} & 1 \end{vmatrix} \tag{9.3}$$

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 (9.3) 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:

$$D_{yx_1} = r_{yx_1} - r_{yx_2} r_{x_1 x_2}, \tag{9.4}$$

$$D_{yx_2} = r_{yx_2} - r_{yx_1} r_{x_1 x_2} \tag{9.5}$$

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 $\left|\widetilde{R}\right|$ рядок і стовпець, що утримують парну кореляцію r_{yx_1}. Отже

$$D_{\overline{yx_1}} = \begin{vmatrix} 1 & r_{yx_2} \\ r_{yx_2} & 1 \end{vmatrix} = 1 - r_{yx_2}^2, \tag{9.6}$$

$$D_{\overline{yx_2}} = \begin{vmatrix} 1 & r_{yx_1} \\ r_{yx_1} & 1 \end{vmatrix} = 1 - r_{yx_1}^2 \tag{9.7}$$

Частинні коефіцієнти кореляції визначаються таким чином:

$$r_{yx_1 \bullet x_2} = \frac{D_{yx_1}}{\sqrt{|R_x| D_{\overline{yx_1}}}} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_2}^2\right)}}, \tag{9.8}$$

$$r_{yx_2 \bullet x_1} = \frac{D_{yx_2}}{\sqrt{|R_x| D_{\overline{yx_2}}}} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{x_1 x_2}^2\right)\left(1 - r_{yx_1}^2\right)}} \tag{9.9}$$

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

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

Отже, щоб визначити, яка з величин X дійсно чинить вплив на величину Y, треба розрахувати частинні коефіцієнти кореляції за допомогою формул (9.8) і (9.9).

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, с не тільки незначним, але навіть має обернений характер.

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

$$\widetilde{R} = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \dots & r_{yx_k} & \dots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \dots & r_{x_1x_k} & \dots & r_{x_1x_n} \\ r_{yx_2} & r_{x_2x_1} & 1 & r_{x_2x_3} & \dots & r_{x_2x_k} & \dots & r_{x_2x_n} \\ r_{yx_3} & r_{x_3x_1} & r_{x_3x_2} & 1 & \dots & r_{x_3x_k} & \dots & r_{x_3x_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_k} & r_{x_kx_1} & r_{x_kx_2} & r_{x_kx_3} & \dots & 1 & \dots & r_{x_kx_n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{yx_n} & r_{x_nx_1} & r_{x_nx_2} & r_{x_nx_3} & \dots & r_{x_nx_k} & \dots & 1 \end{vmatrix} \qquad (9.10)$$

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

$$r_{yx_k \bullet x_1, x_{2,\dots,} x_{k-1}, x_{k+1}, \dots x_m} = \frac{D_{yx_k}}{\sqrt{|R_x| D_{\overline{yx_k}}}}, \qquad (9.11)$$

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

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

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

ЛЕКЦІЯ 10
«Регресійні моделі»

Питання.

1. *Способи добору оптимальних пре дикторів при побудові рівняння множинної лінійної регресії.*
2. *Частинні коефіцієнти кореляції.*

10.1 Способи добору оптимальних пре дикторів при побудові рівняння множинної лінійної регресії

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

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

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

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 $\upsilon = m - k$.

10.2 Приклад аналізу розрахунків за моделлю множинної лінійної регресії з покроковим добором предікторів

Розглянуто 40 водозборів у басейні р. Уссурі (табл. 10.1). До розрахунків залучений пакет статистичних програм „Microstat”.

Необхідно добрати оптимальні предіктори та отримати розрахункове рівняння множинної лінійної регресії для визначення середньобагаторічної величини річного стоку невивчених у гідрологічному відношенні водозборів басейну р. Уссурі. Як потенційні предіктори розглядаються логарифм площі водозборів $lg(F+1)$; норма річних опадів $\overline{X}$; середня висота водозборів $H_{сер}$; заболоченість $f_б$; залісеність $f_л$; умовна довгота λ; умовна широта φ. На першому етапі розраховуються коефіцієнти лінійної парної кореляції між предіктантом та усіма предікторами. Обирається предіктор, який має найбільш тісний зв'язок з предіктантом$(\overline{q})$. У розглянутому випадку таким предіктором визнається умовна довгота λ : $r_{\overline{q}\lambda} = 0{,}692$.

Таблиця 10.1 – Вихідні дані застосовані для річок водозбору р. Уссурі.

N пп	$\overline{q}$, л/с·км^2	$lg(F+1)$	$\overline{X}$, мм	$H_{сер}$, м	$f_б$, %	$f_л$, %	λ, см	φ, см
88	17.2	3.44	813	679	-	100	10.3	10.1
45	8.64	3.03	823	260	19	81	4.0	5.9
46	11.4	1.93	848	340	8	92	3.2	10.3
48	8.72	2.69	764	176	12	78	3.9	8.3
81	10.6	3.83	886	670	0.5	100	8.2	10.0
83	12.3	4.27	881	685	0.5	100	8.8	7.8
85	12.1	4.36	871	601	5	95	8.1	8.7
92	10.3	3.28	823	205	17	83	10.6	10.9
93	10.3	3.67	831	373	4	96	5.5	7.1
96	11.1	3.25	977	445	-	94	10.1	7.4
97	13.8	2.63	1005	591	-	100	10.2	10.9
98	12.5	3.07	1180	568	-	100	10.1	10.2
100	10.74	3.41	914	403	2	98	5.4	10.6
105	12.4	4.12	956	790	0.5	100	12.4	11.4
107	11.4	4.33	919	560	4	93	10.9	11.8
1	10.34	3.71	924	879	-	91	3.6	3.8

Продовження таблиці 10.1

N пп	$\overline{q}$, л/с·км^2	$lg(F+1)$	$\overline{X}$, мм	$H_{сер}$, м	$f_б$, %	$f_л$, %	λ, см	φ, см
2	8.95	4.39	802	435	5	96	4.3	5.2
12	10.8	2.73	904	879	-	91	4.1	1.1
13	10.84	3.24	843	729	-	89	3.3	1.3
16	10.19	3.97	832	558	1	91	4.0	2.5
20	1.6	3.06	907	811	-	98	4.3	1.8

24	10.61	3.06	907	811	-	98	5.4	3.0
25	10.22	3.53	852	578	0.5	94	5.3	3.2
34	10.33	3.39	843	552	-	94	1.3	1.8
36	7.79	3.71	810	402	6	82	1.5	3.0
38	10.78	2.88	892	573	-	97	1.9	2.1
39	10.94	2.97	931	635	-	99	0.7	1.5
41	7.2	2.79	739	235	6	76	1.2	3.4
227	10.7	2.35	907	810	-	100	2.8	1.4
228	8.01	1.26	860	619	-	100	3.7	2.3
230	8.12	1.97	854	599	-	100	0.6	1.6
231	10.08	2.13	853	591	-	98	1.1	2.1
232	10.87	1.54	809	416	0.5	98	1.2	3.4
113	10.1	2.86	889	264	8	88	10.6	13.9
114	8.36	2.16	817	218	7	88	10.3	13.8
119	8.73	3.37	816	189	17	81	10.1	15.0
127	15.9	4.39	1048	629	0.5	94	12.2	15.7
128	13.3	3.49	1022	948	-	97	14.2	13.6
130	13.3	3.05	1040	350	-	99	10.3	13.3
131	11.3	2.70	905	216	10	81	10.2	15.7

Отримане рівняння лінійної парної регресії має наступний вигляд

$$\begin{aligned} &\bar{q} = 0.440\lambda + 8.03; \\ &\sigma_{\tilde{q}} = 1.61 \end{aligned} \qquad (10.1)$$

Підраховуються повна дисперсія предіктанта й регресійна та залишкова складові повної дисперсії

$$\sigma_{\bar{q}}^2 = \frac{\sum_{i=1}^{n}(q_i - \bar{q})^2}{n-1} = \frac{\sum_{i=1}^{40}(q_i - \bar{q})^2}{39} = \frac{189}{39} = 4.85 \qquad (10.2)$$

$$\sigma_p^2 = \frac{\sum_{i=1}^{n}(\tilde{q}_i - \bar{q})^2}{n-1} = \frac{90.4}{39} = 2.32 \qquad (10.3)$$

$$\sigma_{зал}^2 = \frac{\sum_{i=1}^{n}(q_i - \tilde{q}_i)^2}{n-1} = \frac{98.5}{39} = 2.51 \qquad (10.4)$$

Підраховується критерій Фішера

$$F = \frac{189/39}{98.5/38} = \frac{4.85}{2.59} = 1.87 \qquad (10.5)$$

$$\nu = 39; \nu = 38; F_{kp} = 1.7 \; F > F_{kp}$$

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

Надалі розраховуються частинні коефіцієнти кореляції між предіктанотом $\bar{q}$ та предікторами, які не увійшли до рівняння (10.3):

$$\begin{aligned} &r_{\bar{q},lg(F+1)\bullet\lambda} = 0.16; \\ &r_{\bar{q},\bar{x}\bullet\lambda} = 0.44; \\ &r_{\bar{q},H\bullet\lambda} = 0.40; \\ &r_{\bar{q},f_\delta\bullet\lambda} = 0.45; \\ &r_{\bar{q},f_л\bullet\lambda} = 0.46; \\ &r_{\bar{q},\varphi\bullet\lambda} = 0.26 \end{aligned} \qquad (10.6)$$

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

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

$$\begin{aligned} &\bar{q} = 0.105 f_л + 0.420\lambda - 1.66; \\ &\sigma_{\tilde{q}} = 1.44; \qquad R = 0.768 \end{aligned} \qquad (10.7)$$

Коефіцієнт кореляції зростає від 0.692 до 0.768.

