Аналіз статистичних даних

- Форми та методи подання і попереднє статистичне опрацювання числових даних часових послідовностей
    - Попереднє опрацювання даних та подання результатів
    - Формування файлу даних
    - Графічне подання даних
    - Описова статистика – кількісні характеристики даних
    - Побудова гістограми
    - Побудова кумуляти
- Виявлення тенденції часового ряду методами згладжування
    - Методи згладжування часових рядів
    - Метод рухомого середнього
    - Метод зваженого рухомого середнього
    - Властивості рухомого середнього
    - Медіанна фільтрація
    - Нормування часових послідовностей
    - Критерії ефективності згладжування часових рядів
    - Формули для зваженого рухомого середнього
- Кореляційний аналіз часових послідовностей
    - Кореляційне поле
    - Коефіцієнт кореляції
    - Кореляційне відношення
    - Властивості кореляційного відношення
    - Кореляційна матриця
    - Автокореляція
    - Автокореляція в часових рядах
    - Розрахунок автокореляції
- Ієрархічний агломеративний кластерний аналіз багатовимірних даних
- Порядок роботи
- Хід роботи
- Форма звітності

Метою роботи є ознайомлення з основними методами візуалізації, графічного відображення та первинного статистичного опрацювання числових даних, які представлені вибірковою сукупністю або часовим рядом; ознайомлення з основними методами висвітлення поведінки досліджуваного показника, яка подана характером його тренду, за допомогою методів згладжування часових рядів та подання отриманих результатів (візуалізація експериментальних даних - отриманих результатів) засобами табличного процесора **MS Excel (на задовільно)**, на основі мови **Python (добре)** або мови **R (відмінно)**. Метою роботи також є ознайомлення з методами кореляційного аналізу експериментальних даних, які подані часовими послідовностями. Для цього потрібно:

– побудувати кореляційне поле;
– визначити значення коефіцієнта кореляції;
– обчислити кореляційне відношення;
– побудувати графіки автокореляційних функцій;
– розбити одну з послідовностей на три рівні частини;
– побудувати для них кореляційну матрицю;

– знайти коефіцієнти множинної кореляції.

Необхідно розділити задану множину об’єктів, кожен з яких характеризується однаковою сукупністю конкретних ознак, на окремі групи, використовуючи ієрархічний агломеративний кластерний аналіз.

1. Форми і методи подання та попереднє статистичне опрацювання числових даних часових послідовностей

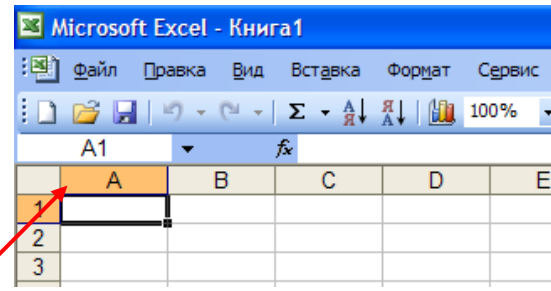
1.1. Попереднє опрацювання даних та подання результатів

Отримані в різних дослідженнях дані переважно характеризують спостережуваний об’єкт в двох аспектах – статичному й динамічному. Статичний аспект дає характеристику об’єкта значеннями конкретних величин, які можуть бути як реальними, тобто характеризують об’єкт таким як він є (кількість елементів конкретної схеми, значення елемента за даною шкалою) або випадковими – з відомим або невідомим законом розподілу їх ймовірності. Динамічний аспект характеризує дані про об’єкт протягом часу, регулярно або нерегулярно, але так, щоб кожне значення було прив’язане до моменту часу його спостереження чи реєстрації. Ці два аспекти виражаються з точки зору їх опрацювання двома класами: вибірковими сукупностями – вибірками і часовими послідовностями або часовими рядами. Для першого класу результатом опрацювання є визначення закону розподілу випадкових значень елементів вибірки – варіант. Зауважимо, що невипадкові значення не потребують знаходження виду та параметрів закону розподілу, а використовуються безпосередньо в розрахунках або їх перед цим усереднюють. Для другого класу, який є різновидом випадкових процесів, результатом опрацювання є аналітичне подання тенденції розвитку досліджуваного показника в часі.

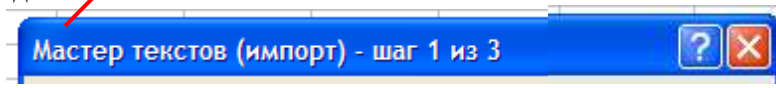
Завершенням таких досліджень є математична модель, в якій ці дві компоненти одного і того ж самого набору даних (вважаємо, що дані прив’язані до часу або до іншого показника) об’єднані аналітично, а сама модель, з точністю до прийнятого критерію адекватності, описує їхню поведінку. Попереднє опрацювання даних фактично дає підстави для побудови такої математичної моделі, проте для цього мають бути реалізовані певні етапи. Для досягнення поставленої мети в цій роботі необхідно забезпечити поетапне виконання низки завдань. Тобто, вирішення цих завдань здійснюється наступними етапами.

1.2. Формування файлу даних

Для цього, отримані дані формуються у файл даних, тобто файл, в якому дані відповідним чином впорядковані (наприклад: в послідовності їх отримання, у відповідності з часом отримання). Такий файл має мати форму таблиці «об’єкт - властивість», «порядковий номер – значення показника». Як правило такий файл є звичайним комп’ютерним файлом, який може бути сформований за допомогою Word або навіть побудований в «Блокноті», але він має мати відповідне розширення, наприклад, *.dat або *.txt. Після формування такого файлу його необхідно внести в книгу Excel. Для цього необхідно у відкритій книзі Excel відкрити доступ до віддалених файлів.



Далі, за допомогою



, вказавши формат вихідних даних і тип розділювача, виконати відповідні кроки. В результаті в книзі буде введено вказаний файл. Для отриманих експериментальних даних може виявитися значний розмір стовпчиків, що у випадку внесення такої таблиці у сторінку Word спричинить незручність сприйняття даних, оскільки дані можуть вимагати декілька незаповнених сторінок. Тоді, отриману таблицю варто «стиснути», подавши її у такій формі. Подання даних у формі стиснутої таблиці здійснюють за допомогою розбиття оригінальної на кількість частин так, щоб в таблиці було як найменше вільних комірок, а сама таблиця не виходила за межі, що визначені текстовим форматом. Не слід дуже ущільнювати стовпчики таблиці, оскільки дані повинні бути читабельними за значеннями, номери та їх значення не повинні бути дуже близько. Це має важливе практичне значення, особливо, коли дані представляються для аналізу в роздрукованому, але без електронного супроводу, представленні. Такі дані для здійснення аналізу сканують і порушення цієї вимоги може призвести до суттєвих помилок.

Таблиця 1.

Подання значень даних у формі стисненої таблиці

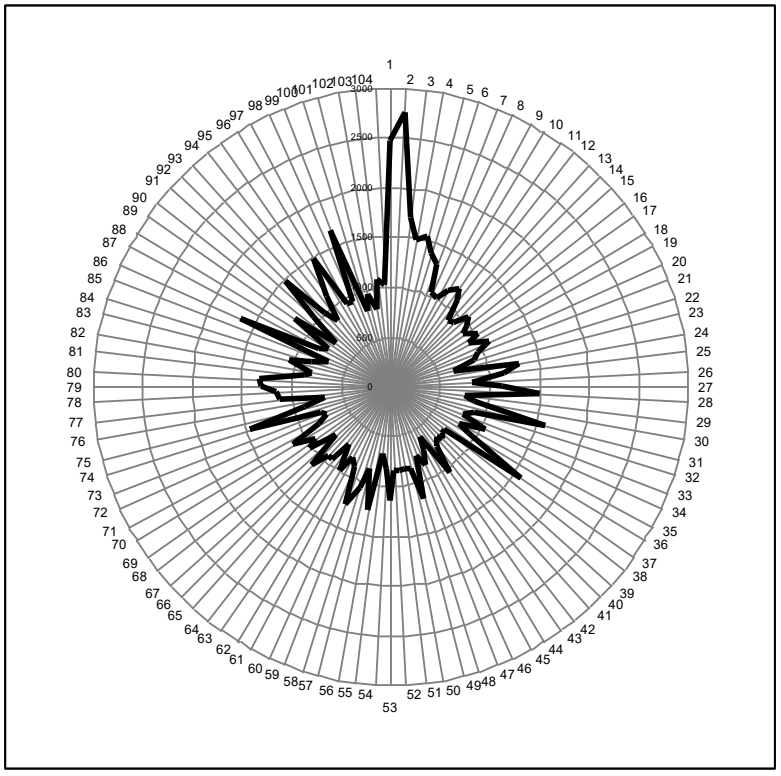
РЕЗУЛЬТАТИ РЕЄСТРАЦІЇ ДИНАМІКИ ПОКАЗНИКА									
№ п/п	Значення показника	№ п/п	Значення показника	№ п/п	Значення показника	№ п/п	Значення показника	№ п/п	Значення показника
1	**2465**	21	**933**	41	**690**	61	**800**	81	**871**
2	**2738**	22	**885**	42	**748**	62	**971**	82	**803**
3	**1698**	23	**663**	43	**1042**	63	**711**	83	**1043**
4	**1482**	24	**1321**	44	**609**	64	**934**	84	**819**
5	**1544**	25	**1148**	45	**717**	65	**937**	85	**651**
6	**1386**	26	**826**	46	**856**	66	**1111**	86	**1654**
7	**1315**	27	**1118**	47	**741**	67	**733**	87	**806**
8	**1032**	28	**1497**	48	**1183**	68	**1006**	88	**725**
9	**998**	29	**757**	49	**836**	69	**921**	89	**1160**
10	**1119**	30	**865**	50	**836**	70	**1140**	90	**695**
11	**1203**	31	**1605**	51	**843**	71	**809**	91	**1039**
12	**1107**	32	**872**	52	**841**	72	**686**	92	**1482**
13	**889**	33	**784**	53	**1150**	73	**744**	93	**1001**
14	**907**	34	**1055**	54	**819**	74	**1462**	94	**835**
15	**1043**	35	**786**	55	**683**	75	**876**	95	**1098**

16	**968**	36	**913**	56	**1265**	76	**683**	96	**1489**
17	**895**	37	**1597**	57	**841**	77	**1108**	97	**935**
18	**1030**	38	**695**	58	**1047**	78	**1146**	98	**915**
19	**893**	39	**716**	59	**1271**	79	**1295**	99	**1675**
20	**1089**	40	**756**	60	**882**	80	**1320**	100	**779**

1.3. Графічне подання даних

Основним типом візуалізації даних у звітах експериментальних та науково-практичних досліджень є графіки. Графік відображає відношення між двома величинами, одна з яких є незалежною змінною і її значення, зазвичай, відкладають вздовж горизонтальної осі – *абсциси*; друга змінна є залежною і її значення відкладають вздовж вертикальної осі – *ординати*. Фактично, будь-який графік складається з декількох основних елементів: осей з надписами, експериментальних точок, ліній, що з'єднують ці точки (їх ще називають *кривими*), пояснюючих написів на рисунку та підпису під рисунком.