Регресійна складова, що показує, яка частина дисперсії величини $\bar{q}$ описується її залежністю від λ та $f_л$, збільшується й становить 2.94, а залишкова складова зменшується до 2.010.

Критерій Фішера приймає значення

$$F = \frac{189/39}{77.3/37} = \frac{4.85}{2.09} = 2.32 \qquad (10.8)$$

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

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

Знов обчислюються частинні коефіцієнти кореляції між предіктантом $\overline{q}$ та предікторами, які не увійшли до рівняння (10.7)

$$\begin{aligned} &r_{\overline{q},lg(F+1)\bullet\lambda,f_л} = 0.25; \\ &r_{\overline{q},\overline{X}\bullet\lambda,f_л} = 0.27; \\ &r_{\overline{q},H\bullet\lambda,f_л} = 0.13; \\ &r_{\overline{q},f_\delta\bullet\lambda,f_л} = 0.14; \\ &r_{\overline{q},\varphi\bullet\lambda,f_л} = 0.00 \end{aligned} \qquad (10.9)$$

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

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

$$t = \frac{r_{xy}}{\sqrt{\sigma^2_{r_{xy}}}} \ , \qquad (10.10)$$

$$\sigma_{r_{xy}} = \frac{1 - r^2_{xy}}{\sqrt{n-1}} \ , \qquad (10.11)$$

де t - критерій Стьюдента.

$$\sigma_{r_{\overline{q},\overline{x}\bullet\lambda,f_л}} = \frac{1 - 0.27^2}{\sqrt{40-1}} = \frac{1 - 0.0729}{6.24} = \frac{0.927}{6.24} = 0.149 \qquad (10.12)$$

$$t = \frac{0.27}{0.149} = 0.85 . \qquad (10.13)$$

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

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

Після виконання перевірки частинних коефіцієнтів кореляції на їх значущість, на основі отриманого розрахункового рівняння виконуються перевірні розрахунки, які зводяться у таблицю 10.2.

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

Таблиця 10.2 – Оцінка похибки перевірних розрахунків

N	$\bar{q}_i$	$\tilde{q}_i$	$\Delta\bar{q} = \bar{q}_i - \tilde{q}_i$	$\delta_i = \frac{\Delta\bar{q}_i}{\bar{q}_i} \cdot 100\%$
1	17.20	12.76	4.44	25.80
2	8.64	8.53	0.11	1.23
3	11.40	10.35	2.04	17.80
4	8.72	8.17	0.54	10.20
5	10.60	12.29	-1.69	-15.90
6	12.30	12.55	-0.24	-1.95
7	12.10	11.72	0.37	3.05
8	10.30	10.83	0.46	4.50
9	10.30	10.74	-0.44	-4.27
10	11.10	10.78	0.31	2.88
11	13.80	11.45	2.34	110.90

Продовження таблиці 10.2

N	$\bar{q}_i$	$\tilde{q}_i$	$\Delta\bar{q} = \bar{q}_i - \tilde{q}_i$	$\delta_i = \frac{\Delta\bar{q}_i}{\bar{q}_i} \cdot 100\%$
12	12.50	11.41	1.08	8.64
13	10.74	10.90	-1.16	-11.90
14	12.40	14.06	-1.66	-13.38
15	11.40	12.27	-0.87	-7.63
16	10.34	10.41	-0.07	-0.75
17	8.95	10.23	-1.28	-14.30
18	10.80	10.62	1.17	10.83

19	10.84	10.08	0.76	7.72
20	10.19	10.58	-0.39	-4.24
21	11.60	10.44	1.15	10.90
22	10.61	10.90	-1.29	-13.40
23	10.22	10.44	-1.22	-13.20
24	10.33	8.76	0.56	10.00
25	7.79	7.58	0.20	2.56
26	10.78	10.33	0.44	4.59
27	10.94	10.03	0.90	10.05
28	7.20	10.83	0.36	5.13
29	10.70	10.02	0.67	10.26
30	8.01	10.40	-2.39	-210.80
31	8.12	10.10	-0.98	-12.06
32	10.08	10.10	-0.02	-0.22
33	10.87	10.14	-2.27	-33.04
34	10.10	10.36	-0.26	-2.57
35	8.36	10.23	-1.87	-22.36
36	8.73	10.67	-1.94	-22.22
37	15.90	13.34	2.55	110.03
38	13.30	14.50	-1.20	-10.00
39	13.30	12.65	0.64	4.86
40	11.30	11.14	0.16	1.42

$$|\delta|_{сер}=10.08$$

ЗМІСТОВНИЙ МОДУЛЬ 3

ЛЕКЦІЯ 11
«Метод сумісного аналізу даних»

Питання.

1. *Просторова дисперсія досліджуваної величини, її географічна та випадкова складові.*
2. *Умова, за якою визначається можливість просторового узагальнення.*
3. *Середня квадратична похибка визначення осередненого параметру.*
4. *Уточнення досліджуваної величини за даними сумісного аналізу.*
5. *Вибір оптимального складу об'єднувальної сукупності.*

11.1. Теоретичні основи методу

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

Навіть при довгих рядах спостережень оцінки окремих статистичних параметрів визначаються з великою погрішністю, тобто є статистично незначущими. До числа таких параметрів відносяться насамперед

коефіцієнти автокореляції $r(1)$ й асиметрії Cs, а також розрахункове відношення Cs/Cv.

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

$$\sigma_{\Pi}^2 = \sigma_{\Gamma}^2 + \sigma_B^2 \ , \qquad (11.1)$$

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

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

σ_B^2 - випадкова складова дисперсії параметра.

При цьому повна просторова дисперсія параметра оцінюється за формулою

$$\sigma_{\Pi}^2 = \frac{\sum_{j=1}^{k}(A_j - A_{CEP})^2}{k-1} \ , \qquad (11.2)$$

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

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

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

A_{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

$$\sigma_B^2 = \frac{\sum_{j=1}^{k}\sigma_{Aj}^2}{k} \ , \qquad (11.3)$$

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

Географічна складова знаходиться за допомогою зворотного розрахунку з (11.1):

$$\sigma_Г^2 = \sigma_П^2 - \sigma_B^2. \qquad (11.4)$$

Якщо виконується умова

$$\frac{\sigma_B^2}{\sigma_П^2} > \frac{\sigma_Г^2}{\sigma_П^2}\ , \qquad (11.5)$$

то можна зробити висновок, що просторовий розподіл досліджуваного параметра в більшій мірі визначається випадковими властивостями поєднуваних вибірок і в меншій - зміною фізико-географічних умов формування стоку по території. Таким чином, при виконанні (11.5) приймається рішення, що вибіркові оцінки параметрів можуть бути осереднені в межах досліджуваної території. Необхідно підкреслити, що *якість об'єднання тим вища, чим менший внесок географічної складової у повну просторову дисперсію параметра.* Географічна складова є, власне кажучи, оцінкою статистичної неоднорідності вихідного матеріалу. Коли оцінки вибіркових параметрів дуже великі, географічна складова дисперсії, обчислена зворотним розрахунком за (11.4), 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

$$\sigma_{CEP} = \sqrt{\frac{\sigma_B^2}{k} + \sigma_Г^2} \qquad (11.6)$$

Величина σ_{CEP} поряд з умовою (11.5) також є критерієм якості об'єднання. Осереднена оцінка параметра визнається статистично достовірною, коли виконується умова

$$A_{CEP} > 2\sigma_{CEP} \qquad (11.7)$$

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 $\sigma_Г^2$ дисперсія осередненого у межах поєднуваної сукупності параметра σ_{CEP}^2 відповідно до виразу (11.6) 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

$$\varepsilon_A = \frac{\sigma_A}{A} \cdot 100\% \ , \qquad (11.8)$$

де σ_A - середнє квадратичне відхилення оцінки параметра A.