а



б

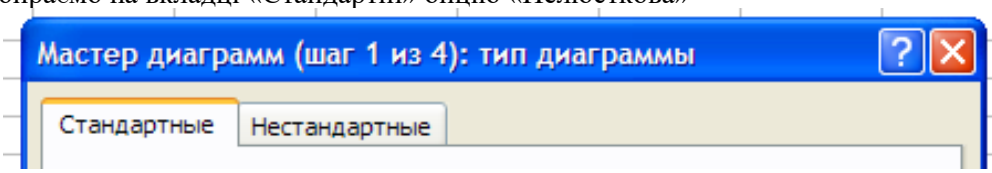
Рис. 1. Графічне подання динаміки показника: *а* – в декартовій системі координат, *б* – в полярній системі координат.

В загальному, щоб побудувати графік необхідно:

– підготувати таблицю з відображуваними даними;
– визначити розміри осей з реперними штрихами і цифрами;
– нанести експериментальні точки і провести криві;
– зробити відповідні написи для осей та шкал на рисунку;
– підготувати підпис до рисунка.

Наявність табличного процесора суттєво спрощує процедуру побудови графіка, переважно з геометричної точки зору, тобто креслення його за точками, в той час, як сама підготовка та редакція й корегування його зовнішнього вигляду залишаються прерогативою фахівця.

На графіку мають бути позначені осі, виділені маркери, має бути назва графіка, назви осей, значення поділок повинно відповідати значенням варіант. Для побудови графіка в полярній системі координат використовуємо «Майстер діаграм» і вибираємо на вкладці «Стандартні» опцію «Пелюсткова»



Редагування графіка – вибір розміру шрифту, параметрів та кольору осей здійснюємо звичайним способом.

1.4. Описова статистика – кількісні характеристики даних

Крім табличного та графічного представлення даних, в їх супровід включають загальні числові та статистичні характеристики, які відносять до описової або дескриптивної статистики.

Описова статистика дає підстави для формування компетенцій щодо вибору шкали вимірювань, автоматизації опрацювання даних при застосуванні різних форматів на етапі їх збору, подання результатів у різних формах, графічного подання результатів, обчислення статистичних параметрів розподілу та оцінки параметрів генеральної сукупності з використанням інформаційних технологій. Вона займається вибором кількісної інформації, яка необхідна (або цікава) для різних людей. Великі масиви даних, перш ніж вони вивчатимуться людиною, мають узагальнюватися або згортатися. Саме це робить описова статистика, яка описує, узагальнює або зводить до бажаного вигляду властивості масивів даних. Описова статистика застосовується для аналізу та інтерпретації статистичних даних, побудови статистичних розподілів та обчислення відповідних числових параметрів, що характеризують досліджувану сукупність. Її використовують для організації збирання інформації, перевірки якості даних та їхньої інтерпретації, зображення статистичного матеріалу.

Описова (дескриптивна) статистика – це найбільш загальні статистичні показники, що описують розподіл даних, приймаючи за норму нормальний розподіл. Це тому, що характерною властивістю нормального розподілу є те, що 68% всіх його спостережень лежать в діапазоні $\bar{x} \pm$ одне стандартне відхилення від середнього арифметичного, а в діапазоні $\bar{x} \pm$ два стандартних відхилення міститься 95% значень вибірки з нормальним розподілом. Ці два показники – середнє арифметичне і стандартне відхилення є основними параметрами нормального розподілу.

Середнє арифметичне є мірою центральної тенденції, що відображає найбільш характерне для даної вибірки значення. Його визначають за формулою

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad (1)$$

де n – обсяг вибірки.

Оманливість цього показника, проілюструємо таким прикладом: в одному купе вагона розмістилася бабуся 60 років з чотирма онуками: одному – 4 роки, двом – по 5 років й одному – 6 років. Середнє арифметичне віку всіх пасажирів цього купе 80/5=16. В іншому купе розташувалася компанія молоді: двоє –15-тилітніх, один – 16-річний та двоє – 17-річних. Середній вік пасажирів цього купе так само дорівнює 80/5=16. Таким чином, за середнім арифметичним пасажири цих купе не відрізняються. Але, якщо звернутися до показника стандартного відхилення, то виявиться, що середній розкид щодо середнього віку в першому випадку виявиться 24,6, а в другому випадку – 1.

Мода (позначається «*Mo*») – це значення, яке найбільш часто зустрічається серед вибірки змінних. Часто застосовується для непараметричних даних і для рангових шкал.

Медіана (позначається «*Me*») – значення, яке ділить навпіл впорядковану множину змінних, тобто для визначення медіани необхідно впорядкувати дані, наприклад, за зростанням. Способи визначення значення медіани для парної і непарної кількості даних відрізняються. Для непарної кількості даних визначають її номер у впорядкованій сукупності за такою формулою:

$$Me(n) = \frac{n+1}{2}. \qquad (2)$$

У випадку парної кількості даних визначають номери двох сусідніх серединних значень $\frac{n}{2}$ і $\frac{n+2}{2}$. Середнє арифметичне цих двох значень вибірки (з номерами $x\left(\frac{n}{2}\right)$ і $x\left(\frac{n+2}{2}\right)$) і буде значенням медіани для парного *n*, тобто:

$$Me(n) = \frac{Me\left(\frac{n}{2}\right) + Me\left(\frac{n+2}{2}\right)}{2}. \qquad (3)$$

Значення середнього, моди та медіани є близькими один до одного. В ідеальному нормальному розподілі вони рівні, оскільки мають однаковий зміст: середина розподілу.

Розмах (інтервал) – показник, який вказує на ширину діапазону значень. Він дорівнює різниці між максимальним і мінімальним значеннями.

Стандартне відхилення (σ, читається «сигма») – є мірою мінливості (варіації) ознаки, яка відображає величину його розкиду відносно середнього арифметичного

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}. \qquad (4)$$

Для більш точного та наочного представлення про варіацію значень показника відносно середнього використовується коефіцієнт варіації:

$$\nu = \frac{\sigma}{\bar{x}} \cdot 100\% . \qquad (5)$$

Коефіцієнт варіації виражає міру мінливості ознаки у відсотках.

Асиметрія – показник, що відображає перекіс розподілу відносно моди вліво або вправо. Це має місце в тих випадках, коли якісь причини сприяють більш частій появі значень, які є більшими або, навпаки, меншими за середнє арифметичне. Для лівосторонньої або додатньої асиметрії в розподілі частіше зустрічаються більш низькі значення, а для правосторонньої або від'ємної – вищі.

Ексцес – показник, що відображає висоту розподілу. У тих випадках, коли якісь причини сприяють появі близьких до середніх значень, утворюється розподіл з додатнім ексцесом. Якщо ж в розподілі переважають крайні значення, причому одночасно і більш низькі, і більш високі, то такий розподіл характеризується від'ємним ексцесом, і в центрі розподілу може утворитися впадина, яка перетворює його в двохвершинний.

Описова статистика включає в себе табулювання (складання таблиць), подання та опис сукупності даних. Засіб аналізу «Описова статистика» використовується для створення одномірного статистичного звіту, який містить інформацію про центральну тенденцію та мінливість даних початкового діапазону.

За допомогою опції «Описової статистики» отримують такі кількісні дані: середнє арифметичне, стандартну помилку, медіану, моду, стандартне відхилення, дисперсію, ексцес, асиметрію, розмах, мінімальне та максимальне значення, суму значень, їх кількість та рівень надійності.

Ці характеристики як правило використовують при первинній ідентифікації даних, тобто їхнього порівняння і виявлення змін. Табличний процесор MS Excel враховує необхідність такого подання характеристик даних і забезпечує процедури їх визначення за допомогою операторів, що зведені нижче в таблицю.

Результати описової статистики	
Показники	Значення
Середнє	932,7837838
Стандартна помилка	41,34955245
Медіана	841
Мода	836
Стандартне відхилення	251,5195083
Дисперсія вибірки	63262,06306
Ексцес	1,371916976
Асиметричність	1,384445279
Інтервал	996
Мінімум	609
Максимум	1605
Сума	34513
Обсяг	37
Рівень надійності (95,0%)	83,86077869

а

Результати описової статистики	
Показники	Значення
Середнє	932,78
Стандартна помилка	41,35
Медіана	841,00
Мода	836,00
Стандартне відхилення	251,52
Дисперсія вибірки	63262,06
Ексцес	1,37
Асиметричність	1,38
Інтервал	996,00
Мінімум	609,00
Максимум	1605,00
Сума	34513,00
Обсяг	37,00
Рівень надійності (95,0%)	83,86

б

Рис. 2. Результати описової статистики: *а* – безпосередньо отримані, *б* – підготовлені до звіту (кількісні дані подані з двома розрядами).

Подання числових значень має бути приведено до відповідної розрядної сітки.

1.5. Побудова гістограми

Якщо отримані дані утворюють звичайну репрезентативну вибірку, тобто вибірку, обсяг якої є достатнім для визначення їхнього розподілу, традиційно, для визначення його вигляду будують гістограму. Гістограма є дуже наближеним відображенням графіка функції щільності закону розподілу даних цієї вибірки. Вона також є діаграмою, яка наочно відображає метод групування даних за деякою істотною ознакою. Методи групування широко використовуються для початкового опрацювання даних, оскільки вони суттєво зменшують обсяг даних і виявляють найбільш характерну для них структуру – частоту значень, розподіл в групах, а в більшості випадків характер розподілу: одно чи багато модальність, положення та наближене значення моди як середини модового інтервалу, наближений вигляд модового інтервалу. Найбільш проблематичним у побудові гістограми є вибір кількості інтервалів групування, тобто кількості груп, на які розбивається вибірка. Для вибору чи визначення кількості інтервалів розбиття існує декілька десятків формул, проте найбільш поширеними є:

формула Стерджеса $k = 1 + \log_2 n$, k – кількість інтервалів;

формула Скотта $h = 3.5 \cdot s \cdot n^{-1/3} = \dfrac{3.5 \cdot s}{\sqrt[3]{n}}$,

де h – ширина інтервалу, s – стандартне відхилення значень ряду.

Побудова гістограми полягає в реалізації наступних кроків:

1. За обсягом вибірки (кількістю даних) n та стандартним відхиленням s визначають кількість або ширину інтервалів. Далі визначають значення меж інтервалів («Інтервал кишень»).

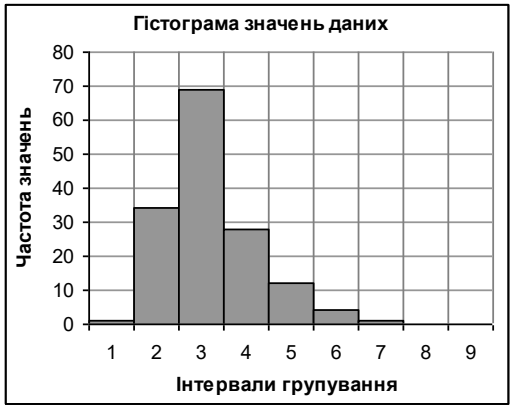
2. Послідовно «книга Excel → Сервіс → Аналіз даних → Гістограма» заповнюють дані в Excel комірках Майстра гістограми,

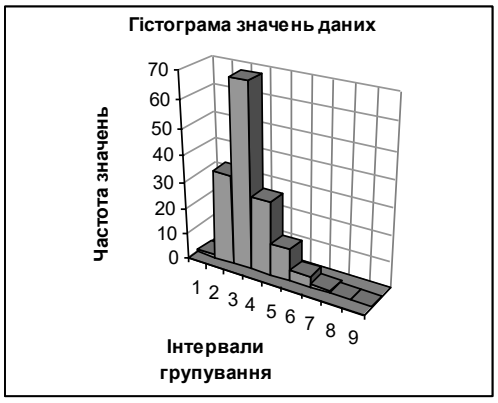
Рис. 3. Майстер побудови гістограми.

а саме:

- у «Вхідний інтервал» вносять дані вибірки;
- в «Інтервал кишень» вносять значення меж інтервалів;
- вказують комірку для «Вихідного інтервалу»;
- відмічають потрібні результати: «Інтегральний процент» та «Вивід графіку» і «ОК».

За даними стовпчика «Частота» в таблиці отриманого результату будують, використовуючи «Майстер діаграм» і опцію «Гістограма», гістограму отриманих даних. На рис. зображені двомірне та трьохмірне подання гістограми.





На підставі зображення гістограми здійснюють аналіз частот значень на предмет апроксимації емпіричної щільності закону розподілу вихідних даних – отриманої гістограми деякою аналітичною функцією. Як таку функцію часто вибирають функцію закону розподілу серед відомих функцій розподілів випадкових величин або використовують відповідні інші математичні функції.

1.6. Побудова кумуляти

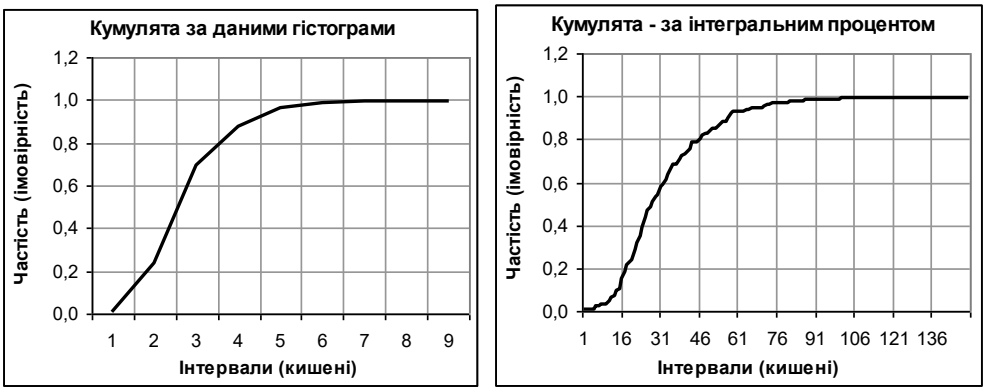
Основними недоліками гістограми є такі: суб'єктивність у виборі кількості інтервалів групування, значення середнього арифметичного, моди і медіани подаються з точністю до розміру інтервалу, обмежена кількість вузлів апроксимації. Проте, деякі з цих недоліків можна усунути використовуючи метод побудови гістограми для побудови *кумуляти* (емпіричного графіка функції закону розподілу), якщо зробити такі зміни.

1. Встановити кількість інтервалів рівною $n-1$, що виключає суб'єктивність у виборі потрібної формули.

2. Побудувати гістограму для такої кількості інтервалів (будуть присутні порожні інтервали, а тому інтерпретація такої гістограми є специфічною).

3. За даними стовпчика «Інтегральний процент» в таблиці отриманого результату будують, використовуючи «Майстер діаграм» і опцію «Графік», графік емпіричної кумуляти (емпіричної функції закону розподілу отриманих даних).

Емпіричну кумуляту можна побудувати і на підставі гістограми, оскільки, оскільки в таблиці її побудови також є стовпчик з даними про інтегральний процент. Відмінність цих двох кумулят в тому, що кумулята побудована за даними гістограми має k вузлів апроксимації і є ломаною кривою, а кумулята побудована за інтегральним процентом має $n-1$ вузлів апроксимації і є більш плавною. На рис. зображені обидві кумуляти.



а *б*

Рис. 4. Кумуляти побудовані: *а* – за даними гістограми;
б – за інтегральним процентом.

Розглянутий підхід до попереднього опрацювання результатів досліджень дає підстави для подання, як самих результатів, так і їхніх основних характеристик, в достатньо інформативній наочній формі – графіки, діаграми та кількісній формі – показниках описової статистики. Результати такого підходу можуть бути достатніми для приведення їх у наукових звітах, дипломних та магістерських кваліфікаційних робота, а також в дисертаційних дослідженнях.

2. Виявлення тенденції часового ряду методами згладжування

При побудові математичних моделей часових рядів, переважно з метою визначення динаміки показника, необхідно в першу чергу виділити тенденцію та відокремити її від випадкових відхилень, зумовлених різними перешкоджаючими факторами. Дослідження часового ряду починається з його графічного представлення. При візуальному способі будується графік часового ряду, на основі якого висувається гіпотеза про його структуру, в першу чергу про форму тренду. Цей підхід дає задовільні результати при відносно монотонних тенденціях. Однак, у випадку значних флуктуацій модельованого процесу можливість помилок у виборі виду функції тренда при цьому підході зростає. Методи виявлення основної тенденції розвитку досліджуваного об’єкта переважно визначаються на основі докладного вивчення фактичного розвитку його динаміки. Вони повинні узгоджуватись з результатами спостережень і статистикою емпіричного матеріалу. Ці методи мають різну логічну змістовність і тому застосовуються до часових рядів в залежності від цілей дослідження. Основна мета полягає в тому, щоб розкривати загальні закономірності розвитку, затушовані окремими, іноді випадковими обставинами. Проте кожен з них має свої особливості. Для виявленні тенденції – характеру розвитку використовують процедуру згладжування часового ряду. Суть її зводиться до заміни фактичних рівнів часового ряду розрахунковими, але з меншими коливаннями, що сприяє більш чіткому проявленню тенденції та її характеру. Саме в цьому випадку, тенденцію зображають гладкою неперервною функцією, яку або її графік називають ***трендом*** часового ряду.

2.1. Методи згладжування часових рядів

Методи згладжування можна умовно розділити на два класи, в основі яких

лежать різні підходи: аналітичний та алгоритмічний.

Аналітичний підхід оснований на припущенні, що дослідник може на підставі візуального аналізу задати загальний вигляд функції, вважаючи що її графік відповідає характеру тенденції. Наприклад, на основі візуального та змістовного аналізу властивостей об'єкта та динаміки часового ряду, поведінку якого він описує, як функція може бути використана: експонента, гіпербола, парабола, степенева функція, поліноми вищих степенів та інші функції.

Наступний етап передбачає аналітичне або статистичне оцінювання статистичних невідомих параметрів вибраної функції, яка в цьому випадку стає математичною моделлю тенденції даного часового ряду.

Іншими словами, аналітичний підхід означає заміну значень рівнів часового ряду значеннями, що теоретично розраховані на підставі явного аналітичного вигляду функції, якою апроксимують візуально визначений тренд.

Цей підхід успішно реалізований в Excel, використовуючи поліноми до 6-го степеня включно, а також степеневу, експонентну та логарифмічну функції.

В **алгоритмічному підході** вигляд тренду отримують за рахунок різних алгоритмів, які практично реалізують процедури згладжування. Ці процедури надають досліднику лише алгоритм розрахунку нового значення часового ряду в будь-який заданий момент часу t.

Ці методи можна класифікувати так:

– просте або звичайне ковзне (рухоме) середнє;
– зважене ковзне (рухоме) середнє;
– експоненціальне згладжування;
– медіанне згладжування.

Найбільш вживаними є методи згладжування часових рядів за допомогою ковзних середніх.

Метод ковзних (рухомих) середніх є одним із найстаріших відомих способів згладжування часового ряду. Він базується на переході від початкових значень ряду до їх середніх значень на інтервалі часу, довжина якого обрана заздалегідь. При цьому сам вибраний інтервал часу ковзає вздовж ряду.

Суть цього методу зводиться до заміни фактичних рівнів ряду послідовностями рівнів, що мають, як правило, значно менші коливання, ніж вихідні дані. Зменшення флуктуації дає можливість наочно виявити основну тенденцію. Часто таку операцію над вихідними даними називають фільтруванням, а оператор її здійснення фільтром. У зарубіжних статистичних пакетах для цих процесів вживають абревіатуру MA – від англійського ковзної середньої (рухоме середнє).

Ковзні середні дозволяють згладити як випадкові, так і періодичні коливання, виявити наявну тенденцію в розвитку процесу і є важливим інструментом при фільтрації компонент часового ряду. Вони можуть бути визначені за допомогою простих ковзних або зважених середніх.

Вибір методу виявлення основної тенденції розвитку залежить від технічних можливостей обчислень та від уміння застосовувати відповідні методи, а також від завдань, які стоять перед дослідженням. Якщо треба дати загальну картину розвитку, його грубу модель, що основана на механічному повторенні одних і тих же дій крок за кроком до послідовності рівнів, то можна обмежитися методом ковзної середньої. Якщо ж мета дослідження полягає в розробці математичної моделі тренду, то сам метод ковзної середньої буде недостатнім. Тоді треба буде використовувати метод кінцевих різниць або метод найменших квадратів.

2.2. Метод рухомого середнього

Метод рухомих (ковзних) середніх дає оцінку середнього рівня за деякий період часу. Чим більше інтервал часу, до якого належить середня, тим більше згладжуватиме рівень, але тим менш точно буде описана тенденція вихідного ряду динаміки. Одержуваний таким чином ряд ковзних середніх поводиться набагато більш гладко, ніж вихідний ряд, за рахунок усереднення відхилень вихідного ряду. Таким чином, ця процедура дає уявлення про загальну тенденцію поведінки ряду. Її застосування особливо корисне для рядів з сезонними коливаннями і незрозумілим характером тренду. Зокрема, перехід до ряду ковзних середніх може бути використаний для виявлення коливальної компоненти часового ряду.

Згладжування ряду динаміки за допомогою ковзної середньої полягає в тому, що обчислюється середній рівень з певного числа перших за порядком рівнів ряду. Ці перші рівні ряду утворюють інтервал постійної величини, в якому рівні розташовані в тому ж порядку, що й у часовому ряді. Заміна першого і останнього наступними сусідніми рівнями за незмінної величини інтервалу створює враження руху цього інтервалу вздовж часового ряду. Оскільки, після кожного кроку, для рівнів охоплених цим інтервалом обчислюється середнє значення, то його використовують для утворення нового ряду. Причому, це обчислене значення в новому ряді відповідає тому моменту часу, якому відповідає значення рівня в середині ковзаючого інтервалу. Для забезпечення такої часової відповідності інтервал має охоплювати непарну кількість рівнів, тобто $w = 2n + 1$, де $n = 1, 2, 3, \ldots$. Таким чином, при обчисленнях середнього рівня межі інтервалу ніби «ковзають» по рівнях часового ряду від його початку до кінця, щоразу відкидаючи один рівень на початку інтервалу і додаючи в кінці наступний. Звідси назва – ***ковзна середня***.