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

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

$$A_j' = \frac{A_j \sigma_{CEP}^2 + A_{CEP} \sigma_j^2}{\sigma_{CEP}^2 + \sigma_j^2} \ , \qquad (11.9)$$

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

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

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

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

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

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

$$\sigma_j^{'} = \frac{\sigma_j \sigma_{CEP}}{\sqrt{\sigma_j^2 + \sigma_{CEP}^2}}\ , \qquad (11.10)$$

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

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

Таким чином, метод С.М. Крицкого та М.Ф. Менкеля дозволяє вирішувати багато задач географічного узагальнення. Наприклад, задача вибору способу географічного узагальнення може бути вирішена при розгляді умови (11.5). 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 (11.7) та аналіз залежності $\sigma_{CEP}^2 = \varphi(k)$. *Зростання географічної складової повної просторової дисперсії параметра буде тим інтенсивнішим, чим ширше межі просторового узагальнення.*

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

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

11.2. Приклади застосування методу сумісного аналізу даних

У табл. 11.1 наведений приклад розподілу складових просторової дисперсії середньобагаторічних значень (норм) річного стоку для басейну р.Уссурі (*Лобода Н.С., Нгуєн Ву Ань, 2006*). Попередні угрупування водозборів у райони (Північний, Центральний, Південний) виділені згідно з результатами районування за синхронністю коливань стоку. Особливості коливань стоку у басейні р.Уссурі обумовлюється різним характером мусонних процесів на півночі та півдні водозбору. У північній частині

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

У межах басейну р.Дністер при обгрунтуванні способу узагальнення статистичних параметрів $Cv, r(1), Cs/Cv$ попередньо були виділені 4 райони за особливостями фізико-геграфічних умов. **Перший** містить у собі гірський Дністер (правобережні притоки), **другий** – передгірський Дністер (лівобережні притоки), **третій** – річки Північної і Центральної Молдови, які характеризуються підвищеним підземним живленням річок за рахунок розвантаження карстових вод. **Четвертий** район утворюють річки Причорноморської низовини з посушливим кліматом і незначною часткою підземного живлення (5%).

Таблиця 11.1 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення норм річного стоку річок в басейні р. Уссурі

Район	Середнє значення параметру $\overline{q}$, $л/с \cdot км^2$	Дисперсія			ε_{CP}, %
		повна	випадкова складова	географічна складова	
НОРМА РІЧНОГО СТОКУ					
Північний	12,3	9,80	0,52 5%	9,28 95%	24,8
ЦЕНТРАЛЬНИЙ	11,2	4,99	0,36 7%	4,63 93%	19,2
Південний	8,94	2,55	0,587 23%	1,97 77%	15,8
Водозбір р.Уссурі	10,3	6,39	0,48 8%	5,91 92%	34,0

Просторова зміна коефіцієнтів варіації обумовлена характером зміни загальної зволоженості території: значення коефіцієнтів варіації збільшуються в міру переходу з зон *надлишкового та достатнього зволоження (райони 1 і 2) в зону недостатнього зволоження (райони 3 та 4).* Але на просторовий розподіл коефіцієнтів варіації впливає такий фактор підстильної поверхні, як гідрогеологічна структура й пов'язане з нею підземне живлення. По перше, у межах водозбору р.Дністер існують карстові зони, які впливають на багаторічну мінливість стоку. По –друге, коефіцієнти варіації річок з площею меньшою другої критичної можуть суттєво відрізнятись від коефіцієнтів варіації великих річок (*Бефані А.М., Мельничук О.М., 1967*). У зв'язку з цим, для об'єднання за методом сумісного аналізу бажано використовувати дані по водозборах з площею більшою $1000 км^2$ (друга критична площа для більшості розглядуваних річок). Певний інтерес являє собою район 1(Карпати та Передкарпаття), де географічна й випадкова складові близькі одна до одної. У таких випадках при практичному застосуванні для водозборів із короткими рядами спостережень рекомендується використовувати не осереднені, а уточнені за формулою (11.9) значення коефіцієнтів варіації. Як випливає з табл.11.2, коефіцієнти варіації верхньої частини водозбору р.Дністер змінюються незначно в зоні надлишкового зволоження й достатнього зволоження, унаслідок чого райони 1 і 2 можливо об'єднати в один. Для району 3 (Середній Дністер) географічна складова приймає від'ємне значення, у зв'язку з чим приймається рівною 0, у той час як випадкова складова дорівнює 100%. Для нижньої частини водозбору р.Дністер через близкість значень випадкової та географічної складових також бажано використовувати уточнені оцінки параметру Cv. Відносна похибка визначення осередненого значення параметру для усіх районів меньша за допустиму (15%).

Просторовий розподіл коефіцієнтів автокореляції визначається в більшій мірі не географічною зональністю, а внеском підземного живлення у формування стоку (табл. 11.3). У гірській частині р.Дністер (зона надлишкового зволоження) та Причорномор'ї (зона недостатнього зволоження), де внесок підземного живлення слабко виражений, коефіцієнти автокореляції близькі до нуля. У межах Волино-Подільського артезіанського басейну (район 2), а також Північної і Центральної Молдови (район 3), де за рахунок близького залягання водоносних горизонтів і наявності карстових вод підземне живлення істотне, відзначаються високі значення коефіцієнтів автокореляції, які наближаються до 0,5. Для оцінки якості об'єднання коефіцієнтів автокореляції r(1) та відношення Cs/Cv використовується (11.7).

Слід зазначити, що за часів колишнього СРСР для території України було рекомендовано використовувати коефіцієнт автокореляції, який дорівнює 0,22. Але по результатах, наведених у табл.11.3, можна зробити висновок, що діапазон значень r(1) набагато ширший. Так, у межах України

на основі методу сумісного аналіза для коефіцієнта автокореляції річного стоку виділено 7 районів, а для коефіцієнта асиметрії –10 (*Лобода Н.С.,2005*).

Таблиця 11.2 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення коефіцієнтів варіації річного стоку річок в басейні р.Дністер

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

Таблиця 11.3 – Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення коефіцієнтів автокорреляції річного стоку річок в басейні р. Дністер

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

Що стосується відношення Cs/Cv, то в зв'язку з великими похибками розрахунку коефіцієнта асиметрії за даними, перевірка гіпотези про можливість районування виконувалося в межах районів, виділених для коефіцієнта варіації (табл.11.2). У гірській зоні басейну р.Дністер відношення Cs/Cv можна прийняти рівним 2 , на лівобережжі р.Дністер - 3, а в середньому плині 1,5. У нижній частині басейну р.Дністер дане відношення рекомендується приймати рівним 1,7 (табл.11.4).

Таблиця 11.4 - Результати застосування методу сумісного аналізу даних до обгрунтування способу просторового узагальнення відношення Cs/Cv річного стоку річок в басейні р.Дністер

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

ЛЕКЦІЯ 12

«Модель факторного аналізу»

Питання.

1. *Основні положення факторного аналізу.*
2. *Аналіз факторних навантажень.*
3. *Вибір кількості головних факторів.*
4. *Інтерпретація головних факторів.*
5. *Виділенння головних чинників формування процесу на різних масштабах.*

12.1.Теоретичні основи методу факторного аналіза

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

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

$$x_j = \sum_{p=1}^{k} l_{jp} f_p + \nu_j \ , \ (j = 1, m) \qquad (12.1)$$

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 (12.1) розгорнути суму, то прийдемо до системи рівнянь (*Школьний Є.П., Лоєва І.Д., Гончарова Л.Д.,1999*):

$$\begin{cases} x_1 = l_{11}f_1 + l_{12}f_2 + ... + l_{1k}f_k + v_1 \\ x_2 = l_{21}f_1 + l_{22}f_2 + ... + l_{2k}f_k + v_2 \\ ..... \\ x_m = l_{m1}f_1 + l_{m2}f_2 + ... + l_{mk}f_k + v_m \end{cases} \quad (12.2)$$

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

Матрична форма рівняння (12.1) має вигляд

$$X = LF + V, \quad (12.3)$$

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

$$K = M\left[X \cdot X'\right] \quad (12.4)$$

Ураховуючи (12.3), можна прийти до такого матричного виразу

$$K = LL' + D, \quad (12.5)$$

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

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

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

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

$$\frac{\partial L_П}{\partial l_{jp}} = 0; \quad \frac{\partial L_П}{\partial d_j} = 0, \tag{12.6}$$

де $L_П$ - функція правдоподібності.

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

$$K_{jj} = \sum_{p=1}^{k} l_{jp}^2 + d_j \text{, при} \quad j = i; \tag{12.7}$$

$$K_{ji} = \sum_{p=1}^{k} l_{jp} l_{ip} \quad \text{при} \quad j \neq i. \tag{12.8}$$

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

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

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

$$h_j^2 = \sum_{p=1}^{k} l_{jp}^2 \tag{12.9}$$

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

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

$$S_p = \frac{\sum_{j=1}^{m} l_{pj}^2}{m} 100\%, \tag{12.10}$$

де m - кількість розглядуваних змінних.

Загальний внесок всіх виділених факторів в сумарну дисперсію досліджуваних змінних дорівнює

$$S = \sum_{p=1}^{k} S_p \ . \qquad (12.11)$$

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

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

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

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

При аналізі синхронності коливань стоку розглядаються не зв'язки між ознаками, а зв'язки між рядами, при цьому використовується Q-техника факторного аналізу, яка може розглядаєтися як варіант класифікаційного аналізу. Внесок кожного фактора у дисперсію змінної у даному випадку представляє собою внесок у дисперсію ряду спостережень за стоком. Не зупиняючись на фізичній інтерпретації факторів, при дослідженні синхронності коливань річного стоку в практиці гідрологічних розрахунків застосовують наступні графічні побудови (*Жук В.А., Євстігнєєв В.М.,1993*). У випадку, коли перших два фактора описують більше 60% загальної дисперсії вихідних даних, на графіку, осі якого являють собою два фактори, проводять вектори з початку координат у точку з координатами, відповідними факторним навантаженням. Довжина вектора розраховується за виразом

$$d_j = \sqrt{l_{j1}^2 + l_{j2}^2} \ , \qquad (12.12)$$

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

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

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

$$\theta_j = \arcsin \frac{l_{j3}}{d_j} \tag{12.13}$$

$$\lambda_j = \arcsin \frac{l_{j2}}{\sqrt{l_{j1}^2 + l_{j2}^2}}, \tag{12.14}$$

де

$$d_j = \sqrt{l_{1j}^2 + l_{2j}^2 + l_{3j}^2} \tag{12.15}$$

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

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

Таким чином, застосування Q-модіфікації факторного аналізу дозволяє стиснути інформацію, яка міститься в кореляційній матриці, й інтерпретувати її. Компактне трактування кореляційної матриці досягається при розгляді кожного угрупування водозборів. При наближенні кута між векторами, спрямованими з початку координат до центрів угрупувань, до $90^\circ C$ виділені угрупування розглядаються як райони з асинхронними коливаннями стоку. Чим меньший кут між угрупуваннями, тим тісніший зв’язок між стоком розглядуваних річок. Усередині угрупувань (районів) можна виділяти окремі групи точок, які по їх територіальному розташуванню і особливостям формування стоку можна інтерпретувати як підрайони.

Розглянемо приклад районування за закономірностями коливань стоку Північно-Західної частини України. Використано 29 рядів річного стоку з

періодом сумісних спостережень з 1955 по 1986 роки, який складає 32 роки. До аналізу залучені дані по водозборах р.Західний Буг, правобережні притоки Прип'яті, правобережні притоки р. Дніпро - рр. Уж, Ірша, Тетерев) та прилеглих територій (лівобережні притоки р. Дністер, верхів'я р.Південний Буг). Установлено, що перші два фактори пояснюють 57,4% сумарної дисперсії вихідних даних, а перші три - 81,9% . Районування за Q - модифікацією факторного аналізу здійснено на підставі графічних побудувань, в яких використовуються результати представлення кореляційної матриці у вагових навантаженнях кожного ряду на виявлений гіпотетичний фактор. За результатами факторного аналізу виділені два угрупування, що утворюють два територіальних райони (західний 1а та східний 1б) з синхронними коливаннями стоку (рис. 12.1).

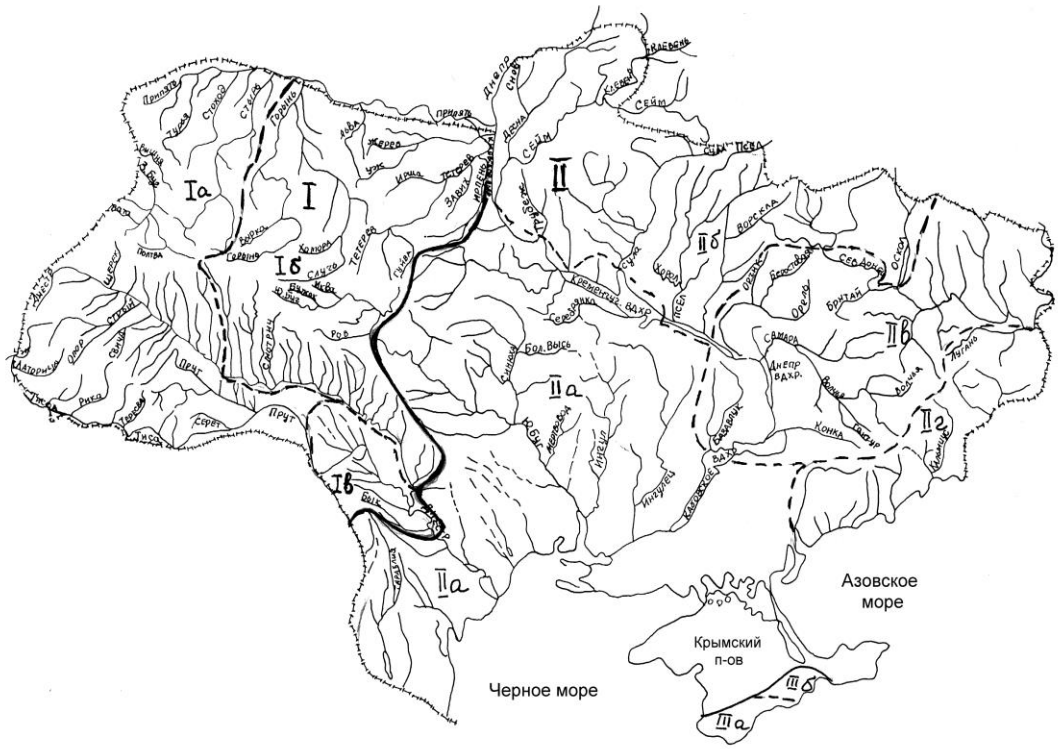


Рисунок 12.1 - Карта-схема районів с синхронними коливаннями річного стоку річок України

В перший район входять водозбори річок бассейну р.Західний Буг, верхів'я р. Дністер, а також такі притоки Прип'яті, як Вижівка, Турія, Стир. Друге угруппування утворюють притоки р.Дністер від р. Серет до р. Ушиця включно, верхів'я р. Південний Буг, правобережні притоки р.Прип'ять,

починаючи від р.Горинь (рис.12.2). Лінія розмежування проходить через вододіл річок Стир – Горинь.