Алгоритм для обчислення простої ковзної середньої є такий

$$\tilde{y}_i = y_1^* + y_2^* + \ldots + y_k^* + \sum_{j=k+1}^{N-2k} \left[\frac{1}{w} \sum_{i=j}^{j+2k+1} y_i \right] + y_{N-k}^* + \ldots + y_{N-1}^* + y_N^* \quad . \qquad (6)$$

Визначення ковзної середньої у випадку парного числа рівнів в ковзаючому інтервалі ускладнюється тим, що тоді середня має бути віднесена тільки до середини між двома моментами часу, що знаходяться всередині інтервалу згладжування, а в такий момент часу спостереження не проводились.

Якщо графічне представлення часового ряду нагадує пряму лінію, то в цьому випадку ковзна середня не спотворює динаміку досліджуваного явища.

У випадку, коли тенденція вихідного ряду, що характеризує досліджуваний процес, не може бути описана лінійним трендом, більш надійним є використання зваженої ковзної середньої.

2.3. Метод зваженого рухомого середнього

Поряд з простими ковзними середніми застосовуються також поліноміальні або зважені середні. Вони дозволяють точніше описати початок форми основну тенденцію ряду, оскільки при обчисленні зваженої середньої кожному рівню ряду в межах інтервалу згладжування приписується певна вагу, що залежить від відстані до середини інтервалу. При побудові зваженої ковзної середньої на кожній активній ділянці значення центрального рівня замінюється на обчислене, що визначене за формулою зваженої середньої арифметичної. Іншими словами, зважена змінна середня відрізняється від простої ковзної середньої тим, що рівні, що входять в інтервал усереднення, підсумовуються з різними вагами.

Проста ковзна середня враховує всі рівні ряду, що входять до інтервалу згладжування, з рівними вагами, а зважена середня приписує кожному рівню вагу, що залежить від відстані цього рівня до рівня, що стоїть всередині цього інтервалу. Це зумовлене тим, що для простої ковзної середньої в інтервалі згладжування обчислення здійснюють на підставі прямої – поліному першого порядку, а для згладжуванні зваженою ковзною середньою використовують поліноми більш високих порядків, переважно другого або третього. Тому метод простої ковзної середньої можна розглядати як окремий випадок методу зваженої ковзної середньої.

Вагові коефіцієнти визначаються за допомогою методу найменших квадратів (МНК), причому немає необхідності щоразу обчислювати їх заново для рівнів ряду, що входять в інтервал згладжування, так як вони будуть однаковими для кожного його положення. Так як ваги симетричні відносно центрального рівня, то в існуючих таблицях використано символічний запис: наведено ваги лише для половини рівнів інтервалу згладжування, причому виділено вагу, що відноситься до рівня, який стоїть у центрі ділянки згладжування. Для решти рівнів ваги не наводяться, тому що вони можуть бути симетрично відображені.

Основними властивостями вагових коефіцієнтів є такі:

1) ваги завжди симетричні відносно центрального рівня;

2) сума ваг в інтервалі згладжування дорівнює одиниці;

3) додатні та від'ємні значення ваг забезпечують для згладженої кривої можливість відтворювати різні вигини кривої тренду.

Алгоритм згладжування зваженим ковзним середнім з розміром „вікна" – інтервалу згладжування $w = 2k + 1$, яке послідовно зміщується вздовж рівнів ряду і усереднює охоплені ним рівні y_i

$$\tilde{y}_i = y_1^* + y_2^* + \ldots + y_k^* + \sum_{j=k+1}^{N-2k} \left[\frac{1}{w} \sum_{i=j}^{j+2k+1} \alpha_i y_i \right] + y_{N-k}^* + \ldots + y_{N-1}^* + y_N^* \quad . \quad (7)$$

В цих формулах значення y_q^*, де $q = 1, 2, \ldots, k, N-k, \ldots, N-1, N$, на початку та в кінці ряду, обчислюються за відповідними інтерполяційними формулами, α_i – ваги, для яких $\sum_{i=1}^{2k+1} \alpha_i = 1$.

2.4. Властивості рухомого середнього

Ковзне середнє, згладжуючи вихідний ряд, дає уявлення про загальну тенденцію поведінки ряду: його тренд і циклічну компоненту, а тому потрібно знати його властивості, до яких відносять наступні.

1. При застосуванні методу ковзних середніх вибір величини інтервалу згладжування повинен бути зроблений на основі змістовних міркувань і бути прив'язаним до періоду можливо існуючих коливальних процесів. Якщо процедура ковзного середнього використовується для згладжування часового ряду за відсутності будь яких коливань, то найчастіше величину інтервалу згладжування вибирають рівною трьом, п'яти або семи. Чим більшим є інтервал усереднення (згладжування), тим більш гладким виглядає графік тренду.

2. Сусідні члени ряду ковзних середніх сильно корельовані, так як в їх формуванні беруть участь одні й ті ж члени вихідного ряду. Це може призвести, до

того, що ряд ковзних середніх може містити циклічні компоненти, які відсутні у вихідному ряді. Це явище носить назву *ефекту Слуцького-Юла*.

3. Як метод усереднення, крім згаданого вище звичайного середнього арифметичного, можна розглядати і зважені ковзні середні, тобто коли значення вихідного ряду в інтервалі згладжування підсумовується з певними вагами. Подібні процедури доцільні, якщо зміна часового ряду в часі носить явно нелінійний характер.

Зараз існують різні таблиці для значень ваг зваженого ковзного середнього. Алгоритми звичайного і зваженого ковзного середнього з врахуванням втрат рівнів на кінцях ряду наведені нижче. Хоча, готових процедур для реалізації цих алгоритмів в середовищі MS Excel немає, проте самі алгоритми досить просто можуть бути реалізовані математичними операціями **Майстра функцій.** Під час реалізації алгоритмів необхідно звернути увагу на індексацію варіант. Приклади застосування простих алгоритмів ковзних середніх наведено нижче.

Лінійне згладжування, коли $\mathbf{w = 3}$.

Наведений алгоритм реалізує ковзне середнє, використовуючи мінімальний інтервал згладжування, причому перша і третя формули обчислюють найбільш відповідні значення втрачені на початку і в кінці ряду.

$$\bar{y}_1 = \frac{5y_1 + 2y_2 - y_3}{6}; \qquad \bar{y}_i = \frac{y_{i-1} + y_i + y_{i+1}}{3}, \qquad \bar{y}_n = \frac{-y_{n-2} + 2y_{n-1} + 5y_n}{6}, \quad (8)$$

$$i = 2,\ 3,\ \ldots,\ n-1.$$

Результат роботи цього алгоритму відображено графіком на рис. 5. для підвищення ефективності згладжування процедуру можна застосувати до вже згладжених рівнів.

Рис. 5. Згладжування простим ковзним середнім для $w = 3$.

Лінійне згладжування для $\mathbf{w = 5}$.

В цьому випадку, в результаті згладжування втрачається по два рівні на початку і в кінці часового ряду. Для усунення цього недоліку в алгоритмі використовуються спеціальні формули перерахунку втрачених значень на підставі реальних, тобто включаючи і самі втрачені значення. Для обчислення втрачених

значень на початку часового ряду використовують першу і другу формулу, а для обчислення втрачених в кінці – четверту і п'яту формули.

$$\bar{y}_1 = \frac{3y_1 + 2y_2 + y_3 - y_5}{5}, \qquad \bar{y}_2 = \frac{4y_1 + 3y_2 + 2y_3 + y_4}{10}, \qquad (9)$$

$$\bar{y}_i = \frac{y_{i-2} + y_{i-1} + y_i + y_{i+1} + y_{i+2}}{5}, \qquad \bar{y}_{n-1} = \frac{y_{n-3} + 2y_{n-2} + 3y_{n-1} + 4y_n}{10},$$

$$\bar{y}_n = \frac{-y_{n-4} + y_{n-2} + 2y_{n-1} + 3y_n}{5}, \qquad \textbf{де} \quad i = 3, 4, \ldots, n-2$$

Результат роботи цього алгоритму зображено наступним графіком на рис. 6.

Ковзне середнє дозволяє проводити операцію згладжування, як лінійними, так і нелінійними алгоритмами різних модифікацій. В літературі з опрацювання результатів експериментальних досліджень саме ці формули широко використовуються.

Ковзне середнє має найбільш просту інтерпретацію і алгоритм, проте його ефективність з точки зору згладжування не є вельми високою, особливо за малих значень розміру інтервалу w. Тому в багатьох випадках застосовують нелінійне згладжування.

Рис. 6. Згладжування ковзним середнім для $w = 5, \quad k = 2$.

Нелінійне згладжування $\mathbf{w = 7}$**.**

У випадку нелінійних змін тренду загальної тенденції використовують нелінійні формули. Наприклад, алгоритм нелінійного згладжування з врахуванням втрачених рівнів на початку і в кінці ряду з величиною інтервалу згладжування має такий вигляд.

$$\bar{y}_1 = \frac{39y_1 + 8y_2 - 4y_3 - 4y_4 + y_5 + 4y_6 - 2y_7}{42},$$

$$\bar{y}_2 = \frac{8y_1 + 19y_2 + 16y_3 + 6y_4 - 4y_5 - 7y_6 + 4y_7}{42},$$

$$\bar{y}_1 = \frac{39y_1 + 8y_2 - 4y_3 - 4y_4 + y_5 + 4y_6 - 2y_7}{42},$$

$$\bar{y}_2 = \frac{8y_1 + 19y_2 + 16y_3 + 6y_4 - 4y_5 - 7y_6 + 4y_7}{42},$$

$$\bar{y}_{n-2} = \frac{y_{n-6} - 4y_{n-5} + 2y_{n-4} + 12y_{n-3} + 19y_{n-2} + 16y_{n-1} - 4y_n}{42},$$

$$\bar{y}_{n-1} = \frac{4y_{n-6} - 7y_{n-5} - 4y_{n-4} + 6y_{n-3} + 16y_{n-2} + 19y_{n-1} + 8y_n}{42},$$

$$\bar{y}_n = \frac{-2y_{n-6} + 4y_{n-5} + y_{n-4} - 4y_{n-3} - 4y_{n-2} + 8y_{n-1} + 39y_n}{42},$$

$$i = 4, 5, \ldots, n-3.$$

Експоненціальне згладжування.

Поряд з простими і зваженими (поліноміальними) середніми для виявлення тенденцій у динаміці часових рядів використовуються експоненціальні середні. Вони формуються під впливом усіх попередніх рівнів ряду таким чином, що більш пізня інформація набуває більшого значення в силу застосування спеціальної системи ваг. В цьому методі згладжене значення визначається лише двома значеннями – біжучим і останнім згладженим рівнем та співвідношенням їхніх ваг, а саме α і $(1-\alpha)$. Важливим параметром і значним недоліком цього методу є оптимальний вибір величини α цього співвідношення.

Рис. 7. Результат нелінійного згладжування для $w = 7$.

Суттєвою ознакою застосування експоненціальних середніх є обґрунтування величини параметра згладжування α. Чим він менший, тим більше згладжуються рівні в ряді, який аналізується. Це означає зростання питомої α. Цей метод згладжування має досить широке застосування в прогнозуванні економічних часових рядів. Метод є однопараметричним, а його єдиний параметр α вибирається за наступної умови:

$$
\begin{aligned}
&0.1 \le \alpha \le 0.3,\\
&\tilde{y}_0 = y_1,\\
&\tilde{y}_1 = \alpha y_1 + (1-\alpha)\tilde{y}_0,\\
&\tilde{y}_2 = \alpha y_2 + (1-\alpha)\tilde{y}_1,\\
&\ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots\\
&y_n = \alpha y_n + (1-\alpha)\tilde{y}_{n-1}.
\end{aligned}
\tag{10}
$$

Початкове $\tilde{y}_0$ значення можна також розрахувати за формулами для ковзного середнього. Загалом алгоритм медіанного згладжування є таким:

$$\tilde{y}_i = \tilde{y}_0^* + \sum_{i=1}^{N}[\alpha\, y_i + (1-\alpha)\tilde{y}_{i-1}], \tag{11}$$

де $\tilde{y}_0^*$ – екстрапольоване значення, α – параметр згладжування;

Результати цього типу згладжування відображені нижче графічно на рис.8. Аналіз графіка експоненціального згладжування показує, що експоненціальне згладжування дає сильне «запізнення», а тому для часових рядів зі значною дисперсією рівнів воно є мало ефективним. Крім того, результат згладжування сильно залежить від параметра згладжування α, для якого на цей час відсутня методика визначення його величини, і хоча існують деякі підходи, проте усі вони є емпіричними.

Рис. 8. Результат експоненціального згладжування.

2.5. Медіанна фільтрація

Медіана відноситься до розподіленого середнього, тобто є значенням ознаки, яка займає місце всередині варіаційного ряду і на відміну від середньої арифметичної, що узагальнює величину показника, залишає значення того показника, який відповідає медіані. Вона є найбільш стійкою характеристикою варіаційного ряду. Як параметр вибірки, вона є однією з числових характеристик її структури та іноді використовується як альтернатива середньому арифметичному значенню елементів вибірки.

Медіанне згладжування є не обчислювальною нелінійною процедурою, оскільки для нього не виконується одна з аксіом лінійності, а саме: медіана суми двох довільних послідовностей не рівна сумі їхніх медіан. Характерною особливістю медіанного згладжування є те, що монотонні ділянки послідовності даних та різкі перепади вона залишає без змін, а для немонотонних ділянок в межах розміру ковзного інтервалу згладжування залишає лише центроване значення рівне їхній медіані, тобто ефективно усуває ті рівні, які порушують монотонність.

Зміст алгоритму медіанного згладжування часових рядів полягає у визначені значення медіани для рівнів інтервалу згладжування. Далі значенням медіани заміняють значення того рівня часового ряду, якому відповідає середина інтервалу згладжування. Іншими словами, медіана відповідає моменту часу t за умови, що межі інтервалу згладжування відповідають моментам часу $[t-k,\ t+k]$, тобто, коли інтервал згладжування є рівним $w=2k+1$. Медіанне згладжування повністю усуває поодинокі екстремальні або аномальні значення рівнів, які віддалені один від одного як мінімум на половину величини інтервалу згладжування; зберігає різкі перепади в тенденції (ковзне середнє та експоненціальне згладжування їх змазує); ефективно усуває поодинокі рівні з дуже великими або дуже малими значеннями, які мають випадковий характер і різко виділяються серед інших рівнів.

Медіанна фільтрація використовується досить рідко, хоча вона є вельми ефективною для часових рядів, рівні яких мають суттєво асиметричний розподіл. Загальним алгоритмом медіанної фільтрації є такий

$$\bar{y}_i = \max\{\min(y_{i-1},\ y_i),\ \min(y_i,\ y_{i+1}),\ \min(y_{i-1},\ y_{i+1})\}, \qquad (12)$$

де індекс i приймає значення $i=2,\ 3,\ 4,\ \ldots,\ n-1$.

Результати застосування медіанної фільтрації відображені на наступному графіку (рис.9).

Рис. 9. Ефективність застосування медіанної фільтрації.

При повторному застосуванні медіанного згладжування (за постійного розміру інтервалу вікна) перепади в поведінці рівнів зберігаються до тих пір, поки є зміни в згладжуваному ряді. Основними властивостями медіани є такі: кількість додатних відхилень від медіани є рівна кількості від'ємних та сума абсолютних відхилень варіант вибірки відносно медіани є мінімальною. Вперше ковзна оцінка

медіани, тобто медіанне згладжування, для аналізу часових рядів була запропонована і застосована Тьюки в 1971 р. Він також вказував, що медіанне згладжування зберігає в часових рядах більш різкі зміни їх рівня (тобто перепади). Початкове $\bar{y}_1$ і кінцеве $\bar{y}_n$ значення обчислюють за формулами для ковзного середнього. Такий розрахунок можна провести як до виконання операції медіанної фільтрації, так і після неї. Крім того, як початковий і кінцевий рівні можна використати початкове y_1 і кінцеве y_n оригінальні значення. В середовищі пакету MS Excel значення для медіанної фільтрації, тобто реалізацію цього алгоритму можна здійснити, використовуючи процедуру знаходження медіани, тобто

$$\tilde{y}_i = \textbf{МЕДІАНА}\left(y_{i-1}, y_i, y_{i+1}\right). \qquad (13)$$

Медіанну фільтрацію переважно здійснюють за трьома рівнями, тобто інтервал згладжування має $w = 3$, проте можна використовувати будь-які непарні додатні цілі числа.

2.6. Нормування часових послідовностей

Якщо, деяку якість досліджуваного об’єкта необхідно описати числом, то таке число x формується як сума балів. Нехай, це число змінюється від деякого мінімального значення $x_{\min}$ (відображення відсутності цієї якості) до деякого максимального значення $x_{\max}$ (найвищий ступінь прояву її наявності, прояву тощо). Отримання такого числа дає підстави для вирішення проблеми порівняння двох об'єктів, але тільки за показником, якому відповідає це число. Проте, треба завжди пам'ятати, в яких межах змінюється показник. Найбільш часто застосовується лінійне перетворення, яке полягає в тому, що значення рівнів часового ряду приводять до інтервалу значень $[0, 1]$, використовуючи такий алгоритм

$$y_i^н = \frac{y_i - y_{\min}}{y_{\max} - y_{\min}}, \qquad (14)$$

де $y_i^н$ – нормоване значення, y_i – біжуче значення рівня, $y_{\min}$ та $y_{\max}$ – найменше та найбільше значення рівнів цього часового ряду.

Нормування часових послідовностей в першу чергу дає можливість порівняти показники, що отримані для різних об’єктів, оскільки при такому нормуванні структура рядів (пропорційність між рівнями в рядах) залишається незмінною. Це дає можливість порівнювати обчислені показники та побудовані на таких даних моделі.

2.7. Критерії ефективності згладжування часових рядів

Використання того чи іншого методу згладжування залежить, з одного боку від забезпечення вимог самої задачі, а з іншого – від структури самого ряду. Тому, важливе значення має визначення ефективності існуючих методів. Для цього використовують різні критерії, серед яких найбільш адекватними цій задачі є: критерій поворотних точок та коефіцієнт кореляції між вихідним та згладженим рядами.

Критерії поворотних точок. Для оцінювання ефекту згладжування пропонується використати критерій поворотних точок, зміст якого полягає в звичайному підрахунку рівнів, значення яких є більшими або меншими ніж два сусідні. Цей критерій легко реалізується в Excel.

Коефіцієнт кореляції. Для оцінювання тісноти зв'язку між вихідним (оригінальним рядом) та згладженим використовують коефіцієнт кореляції. Для його знаходження в Excel є спеціальна опція в надбудові «Аналіз даних».

Формула для обчислення поворотних точок

=ЕСЛИ((I3>I2);(I3>I4);ИЛИ(ЕСЛИ((I3<I2);(I3<I4))))

2.8. Формули для зваженого ковзного середнього

В задачах з використанням простого ковзного середнього обчислюється середнє значення рівнів, що входять в інтервал згладжування. Ефект в цьому випадку, навіть при використанні великих інтервалів згладжування, є не дуже значним. Тому застосовують повторне згладжування, збільшуючи для кожного повторного підходу розмір вікна. В цьому випадку ефект є суттєвим.

В таблицях наведені значення ваг для ковзних середніх. Таблиці розроблені різними авторами і дають різний ефект згладжування. Наприклад, для простої ковзної середньої з розміром інтервалу згладжування $W=5$ формули для розрахунку втрачених рівнів мають вигляд, поданий наступними виразами (9). Іншими словами, приведені таблиці з формулами з Кендела реалізують прості ковзні середні, оскільки для інтервалу згладжування всередині часового ряду, що відповідає середньому стовпчику таблиці, всі ваги рівні 1.

Згладжування формулами з Кендела

N=3

1	2	n
5	1	-1
2	1	2
-1	1	5
6	3	6
0,913	0,577	0,913

N=5

1	2	3	n-1	n
3	4	1	0	-1
2	3	1	1	0
1	2	1	2	1
0	1	1	3	2
-1	0	1	4	3
5	10	5	10	5
0,775	0,548	0,447	0,548	0,775

N=7

1	2	3	4	n-2	n-1	n
13	5	7	1	1	-1	-5
10	4	6	1	2	0	-2
7	3	5	1	3	1	1
4	2	4	1	4	2	4
1	1	3	1	5	3	7
-2	0	2	1	6	4	10
-5	-1	1	1	7	5	13
28	14	28	7	28	14	28
0,681	0,535	0,423	0,378	0,423	0,535	0,681

N=9

1	2	3	4	5	n-3	n-2	n-1	n
17	56	22	32	1	8	-2	-16	-7
14	47	19	29	1	11	1	-7	-4
11	38	16	26	1	14	4	2	-1
8	29	13	23	1	17	7	11	2
5	20	10	20	1	20	10	20	5
2	11	7	17	1	23	13	29	8
-1	2	4	14	1	26	16	38	11

-4	-7	1	11	1	29	19	47	14
-7	-16	-2	8	1	32	22	56	17
45	180	90	180	9	180	90	180	45
0,615	0,511	0,422	0,357	0,333	0,357	0,422	0,511	0,615