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

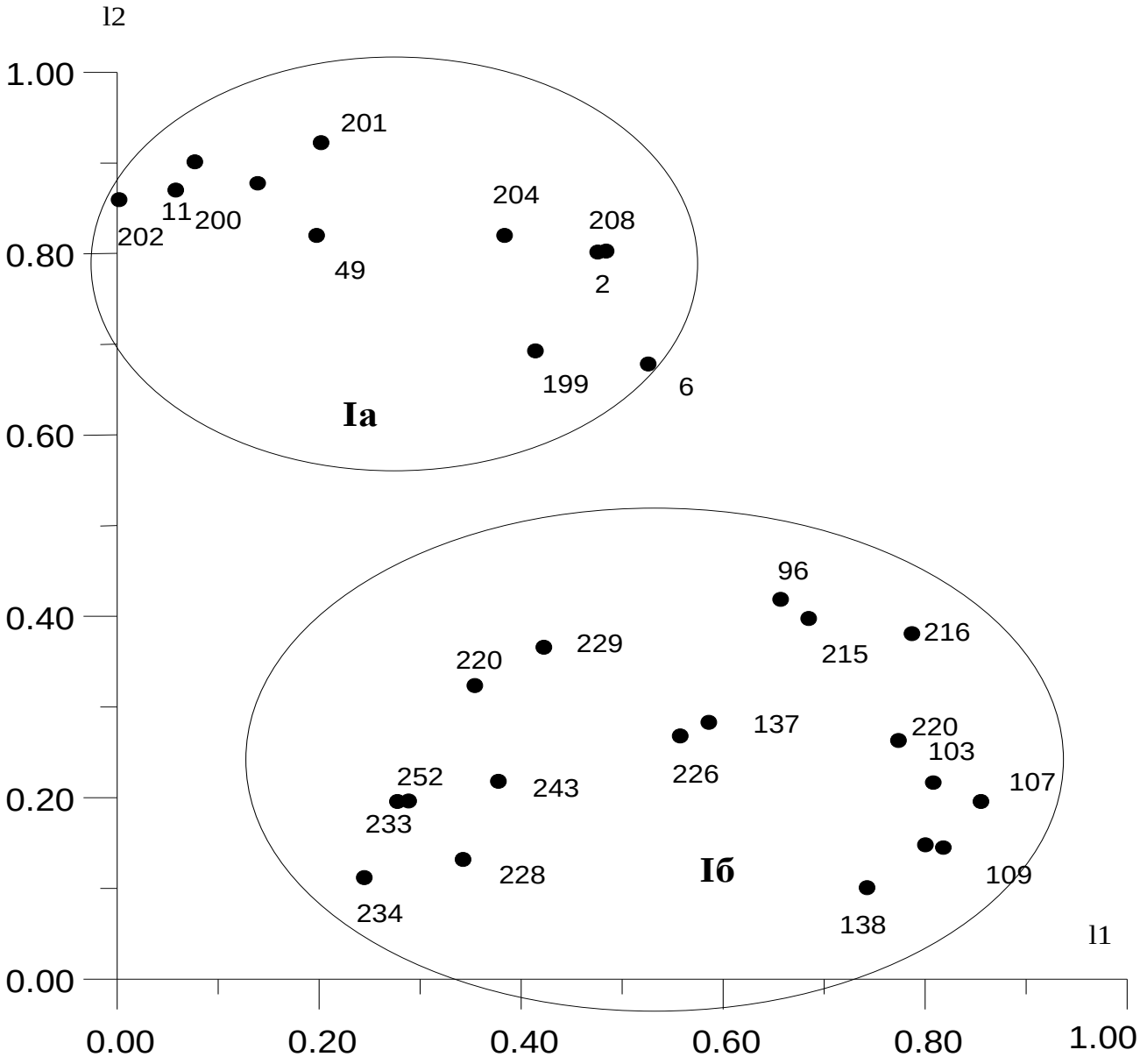


Рисунок 12.2 – Виділення угрупувань з синхронними коливаннями річного стоку за двома факторами для річок Північно-Західної України (біля точок – номера водозборів)

Аналогічним чином були виділені райони IIIа (Західний) та IIIб (Східний) для водозборів Гірського Криму (рис.12.3). Внесок перших двох факторів становить 81%, а кут між двома угрупуваннями близький до $90^{\circ}C$. Це вказує на суттєву різницю у коливаннях стоку заходу і сходу, оскільки ($r = \cos(90^{\circ}) = 0$). Окреме положення займає водозбір р.Салгір-м.Сімферополь, особливість коливань стоку на цьому водозборі пов'язана з інтенсивним використанням стоку для водогосподарських потреб, яке порушує загальну закономірність коливань стоку, обумовлену коливаннями кліматичних факторів. Як правило, такі водозбори мають низький рівень інформативності, який описується величиною d. Наприклад, для водозбору р.Салгір- м.Сімферополь ця величина становить 0,45. Різниця між західною й східною частинами обумовлена кліматичними факторами, насамперед, опадами. Західні схили знаходяться під значним впливом середиземноморських циклонів, які забезпечують приплив зволожених повітряних мас. Захищенність Кримськими горами південно-західного узбережжя забезпечує формування субтропічного клімату, що приводить до значного розбігу між умовами формування стоку заходу і сходу.

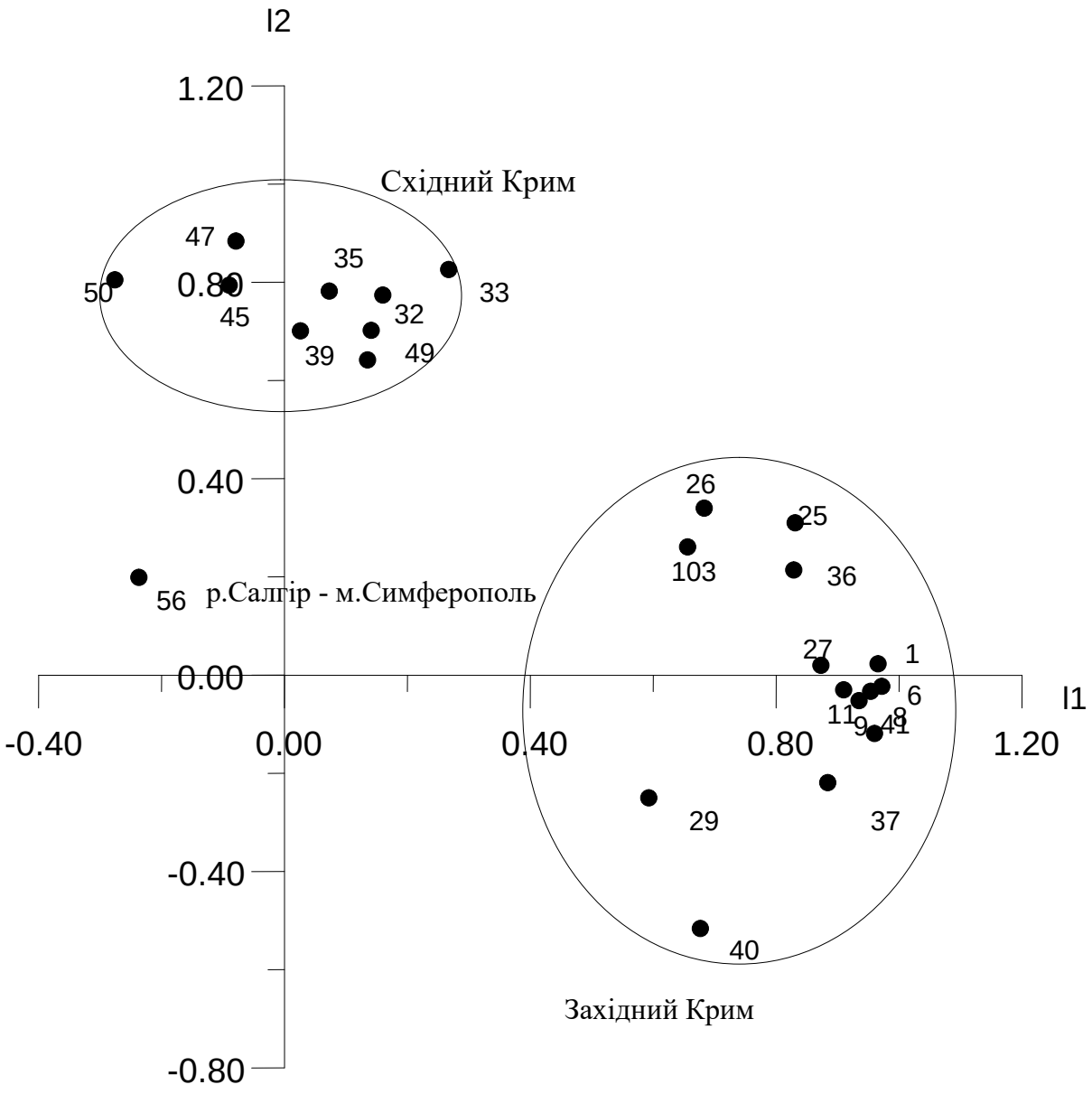


Рисунок 12.3 – Виділення угрупувань з синхронними коливаннями річного стоку гірського Криму за двома першими факторами
(біля точок – номера водозборів)

ЛЕКЦІЯ 13

«Дискримінантний аналіз як метод прогнозу або вирішальне правило»

Питання.

1. *Схема побудови розв'язувального правила.*
2. *Форми дискримінант них функцій.*
3. *Квадратична, лінійна та спрощена дискримінантна функція.*
4. *Число Махаланобіса.*
5. *Схеми прогнозування та районування.*

13.1. Схема побудови розв'язувального правила

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

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

У теорії розпізнавання образів задача класифікації формулюється таким чином: на основі відомостей про окремих представників різних класів навчаючої системи із характерними для них ознаками (предікторами) 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 котрого 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. Ясно, що

$$V_1 \cup V_2 = V, V_1 \cap V = \mathrm{O} \qquad (13.1)$$

Поперед усього, як зазначалося вище, треба побудувати поверхню, яка б розділяла підпростори R_{V1} і R_{V2}. Наведемо для пояснення прості приклади.

Нехай простір буде двовимірним $R_V = R_V (x_1, x_2)$ (рис.13.1). Тоді ми маємо на площині (x_1, x_2) лінію $x_1 = f (x_2)$, що розділяє підпростір R_{V1} від підпростору R_{V2}. Досить простим буде і випадок трьохвимірного простору $R_V = R_V(x_1, x_2,x_3)$ (рис.13.2). У цьому випадку підпростори R_{V1} і R_{V2} розділяє деяка поверхня у трьохвимірному просторі, рівняння якої має вид $x_3=\varphi(x_1, x_2)$.