N=11

1	2	3	4	5	6	5	4	3	2	1
7	15	25	10	15	1	5	0	-5	-5	-3
6	13	22	9	14	1	6	1	-2	-3	-2
5	11	19	8	13	1	7	2	1	-1	-1
4	9	16	7	12	1	8	3	4	1	0
3	7	13	6	11	1	9	4	7	3	1
2	5	10	5	10	1	10	5	10	5	2
1	3	7	4	9	1	11	6	13	7	3
0	1	4	3	8	1	12	7	16	9	4
-1	-1	1	2	7	1	13	8	19	11	5
-2	-3	-2	1	6	1	14	9	22	13	6
-3	-5	-5	0	5	1	15	10	25	15	7
22	55	110	55	110	11	110	55	110	55	22
0,564	0,486	0,416	0,357	0,316	0,302	0,316	0,357	0,416	0,486	0,564

N=13

1	2	3	4	5	6	7	6	5	4	3	2	1
25	44	19	32	13	20	1	8	1	-4	-5	-16	-11
22	39	17	29	12	19	1	9	2	-1	-3	-11	-8
19	34	15	26	11	18	1	10	3	2	-1	-6	-5
16	29	13	23	10	17	1	11	4	5	1	-1	-2
13	24	11	20	9	16	1	12	5	8	3	4	1
10	19	9	17	8	15	1	13	6	11	5	9	4
7	14	7	14	7	14	1	14	7	14	7	14	7
4	9	5	11	6	13	1	15	8	17	9	19	10
1	4	3	8	5	12	1	16	9	20	11	24	13
-2	-1	1	5	4	11	1	17	10	23	13	29	16
-5	-6	-1	2	3	10	1	18	11	26	15	34	19
-8	-11	-3	-1	2	9	1	19	12	29	17	39	22
-11	-16	-5	-4	1	8	1	20	13	32	19	44	25
91	182	91	182	91	182	13	182	91	182	91	182	91
0,524	0,406	0,406	0,355	0,314	0,287	0,277	0,287	0,314	0,355	0,406	0,463	0,524

N=15

1	2	3	4	5	6	7	8	7	6	5	4	3	2	1
29	91	161	35	119	49	77	1	35	7	-7	-7	-49	-35	-13
26	82	146	32	110	46	74	1	38	10	2	-4	-34	-26	-10
23	73	131	29	101	43	71	1	41	13	11	-1	-19	-17	-7
20	64	116	26	92	40	68	1	44	16	20	2	-4	-8	-4
17	55	101	23	83	37	65	1	47	19	29	5	11	1	-1
14	46	86	20	74	34	62	1	50	22	38	8	26	10	2
11	37	71	17	65	31	59	1	53	25	47	11	41	19	5
8	28	56	14	56	28	56	1	56	28	56	14	56	28	8
5	19	41	11	47	25	53	1	59	31	65	17	71	37	11
2	10	26	8	38	22	50	1	62	34	74	20	86	46	14
-1	1	11	5	29	19	47	1	65	37	83	23	101	55	17
-4	-8	-4	2	20	16	44	1	68	40	92	26	116	64	20
-7	-17	-19	-1	11	13	41	1	71	43	101	29	131	73	23
-10	-26	-34	-4	2	10	38	1	74	46	110	32	146	82	26
-13	-35	-49	-7	-7	7	35	1	77	49	119	35	161	91	29
120	420	840	210	840	420	840	15	840	420	840	210	840	420	120
0,492	0,442	0,395	0,352	0,314	0,285	0,265	0,258	0,265	0,285	0,314	0,352	0,395	0,442	0,492

Згладжування формулами з Полларда. Застосування цих формул продемонструємо таким прикладом. Нехай маємо фрагмент таблиці з рівнями деякого часового ряду. Обчислимо для рівня №20 його згладжене значення за

формулою, ваги якої приведені в стовпчику 9, тобто для величини інтервалу згладжування $w = 9$.

Номер рівня	15	16	17	18	19	**20**	21	22	23	24	25
Значення рівня	859	590	538	823	619	**564**	708	569	504	976	567

В межі такого інтервалу включені рівні з такими номерами: 16, 17, 18, 19, 20, 21, 22, 23, 24. Тоді ваги для стовпчика 9 розподіляться в такий спосіб:

w_0^9 = 0,33114 для рівня № 20,

w_1^9 = 0,266557 для рівнів № 19 і № 21,

w_2^9 = 0,11847 для рівнів № 18 і № 22,

w_3^9 = -0,00987 для рівнів № 17 і № 23,

w_4^9 = -0,04072 для рівнів № 16 і № 24,

а сама формула матиме такий вигляд

y_{20} = (-0,04072) × 590 + (-0,00987) × 538 + 0,11847 × 823 + 0,266557 × 619 + + 0,33114 × **564** + 0,266557 × 708 + 0,11847 × 569 + (-0,00987) × 504 + + (-0,04072) × 976 = $\tilde{y}_{20}$.

	3	**5**	**7**	**9**	**11**	**13**	**15**
w_0	0,78082	0,559441	0,412587	0,33114	0,277945	0,240057	0,211541
w_1	0,10959	0,293706	0,293706	0,266557	0,238693	0,214337	0,193742
w_2		-0,07343	0,058741	0,11847	0,141267	0,147356	0,145904
w_3			-0,05874	-0,00987	0,035723	0,065492	0,082918
w_4				-0,04072	-0,02679	0	0,024027
w_5					-0,02786	-0,02786	-0,01413
w_6						-0,01935	-0,0245
w_7							-0,01373

Як видно цієї формули, ваги розташовані симетрично відносно середини інтервалу згладжування.

3. Кореляційний аналіз часових послідовностей

У практичній діяльності системного аналітика часто зустрічаються ситуації, в яких оцінку однієї з властивостей об'єкта необхідно здійснювати з врахуванням оцінки іншої властивості. У цьому випадку виникає необхідність визначення взаємного впливу властивостей. Закономірності такого впливу досить складно описати математичними моделями. В подібних ситуаціях використовують кореляційну оцінку показників, встановлюючи елемент якісної, експертної оцінки впливу одного показника на інший. Метою дослідника при вирішенні зазначеної задачі є не тільки знаходження кореляційної залежності між двома властивостями об'єкта, а й отримання якісної (експертної) оцінки впливу однієї властивості на іншу.

Під кореляційним аналізом розуміють групу методів, що дозволяють виявляти наявність і ступінь взаємозв'язку між кількома параметрами, що змінюються випадковим чином. Міра такого взаємозв'язку оцінюється спеціальними числовими характеристиками, а також їх статистиками, що визначають ступінь

близькості цього взаємозв'язку до функціонального, який може існувати між параметрами, що володіють детермінованим характером зміни.

Кореляційний зв'язок з'являється, коли одному і тому ж значенню аргументу (незалежної змінної) відповідає низка значень функції (залежної змінної). Тоді зв'язок виявляється у вигляді тенденції зміни середніх значень функції залежно від змін аргументу. Цим кореляційний зв'язок відрізняється від функціонального, який виникає у разі, коли заданому значенню аргументу відповідає цілком певне значення функції. Кореляційний зв'язок є неповним, оскільки залежність між функцією і аргументом в кожному конкретному випадку схильна до впливу з боку інших чинників (які найчастіше мають мінливий характер).

Найбільш повно в статистиці розроблена методологія парної кореляції, що розглядає вплив варіації однієї факторної ознаки на результатну.

Дослідження парної кореляції здійснюється на основі кореляційного аналізу, передбачає послідовне вирішення низки завдань:

- виявлення зв'язку;
- опис зв'язку в табличній і графічній формах;
- вимірювання тісноти зв'язку;
- формулювання висновків про характер існуючого зв'язку.

Основні завдання кореляційного аналізу – це визначення і вираження форми аналітичної залежності результативної ознаки у від факторних ознак x_i.

Відмінною рисою кореляційного аналізу є вимірювання тісноти зв'язку між y та x. Його основними числовими характеристиками є коефіцієнт кореляції і кореляційне відношення.

Виділяють такі етапи кореляційного аналізу:

– виявлення взаємозв'язку між ознаками;

– визначення форми зв'язку;

– визначення сили (тісноти) і напрямку зв'язку .

Інтерпретуючи результати кореляційного аналізу потрібно врахувати, що коефіцієнт кореляції є статистичним показником, який не вказує на те, що досліджувані величини знаходяться в причинно-наслідковому зв'язку. Тому, будь-яке трактування кореляційної залежності повинно ґрунтуватися на інформації про суть і характер досліджуваних експериментальних даних та процесів, яким вони відповідають. До переваг кореляційного аналізу можна віднести можливість створення нового правила взаємодії функцій одна з одною, а також оцінку взаємодії функцій, які отримані невідомим шляхом. Недоліками є те, що всі результати, отримані за допомогою цієї методики можна використовувати лише в області цього дослідження або близької до нього. Після виявлення стохастичних зв'язків між досліджуваними змінними величинами дослідник приступає до математичного опису виявленої ним і цікавої для нього залежності.

3.1. Кореляційне поле

У своїй практиці дослідник часто стикається з необхідністю встановлення факту існування функціональних чи інших залежностей між отриманими експериментальними даними. Проте, вже на підставі візуального аналізу поля кореляції можна висунути гіпотезу щодо існуючого зв'язку між усіма можливими значеннями X та Y: лінійний або нелінійний, сильний, слабкий або відсутній. Переважно такий зв'язок є випадковим. Графічне подання взаємозв’язку між двома досліджуваними послідовностями називається *кореляційним полем* або *полем кореляції* або *діаграмою розсіювання.* Графічний метод забезпечує наочне

зображення форми зв'язку між цими послідовностями. Для цього, в прямокутній системі координат будують графік – по осі ординат відкладають індивідуальні значення однієї послідовності, вибраної в якості результативної ознаки Y, а по осі абсцис – індивідуальні значення іншої – факторної ознаки X. Саме сукупність точок результативної і факторної ознак називається *полем кореляції*.

Отже, кореляційне поле – це сукупність точок у прямокутній системі координат, абсциса кожної з яких відповідає значенню факторної ознаки (x), а ордината – значенню результативної ознаки (y) певної одиниці спостереження. Кількість точок на графіку відповідає кількості одиниць спостереження. Використовується для аналізу наявності та характеру (напряму) зв'язку між результатами двох вибірок спостережень. Розміщення точок на графіку свідчить про наявність і напрям зв'язку. Загалом, локалізація точок кореляційного поля вказує на наявність прямого, оберненого зв'язку між ознаками або його відсутність, а також на форму лінії регресії (рис. 10). Розміщення точок на кореляційному полі дозволяє судити про характер залежності, наприклад: лінійна, параболічна, гіперболічна, логічна, логарифмічна, експонентна, показникова або відсутність залежності.

Кореляційні зв'язки можна вивчати на якісному рівні з діаграм розсіяння емпіричних значень змінних X та Y і відповідним чином їх інтерпретувати. Так, наприклад, якщо підвищення рівня однієї змінною супроводжується підвищенням рівня іншої, то йдеться про *додатну* кореляцію або прямий зв'язок (рис. 11,). Якщо ж зростання однієї змінної супроводжується зниженням значень іншої, то маємо справу з *від'ємною* кореляцією або оберненим зв'язком (рис. 12). Нульовою або відсутньою називається кореляція за відсутності зв'язку змінних (рис. 10). Проте нульова загальна кореляція може свідчити лише про відсутність *лінійної* залежності, а не взагалі про відсутність будь якого *статистичного* зв'язку. При цьому функцію, графік якої відповідає розміщенню точок називають теоретичною лінією регресії. Для вибору тієї чи іншої форми кореляційної залежності, треба порівняти уявну емпіричну лінію регресії з графіками відомих функцій. Методи кореляційного аналізу широко застосовуються для виявлення та опису стохастичних залежностей між випадковими величинами, якими переважно є зібрані або експериментальні дані.