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

$$F(x_1,x2,\ldots,x_n) = 0. \qquad (13.2)$$

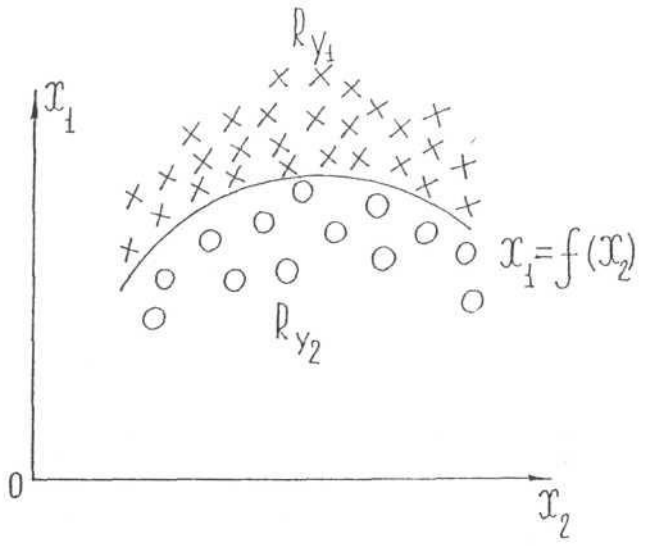


Рисунок 13.1 - Образи і розділяюча функція в двовимірному просторі

Надалі отримується правило, за допомого якого є підстава віднести вектор X, що підлягає розпізнаванню, до підпростору R_{V1}, або підпростору R_{V2}. Це правило називають розв'язувальним правилом. Якщо відповідно до нього приймається рішення, що $X \in R_{V1}$, то явище прогнозується, якщо приймається рішення, що , $X \in R_{V2}$, то явище не прогнозується. Етап, який складається з побудови розділяючої гіперповерхні та розв'язувального правила, носить назву етапа навчання.

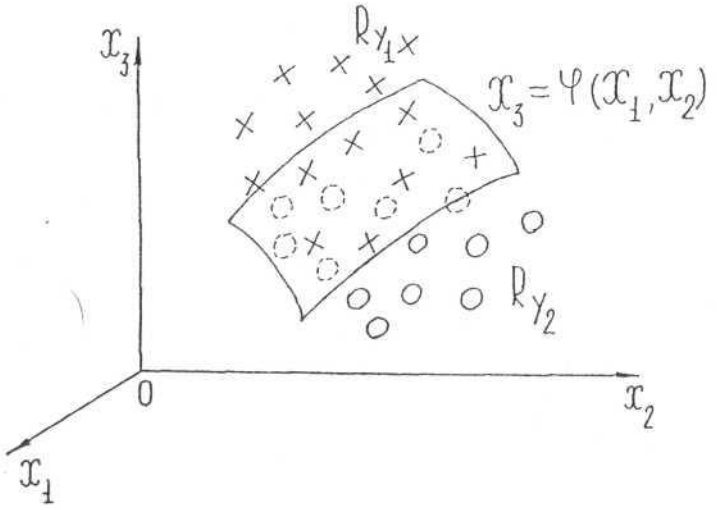


Рисунок 13.2 – Образи та розділяюча поверхня в трьохвимірному просторі

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 розділимість класів, а з іншої сторони, допускає їх часткову "перекриваємість". Таким чином, рішення про віднесення явища або об'єкту до того чи іншого класу приймається з тією чи іншою долею похибки. "Перекриваємість" класів виражається у тому, що інтервал числових значень ознак у об'єктів, які належать до різних класів, перекивається. Це утруднює віднесення об'єкта або явища до визначеного класу, якщо їх кількісна ознака лежить у області перекриття. Наприклад, виділені два гідрологічних райони, які відрізняються один від одного кількісними характеристиками стоку. Але розподіл стокових величин у просторі підкорюється закону географічної зональності. Отже існують "приграничні" водозбори, на яких розглядувані характеристики можуть попадати у область "перекриття". Коли класифікація виконується по одній із ознак, то водозбір буде віднесений до першого класу (району А), якщо ознака x буде меньшою за x_0, і до другого (району В), якщо $x > x_0$. Неправильні рішення позначимо через δ_a та δ_b. У якості x_0 можна вибрати середину інтервалу перекриття або абсцису точки перетину кривих щільності розподілу x_0. Але більш доцільно вибрати таку точку x^*_0,

для якої буде виконуватись $\delta_a = \delta_b$. За цієї умови максимальна похибка (найбільша з двох) має минимальне 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.

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 обгрунтовано прийняти гіпотезу H_1 або H_2. Всіляка процедура перевірки гіпотез передбачає, що приймаючи те чи інше рішення, ми можемо припустити помилку 1-го чи 2-го роду. Нагадаємо, що помилку 1-го роду ми припускаємо, коли відкидаємо правильну гіпотезу. Помилка 2-го роду пов'язана з прийняттям невірної гіпотези. Помилку другого роду ще називають “похибкою хибної тривоги”.

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

$$P\left(x_{1,}x_2,...,x_n / V_1\right) \tag{13.3}$$

та

$$P\left(x_{1,}x_2,...,x_n / V_2\right) \tag{13.4}$$

Позначимо ймовірність помилки 1-го роду через P_a, а 2-го роду через P_b. Знаючи ймовірності (13.3) та (13.4), а також апріорні ймовірності $P(V_1)$ і $P(V_2)$ класів V_1 і V_2, можна розрахувати ймовірності помилок 1-го й 2-го роду.

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

$$\lambda\left(x_{1,}x_{2,...,}x_n\right) = \frac{P\left(x_{1,}x_{2,...,}x_n / V_1\right)}{P\left(x_{1,}x_{2,...,}x_n / V2\right)} \tag{13.5}$$

а величину

$$\frac{\delta_b P(V_2)}{\delta_a P(V_1)} = \theta \tag{13.6}$$

називають порогом.

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

$$\text{вектор } X \in V_1, \text{ якщо } \lambda(x_1,x_2,\ldots,x_n) > \theta \tag{13.7}$$

$$\text{вектор } X \in V_2, \text{ якщо } \lambda(x_1,x_2,\ldots,x_n) < \theta. \tag{13.8}$$

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

$$\text{вектор } X \in V_1, \text{ якщо } \lambda(x_1,x_2,\ldots,x_n) > 1,$$
$$\text{вектор } X \in V_2, \text{ якщо } \lambda(x_1,x_2,\ldots,x_n) < 1 \tag{13.9}$$

Отже розв'язувапьне правило базується на нерівності

$$\frac{P(x_1,x_2,\ldots,x_n / V_1)}{P(x_1,x_2,\ldots,x_n / V_2)} > \frac{\delta_b P(V_2)}{\delta_a P(V_1)}, \tag{13.10}$$

яка може бути представленою у логарифмичному виді

$$\ln \frac{P(x_1,x_2,\ldots,x_n / V_1)}{P(x_1,x_2,\ldots,x_n / V_2)} > \ln \frac{\delta_b P(V_2)}{\delta_a P(V_1)} \tag{13.11}$$

Функцію виду

$$F(x_1,x_2,\ldots,x_n) = \ln P(x_1,x_2,\ldots,x_n / V_1) - \ln P(x_1,x_2,\ldots,x_n / V_2) + \frac{\delta_a P(V_1)}{\delta_b P(V_2)} \tag{13.12}$$

називають дискримінантною функцією .

Якщо використовується дискримінантна функція, то розв'язувальне правило приймає вид:

$$X \in V_1, \text{якщо} F(x_1,x_2,\ldots,x_n) > 0; \tag{13.13}$$

$$X \in V_2, \text{якщо} F(x_1,x_2,\ldots,x_n) < 0; \tag{13.14}$$

Ясно, що рівняння $F(x_1,x_2,\ldots,x_n) = 0$ 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 множнії векторів-предікторів класів V_1 і V_2, отримання їх аналітичною виду - це дуже складна задача. Тому вважають, що вид законів розподілу є відомим. У такому разі задача зводиться до необхідності на основі вибірок векторів-предікторів отримати оцінки параметрів цих законів. Ця процедура носить назву відновлення закону розподілу. При практичних реалізаціях цих методів вважають найбільш часто, що класи векторів-предікторів підпорядковуються умовним нормальним законам розподілу. У дійсності ці припущення строго не виконуються. Дуже добре відомо, що для багатьох метеорологічних величин, які виступають у ролі предікторів, нормальний закон ошоділу 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*).