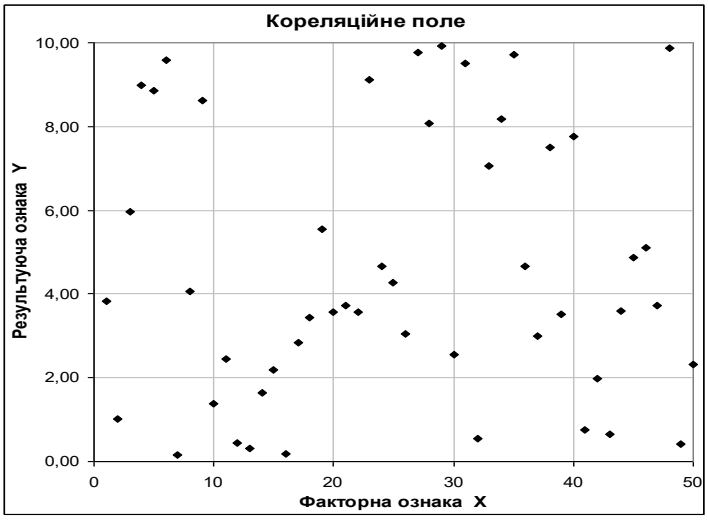


Рис. 10. Візуальна оцінка характеру зв'язку вказує на його відсутність.

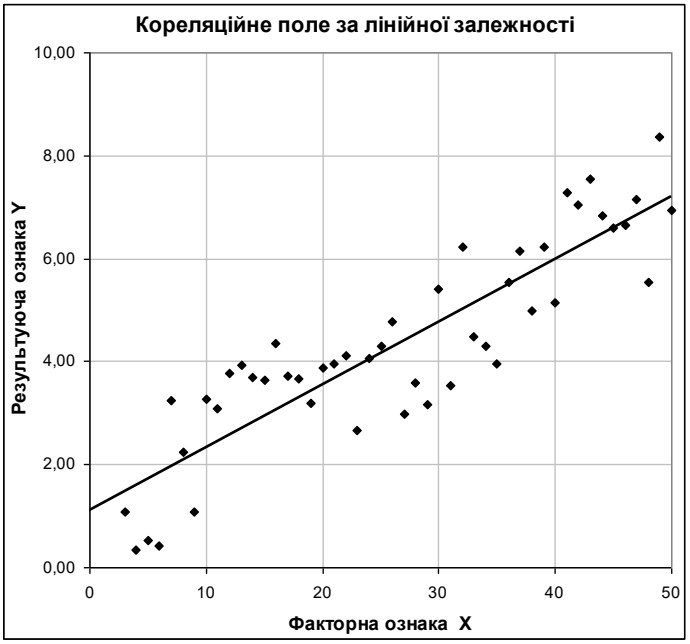


Рис. 11. Прямий лінійний зв'язок.

Для експериментального вивчення залежності між випадковими величинами Y і X проводять деяку кількість незалежних дослідів. Результат i-го досліду дає пару значень (x_i, y_i), де $i = 1, 2, \ldots, n$. Отже, досліджувані послідовності можна подати так: $X = \langle x_1, x_2, \ldots, x_n \rangle$; $Y = \langle y_1, y_2, \ldots, y_n \rangle$. Якщо послідовності подати у вигляді функцій, що залежать від одного аргументу, то, провівши кореляційний аналіз, можна встановити взаємний вигляд зв'язку між ними та його величину, при цьому обсяг даних має бути однаковий. Про наявність чи відсутність кореляції між двома випадковими величинами якісно можна судити з вигляду поля кореляції, відобразивши експериментальні пари точок на координатну площину.

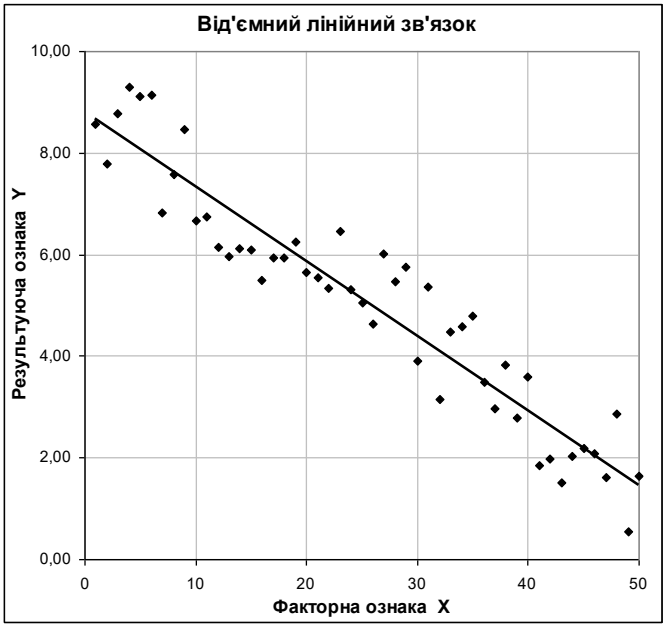


Рис. 12. Обернений лінійний зв'язок.

3.2. Коефіцієнт кореляції

Для кількісної оцінки тісноти зв'язку служить вибірковий *коефіцієнт кореляції*. Вибірковий коефіцієнт кореляції r за абсолютною величиною не перевищує одиниці. Для незалежних випадкових величин коефіцієнт кореляції дорівнює нулю, але він може бути рівний нулю для деяких залежних величин, які при цьому називаються некорельованими.

Для випадкових величин, що мають нормальний розподіл, відсутність кореляції означає і відсутність будь-якої залежності.

Вибірковий коефіцієнт кореляції не змінюється при зміні початку відліку і масштабу величин . Коефіцієнт кореляції характеризує не довільну залежність, а тільки лінійну. Лінійна ймовірнісна залежність випадкових величин полягає в тому , що при зростанні однієї випадкової величини інша має тенденцію зростати (або спадати) за лінійним законом.

Коефіцієнт кореляції характеризує ступінь тісноти лінійної залежності. У загальному випадку, коли величини X та Y пов'язані деякою стохастичною залежністю, коефіцієнт кореляції може мати значення в межах $-1 \le r \le +1$.

Відзначимо властивості коефіцієнта кореляції:

- коефіцієнт парної кореляції обчислюється для кількісних ознак;
- коефіцієнт кореляції симетричний, тобто не змінюється, якщо X та Y поміняти місцями;
- коефіцієнт кореляції є величиною безрозмірною;
- коефіцієнт кореляції не змінюється при зміні одиниць виміру ознак X та Y;
- величина коефіцієнта кореляції не змінюється від додавання до X та Y невипадкових доданків;
- величина коефіцієнта кореляції не змінюється від множення X та Y на додатні числа;
- якщо одну з величин, не змінюючи іншу, помножити на –1, то на –1 треба помножити і коефіцієнт кореляції.

Схема застосування кореляційного аналізу з практичною метою приблизно така: є кілька параметрів, які спостерігаються протягом деякого проміжку часу, про які, за результатами спостережень (або з будь-яких апріорних міркувань), можна припустити, що вони можуть бути взаємопов'язані будь-яким чином.

Обчислення коефіцієнта кореляції здійснюють за такою формулою

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i-1}^{n} (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i-1}^{n} (y_i - \bar{y})^2}}. \qquad (15)$$

В статистичній літературі рекомендують використовувати для обчислення коефіцієнта кореляції наступний вираз

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] \left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right]}}. \qquad (16)$$

В цьому випадку немає потреби обчислювати відхилення біжучих значень від середньої величини, а це виключає помилки в розрахунках при заокругленні середніх величин.

Коефіцієнт кореляції r_{xy} є випадковою величиною, оскільки обчислюється для випадкових величин. Стосовно нього можна висувати і перевіряти такі гіпотези:

1. Коефіцієнт кореляції значимо відрізняється від нуля, тобто між величинами є взаємний зв'язок. Для перевірки цієї гіпотези обчислюють тестову статистику за такою формулою

$$\xi = \left(0.5 \cdot \ln\left(\frac{1 + r_{xy}}{1 - r_{xy}} \right) - \frac{|r_{xy}|}{2(n-1)} \right) \sqrt{n-3}. \qquad (17)$$

Обчислене значення ξ порівнюється з табличним значенням коефіцієнта Стьюдента $t(p = 0.95,\ f = \infty) = 1.96$. Якщо тестова статистика є більшою за табличне значення, то коефіцієнт кореляції значимо відрізняється від нуля. З формули випливає, що, чим більше вимірів n, тим більшою є тестова статистика, а, отже, коефіцієнт кореляції значимо відрізняється від нуля.

2. Значення коефіцієнта кореляції є значимим, якщо обчислений коефіцієнт детермінації перевірити за допомогою критерію Стьюдента

$$t_{розр} = r_{xy} \cdot \sqrt{\frac{n-k-1}{1-r_{xy}^2}}. \qquad (18)$$

Табличне значення *t*-критерію Стьюдента за довірчої ймовірності 0,95 і для числа ступенів свободи $\gamma = (n-k-1)$ порівнюється з обчисленим значенням. Якщо обчислене значення перевищує табличне, то коефіцієнт кореляції визнається значимим. Оцінка значущості коефіцієнта парної кореляції з використанням *t*-критерію Стьюдента може бути обчислена за такою формулою

$$t_{розр} = \sqrt{\frac{r_{xy}^2}{1-r_{xy}^2} \cdot (n-2)}. \qquad (19)$$

Обчислене за цією формулою значення t порівнюється з критичним значенням *t*-критерію, яке береться з таблиці значень *t* Стьюдента з врахуванням заданого рівня значущості і числа ступенів свободи (*n*-2).

3. Відмінність між двома коефіцієнтами кореляції є значимою, якщо тестова статистика

$$\zeta = 0.5 \cdot \ln\left(\frac{(1+r_1)(1-r_2)}{(1-r_1)(1+r_2)} \right) \cdot \frac{1}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}, \qquad (20)$$

яку також порівнюють з табличним значенням $t(p,\infty)=1.96$

Сам по собі коефіцієнт кореляції не має змістовної інтерпретації. Проте його квадрат $R=r^2$, який називають *коефіцієнтом детермінації* (позначається R і зазвичай виражається у %), має простий зміст – це показник того, наскільки зміни залежної ознаки пояснюються змінами незалежної.

З визначення коефіцієнта детермінації випливає, що він приймає значення в діапазоні від 0 % до 100 %.

Якщо дві змінні функціонально лінійно залежні (точки на кореляційному полі лежать на одній прямій), то можна сказати, що зміна однієї з них повністю пояснюється зміною іншої, а це якраз той випадок, коли коефіцієнт детермінації дорівнює 100 % (при цьому коефіцієнт кореляції може дорівнювати як +1, так і – 1) .