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

Будемо вважати, шо вектори-предіктори

$$X_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{nj} \end{pmatrix} \tag{13.15}$$

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

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix} \tag{13.16}$$

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

$$P(x_1, x_2, \dots, x_n / V_1) = \frac{1}{(2\pi)^{n/2} |K_1|^{1/2}} \exp\left[-\frac{1}{2} (X - \mu_1) K_1^{-1} (X - \mu_1) \right] \tag{13.17}$$

$$P(x_1, x_2, \dots, x_n / V_2) = \frac{1}{(2\pi)^{n/2} |K_2|^{1/2}} \exp\left[-\frac{1}{2} (X - \mu_2) K_2^{-1} (X - \mu_2) \right] \tag{13.18}$$

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

Після підстановки (13.17) і (13.18) до дискримінантної функції (13.12) прийдемо до такого рівняння:

$$\begin{aligned} F(x) = & -\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |K_1| - \frac{1}{2} (X - \mu_1)' K_1^{-1} (X - \mu_1) + \frac{n}{2} \ln 2\pi + \\ & + \frac{1}{2} \ln |K_2| + \frac{1}{2} (X - \mu_2)' K_2^{-1} (X - \mu_2) + \ln \frac{P(V)}{P(V_2)} \end{aligned} \tag{13.19}$$

Після скорочень дискримінантна функція набуде виду

$$F(x)=\frac{1}{2}\left[(X-\mu_2)K_2^{-1}(X-\mu_2)-(X-\mu_1)K_1^{-1}(X-\mu_1)+\ln\frac{|K_2|}{|K_1|}\right] \quad (13.20)$$
$$+\ln\frac{P(V_1)}{P(V_2)}$$

Вважається, що ціни помилок першого і другого роду однакові $\delta_a=\delta_b$. Дискримінантна функція F(x), яка визначається формулою (13.20), 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 (13.20) утримує 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 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:

$$K_1 = K_2 = K \ , \quad (13.21)$$

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

$$K = = \frac{K_1+K_2}{2} \quad (13.22)$$

Якщо прийняти умову (13.22) у дискримінантній функції (13.20) і вважати, що $P(V_1) = P(V_2)$, то отримаємо:

$$F(x)=\frac{1}{2}\left[(X-\mu_2)'K^{-1}(X-\mu_2)-(X-\mu_1)'K^{-1}(X-\mu_1)\right]= \quad (13.23)$$
$$=(\mu_1-\mu_2)'K^{-1}X+\frac{1}{2}\left[\mu_2'K^{-1}\mu_2-\mu_1'K^{-1}\mu_1\right]$$

Дискримінантна функція (13.23) є лінійною дискримінантною функцією.

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

Тоді

$$K = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix} \tag{13.24}$$

а її обернена матриця

$$K^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{pmatrix} \tag{13.25}$$

тобто операція обернення матриць коваріацій значно спрощується. При цьому спрощується й процедура розрахування коефіцієнтів квадратичних форм у *квадратичній дискримінантній функції (13.12).*

Коли виконується умова (13.21) і, крім того, можна вважати коваріаційну матрицю діагональною, значно спрощується вид лінійної дискримінантної функції (13.23). У цьому випадку вона дорівнює :

$$F(x) = \sum_{i=1}^{n} \frac{\mu_{1i} - \mu_{2i}}{\sigma_i^2} x_i + \sum_{i=1}^{n} \frac{\mu_{2i}^2 - \mu_{1i}^2}{\sigma_i^2} \tag{13.26}$$

Рівняння (13.26) є спрощеною дискримінантною функцією.

Звичайно, при практичному використанні розглянутих дискримінантних функцій замість складових векторів математичних сподівань і дисперсій предикторів 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

$$\Delta = (\mu_1 - \mu_2)' K^{-1} (\mu_1 - \mu_2), \tag{13.27}$$

а при використанні (13.26)

$$\Delta = \sum_{i=1}^{n} (\mu_{1i} - \mu_{2i})^2 / \sigma_i^2 . \qquad (13.28)$$

Чим більше число Махаланобіса, тим меньше 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:

$$D' = D_{F \geq 0} \qquad (13.29)$$

де D' – прогнозна дата появи досліджуваного явища; $D_{F \geq 0}$ – дата, на яку дискримінантна функція F за конкретних гідрометеорологічних умов стає рівною або більшою за нуль.

Оцінка точності прогнозу виконується шляхом розрахунку забезпеченості допустимої похибки прогнозу

$$p = \frac{n - m}{n} \cdot 100\% , \qquad (13.30)$$

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%.

ЛЕКЦІЯ 14

«Метод головних компонентів»

Питання.

3. *Розкладання полів випадкових величин у базисі власних векторів.*
4. *Базисне матричне рівняння.*
5. *Власні вектори, власні числа, компоненти.*
6. *Властивості власних векторів.*

14.1. Теоретичні основи методу головних компонентів

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

$$R_X - \lambda_i U_i = 0, \tag{14.1}$$

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

U_i - власний вектор матриці кореляцій;

λ_i - відповідне власному вектору власне значення.

Матриця R_X має m корінів або m власних чисел λ, які є дійсними, позитивними й простими. Для знаходження m власних векторів, що відповідають m власним числам, необхідне рішення 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

$$U' \cdot \varphi_i = Z_i, \tag{14.2}$$

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

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

Z_i - вектор головних компонентів, який є результатом лінійного перетворення φ_i – того поля відповідним власним вектором. Оскільки власні вектори ортонормовані, головні компоненти поля є статистично незалежними. Рівність (14.2) означає, що вхідне поле розкладене на m незалежних компонентів.

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

$$z_{ip} = \sum_{k=1}^{m} U_{pk} \varphi_{ik} \; ; \; p = \overline{1,m} \; , \tag{14.3}$$

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

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

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

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

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

$$\lambda_P = \sigma^2_{Zp}, \tag{14.4}$$

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

$$\sum_{p=1}^{m} \sigma^2_{Zp} = \sum_{p=1}^{m} \lambda_p = t_R R_x = m, \tag{14.5}$$

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

Питання.

1. *Виділення головних компонент.*
2. *Власні числа та дисперсії головних компонент та вихідних даних, зв'язок між ними.*
3. *Вирішення задачі стиснення вихідної інформації за методом головних компонентів.*

Якщо розкладання за ПОФ виконувати в базисі власних векторів вихідного центрованого поля ΔX_i, то (13.2) $U' \cdot \varphi_i = Z_i$ буде мати вигляд

$$W' \Delta X_i = Z_i, \tag{14.6}$$

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

$$\sum_{p=1}^{m} \sigma_{Zp}^2 = \sum_{p=1}^{m} \lambda_p = t_R K_x = \sum_{p=1}^{m} \sigma_{Xp}^2. \tag{14.7}$$

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 убуваючому 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 оцінюєтьсза 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. При використанні методу головних компонент вхідна інформація не тільки заміщається малим числом статистичних функцій, але й зберігає в цих функціях їх фізичне навантаження. Отримані в результаті розкладання функції є відображенням реальних фізичних процесів, які обумовлюють просторово-часовий розподіл досліджуваних гідрометеорологічних величин.

Питання:

1. *Базисні функції, їх фізична інтерпретація, визначення районів із синхронними коливаннями стоку.*
2. *Амлітудні функції, їх фізична інтерпретація.*
3. *Вирішення задачі фільтрації вихідних даних на основі методу головних компонентів.*

Аналізується кореляційна матриця 20 рядів річного стоку басейну р.Уссурі за період сумісних спостережень, починаючи з 1960 року і закінчуючи 1986 роком, тобто 27 років. Внесок перших компонент у загальну дисперсію становить: 63% для першої компоненти, 20% - для другої та 5% - для третьої (*Лобода Н.С., Нгуєн Ву Ань, 2006*).Оскільки для виявлення просторових закономірностей зміни базисних функцій перших компонент розкладання за природними ортогональними функціями необхідні дані про координати центрів тяжіння водозборів, досліджуваний водозбір був вкритий координатною "сіткою" й положення центрів тяжіння було представлене у вигляді умовних координат (рис.14.1).Знак вагових коефіцієнтів першої компоненти розкладання не змінюється (табл.14.1), що інтерпретується як однорідний вплив найбільш масштабного фізичного процесу на формування полів річного стоку. Зазвичай, перша компонента розкладання розглядається як статистичне відображення дії атмосферних процесів глобального масштабу. Вагові коефіцієнти другої компоненти змінюють знак, що у літературних джерелах (*Д.Л.Смірнов, В.Л. Скляренко,1986*) інтерпретується як існування різниці у закономірностях коливань гідрологічних характеристик на різних ділянках території.

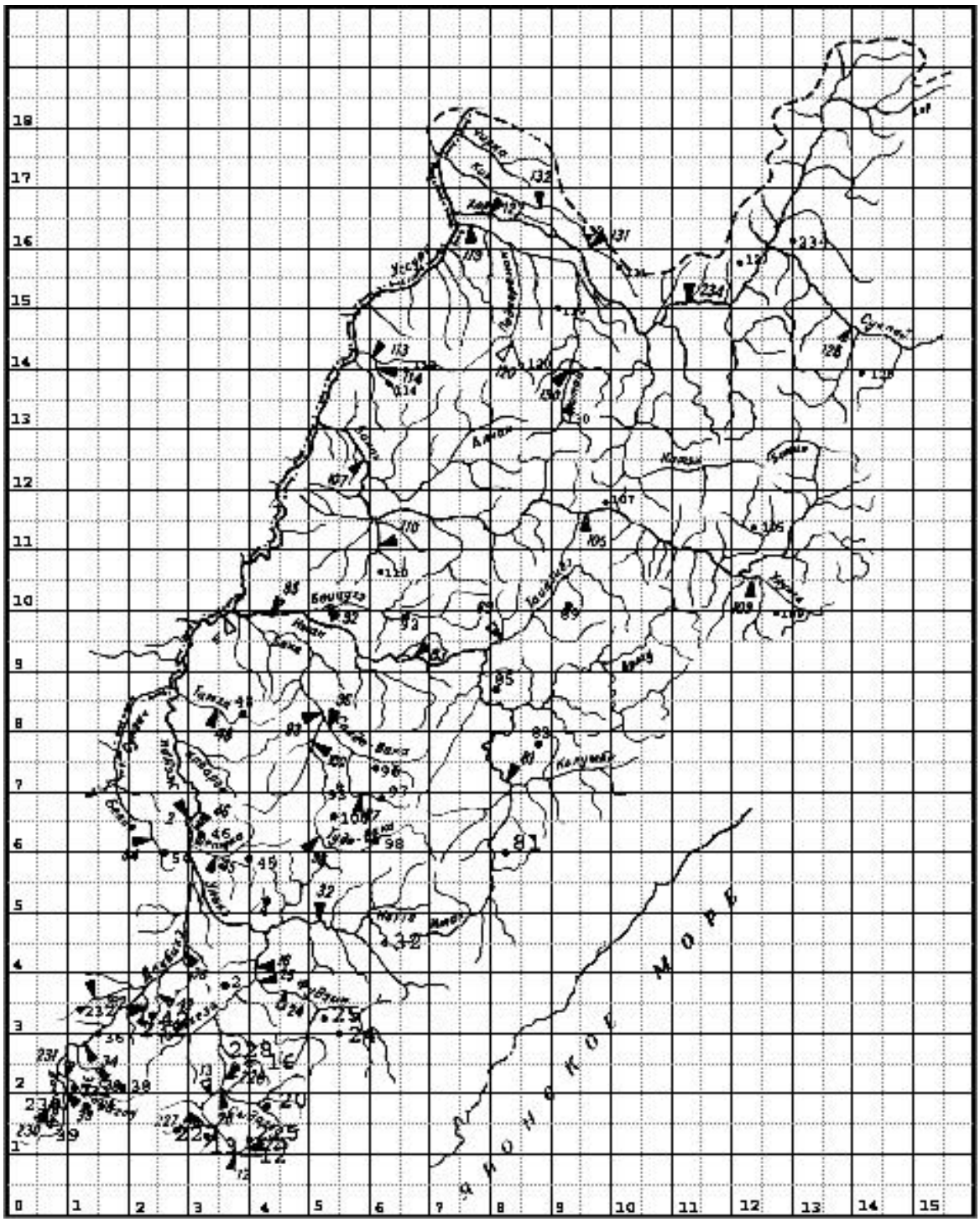


Рисунок 14.1 - Координатна сітка, гидрологічні пости та центри тяжіння водозборів для басейну р.Уссурі

Таблиця 14.1 - Значення перших базисних функцій

Номер за картою	Річка –пост	U_1	U_2	U_3
2	р.Уссурі-пос.Кировський	0,2582	0.2184	0,0398
12	р.Улахе-с.Березняки	0,2587	0.2592	0,4140
20	р.Сидагоу-с.Извілінка	0,2208	0,2684	0,3047
25	р.Фудзін-с.Уборка	0,2566	0,2193	0,1200
36	р.Даубіхе-с.Яковлівка	0,2266	0,2933	-0,0678
42	р.Хоніхеза-с.Варфоломєєвка	0,2465	0,2414	0,0768
45	р.Шетуха-с.Криловка	0,2359	0,1001	-0,3672
48	р.Тамга-с.Тамга	0,2446	0,0027	-0,3074
83	р.Іман-с.Картун	0,2317	-0,0645	-0,1489
85	р. Іман-пос. Вагутон	0,2406	-0,0745	-0,1916
89	р.Сібічі-с. Сібічі	0,2078	-0,2371	-0,1818
93	р.Вака-с.Ракітне	0,1819	0,0402	-0,2741
98	р.Тудо-Вака-с.Аріадне	0,2067	0,0836	-0,3054
107	р.Бікін-ст.Звєньєва	0,2290	-0,2414	0,0145
114	р.Горбун-с.Пушкіно	0,1943	-0.3282	0,1602
119	р.Подхорьонок-с.Дормідонтівка	0,1914	-0,3282	0,1762
120	р.Правий Подхорьонок – лзу. Медвежий Ключ	0,2159	-0,3268	0,1797
128	р.Сукпай-мет.ст.Сукпай	0,2337	-0,2188	0,2627
131	р.Кія-с.Марусіно	0,2038	-0,3472	0,1932
204	р.Каменка-пос. Каменський	0,1548	0,2030	0,1436

Інакше кажучи, другий за значущістю фізичний процес обумовлює прояв несинхронності у коливаннях характеристик стоку. Положення нульової ізолінії ($U_2 = 0$) розглядається як межа між районами з несинхронними коливаннями цих характеристик. Нульова ізолінія поділяє територію водозбору р. Уссурі навпіл (рис.14.2). Різницю у коливаннях стоку між північчю і півднем можна пояснити тим, що південна частина басейну р.Уссурі знаходиться у області дії субтропічних мусонів, а північна розташована у області впливу мусонів помірних широт. Тобто другий за значимістю фізичний процес може бути інтерпретований як атмосферний процес синоптичного масштабу.

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

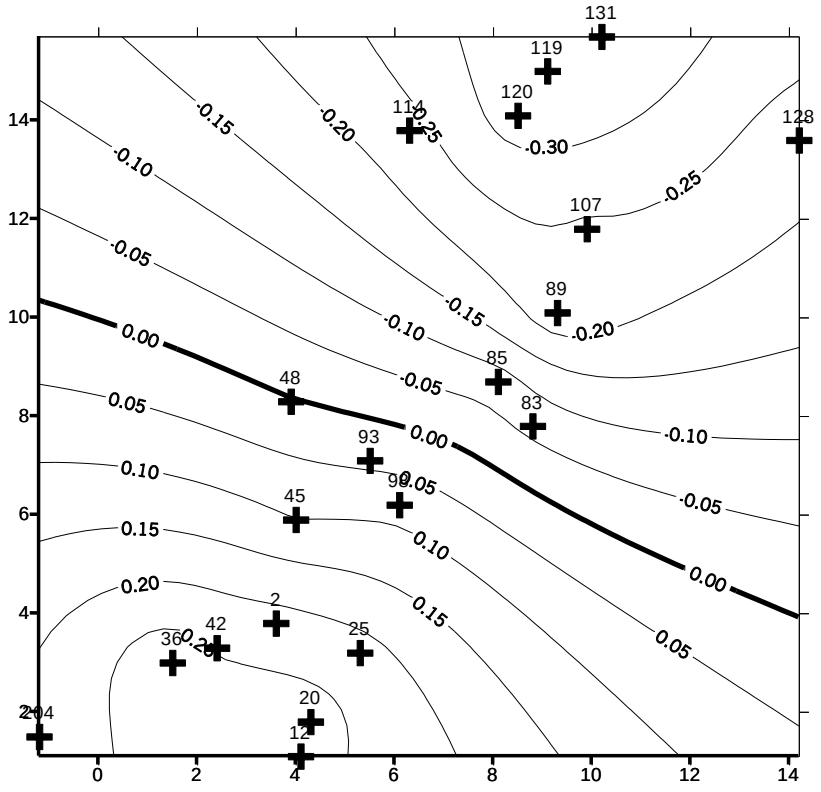


Рис. 14.2 Схема розподілу U_2 у басейні р.Уссурі

(+ - центри тяжіння водозборів)

ЛІТЕРАТУРА

Основна

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

2.Лобода Н.С. Методи просторового узагальнення гідрологічної інформації (Конспект лекцій). – Одеса. – Екологія, 2008 – 86 с.

3.Лобода Н.С. Методи статистичного аналізу у гідрологічних розрахунках і прогнозах. Навчальний посібник. – Одеса: Екологія. -2010. – 184 с.

4. http://library-odeky.16mb.com

Додаткова

1. Сніжко С.І. Оцінка та прогнозування якості природних вод: Підручник. – К.:Ніка-Центр, 2001. – 204 с.