Чим вище за модулем (за абсолютною величиною) значення коефіцієнта кореляції, тим сильніший зв'язок між ознаками.

Прийнято вважати, що коефіцієнти кореляції, які за модулем більше 0.7, вказують на сильний зв'язок (при цьому коефіцієнти детермінації > 50%, тобто одна ознака визначає іншу більш, ніж наполовину).

Коефіцієнти кореляції, які по модулю менше 0,7 , але більше 0,5 , говорять про зв'язок середньої сили (при цьому коефіцієнти детермінації менше 50 % , але більше 25%). Нарешті, коефіцієнти кореляції, які по модулю менше 0,5 , говорять про слабкий зв'язок (при цьому коефіцієнти детермінації менше 25%).

3.3. Кореляційне відношення

При відхиленні парної статистичної залежності від лінійної коефіцієнт кореляції втрачає свій сенс як характеристика ступеня тісноти зв'язку. У цьому випадку користуються таким виміром зв'язку як ***кореляційне відношення***. Кореляційне відношення застосовують в тих випадках, коли:

- між парою досліджуваних ознак відзначається нелінійна залежність;
- характер вибіркових даних (кількість, щільність розташування на кореляційному полі) допускає, по-перше, їх групування по осі ординат, по-друге, можливість підрахунку "окремих" математичних очікувань всередині кожного інтервалу групування.

Послідовність методики обчислення кореляційного відношення. Нехай між X та Y існує нелінійна залежність, зображена на рис. 13.

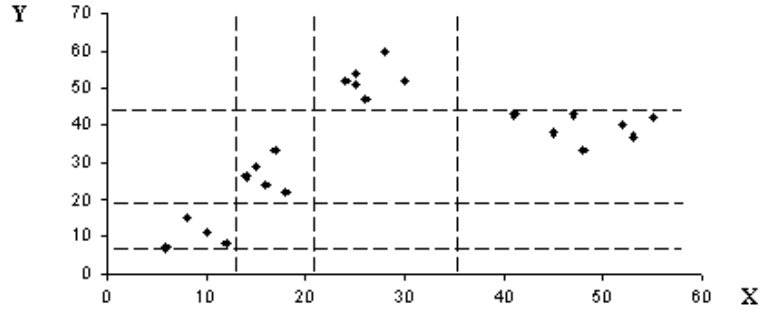


Рис. 13. Нелінійна залежність між компонентами X і Y.

1. Розіб'ємо кореляційне поле за змінною X на L інтервалів групування, що не перетинаються, які можуть мати різну довжину.

2. Знайдемо "частинні" математичні сподівання відгуку Y в кожній з L виділених груп

$$\overline{m}_{Y_j} = \frac{1}{n_j} \cdot \sum_{k=1}^{n_j} y_j \,, \tag{21}$$

де $j = \overline{1, L}$, $k = \overline{1, n_j}$, n_j – кількість елементів вибірки в j-му інтервалі групування.

3. Знайдемо математичне сподівання частинних групувань відгуків, використовуючи "частинні" $\overline{m}_{Y_j}$

$$\overline{m}_Y = \frac{1}{n} \cdot \sum_{j=1}^{L} n_j \cdot \overline{m}_{Y_j} \,. \tag{22}$$

4. Обчислимо групову дисперсію вихідної змінної Y

$$\sigma^2_{\overline{m}_Y} = \frac{1}{n} \cdot \sum_{j=1}^{L} n_j \cdot \left(\overline{m}_{Y_j} - \overline{m}_Y\right)^2 , \tag{23}$$

і дисперсію, яка отримана за не згрупованим відгуком

$$\overline{\sigma}^2_Y = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(Y_i - \overline{m}_Y\right)^2 . \tag{24}$$

5. Кореляційне відношення для залежної змінної Y і незалежної змінної X може бути отримане з відношення

$$\overline{\rho}_{Y \cdot X} = \frac{\overline{\sigma}_{\overline{m}_Y}}{\overline{\sigma}_Y} . \tag{25}$$

3.4. Властивості кореляційного відношення

Кореляційне відношення не має властивості симетрії, тобто $\overline{\rho}_{Y \cdot X} \neq \overline{\rho}_{X \cdot Y}$. Крім того, $\overline{\rho}_{Y \cdot X}$ не від'ємне, оскільки вважається, що воно є результатом добування кореня квадратного з $\rho^2_{Y \cdot X}$. Кореляційне відношення $0 \le \rho_{Y \cdot X} \le 1$.

З $\rho_{Y \cdot X} = 1$ випливає, що між Y та X існує однозначна функціональна залежність. Протилежне твердження в загальному випадку не вірне.

Відсутність кореляційного зв'язку між Y та X означає, що умовні середні $\overline{m}_{Y_j}$ зберігають від групи до групи постійне значення, що дорівнює загальному середньому $\overline{m}_Y$. Тому, необхідно також відзначити, що між $\rho_{Y \cdot X}$ і $\rho_{X \cdot Y}$ немає будь-якої певної залежності. Некорельованість Y від X не означає некорельованість X від Y. І, нарешті, дослідження показали, що $\rho_{YX} \ge \left| r_{YX} \right|$. Умова рівності виконується у випадку лінійної залежності X та Y. Всі зауваження щодо змістовної інтерпретації $\rho_{Y \cdot X}$ аналогічні інтерпретації значень отриманих даних.

3.5. Кореляційна матриця

У випадку великої кількості спостережень, коли коефіцієнти кореляції необхідно послідовно обчислювати для декількох вибірок, для зручності отримані коефіцієнти зводять в таблиці, які називають *кореляційними матрицями*.

Кореляційна матриця – це квадратна таблиця, в якій на перетині відповідних рядка і стовпця знаходиться коефіцієнт кореляції між відповідними параметрами.

У MS Excel для обчислення кореляційних матриць використовується процедура **КОРЕЛЯЦІЯ** з пакету **Аналіз даних**. Процедура дозволяє отримати кореляційну матрицю, що містить коефіцієнти кореляції між різними параметрами.

Для реалізації процедури необхідно:

– виконати команду **Сервіс** → **Аналіз даних** ;

– в списку **Інструменти аналізу** вибрати рядок **КОРЕЛЯЦІЯ** і натиснути кнопку ОК;

– в діалоговому вікні вказати **Вхідний інтервал**, тобто ввести посилання на клітинки, які містять аналізовані дані. Вхідний інтервал повинен містити не менше двох стовпців;

– в розділі **Групування** перемикач встановити відповідно до введених даних, тобто за стовпцями чи за рядками;

– вказати **Вихідний інтервал**, тобто ввести посилання на клітинку, для якої будуть виведені результати аналізу. Розмір вихідного діапазону буде визначений автоматично, і на екран буде виведене повідомлення у разі можливого накладення вихідного діапазону на вхідні дані. Натиснути кнопку ОК .

В результаті у вихідний діапазон буде виведена кореляційний матриця, в якій на перетині кожних рядка і стовпця знаходиться коефіцієнт кореляції між відповідними параметрами. Значення коефіцієнтів кореляції рівне $+1$, що розміщені вздовж діагоналі, вказує на те, що кожен стовпець у вхідному діапазоні повністю корелює сам з собою.

В процесі інтерпретації кожен коефіцієнт кореляції між відповідними параметрами розглядається окремо. Зазначимо, що хоча в результаті буде отримана трикутна матриця, кореляційна матриця є симетричною, оскільки в порожніх клітинках в правій верхній половині таблиці знаходяться ті ж самі коефіцієнти кореляції, що і в нижній лівій (симетрично розташовані відносно діагоналі.

3.6. Автокореляція

Для вивчення природи динаміки рівнів, які відповідають різним часовим інтервалам часто використовується *поняття автокореляції*, яка характеризує не тільки взаємозалежність рівнів одного й того ж ряду, що відносяться до різних моментів спостережень, але і ступінь стійкості розвитку процесу в часі. Корельованість рівнів часових послідовностей із застосуванням парного коефіцієнта кореляції правильно показує тісноту зв'язку лише в тому випадку, якщо в кожній з них відсутня автокореляція. Існування залежності між попередніми і наступними рівнями часової послідовності в статистичній літературі називають *автокореляцією*.

Застосування методів класичної теорії кореляції в часових послідовностях пов'язано з тим, що для більшості часових послідовностей має місце залежність наступних рівнів від попередніх. Так як методика кореляційного аналізу ґрунтується на принципі статистичної незалежності даних, наявність автокореляції може призвести до помилкового визначення суттєвості та довірчих меж коефіцієнтів регресії і до інших наслідків, що ставить під сумнів результати аналізу. Тому, якщо аналіз проводиться за даними за різні періоди, необхідно переконатися у відсутності автокореляції в досліджуваних рядах динаміки.

Тому, перш ніж корелювати такі послідовності за рівнями, необхідно перевірити кожну з них на наявність або відсутність в них автокореляції.

3.7. Автокореляція в часових рядах

Явище автокореляції має місце у тих випадках, коли кореляційний аналіз проводиться за даними за певні періоди, може виявитися явище автокореляції, тобто зв'язок між даними за попередні і подальші періоди. За наявності тенденції і циклічних коливань значення кожного наступного рівня ряду залежать від попередніх значень. Кореляційну залежність між послідовними рівнями часового ряду називають *автокореляцією* рівнів ряду. Кількісно її можна виміряти за допомогою лінійного коефіцієнта кореляції між рівнями вихідного часового ряду y_t і рівнями цього ряду, зсунутими на кілька кроків у часі $y_{t-\tau}$.

Ступінь тісноти статистичного зв'язку між рівнями часового ряду, зсунутими на τ одиниць часу визначається величиною коефіцієнта кореляції $r(\tau)$. Оскільки, $r(\tau)$ вимірює тісноту зв'язку між рівнями одного і того ж часового ряду, його прийнято називати *коефіцієнтом автокореляції*. При цьому довжину часового зсуву називають зазвичай лагом (τ). Коефіцієнт автокореляції обчислюється за безпосередніми даними рядів динаміки, коли фактичні рівні одного ряду розглядають як значення факторної ознаки, а рівні цього ж ряду зсунуті на один період, приймають за результативну ознаку (цей зсув називається лагом).

Число періодів, за якими розраховується коефіцієнт автокореляції, називають лагом. Із збільшенням лага число пар значень, за якими розраховується коефіцієнт автокореляції, зменшується. Максимальний лаг повинен бути не більше (n/4). Примітка: щоб уникнути плутанини, слід звернути увагу на порядок, за яким буде проводитися зміна рівнів, а саме, вниз або вгору. Відповідно і в формулах за різними джерелами, ряд із зсувом відображають y_{t-1} або y_{t+1}.

Коефіцієнт автокореляції характеризує тісноту тільки лінійного зв'язку поточного й аналізованого рівнів ряду. Тому за коефіцієнтом автокореляції можна судити про наявність лінійної (або близькою до лінійної) тенденції. Для деяких часових рядів, які мають сильну нелінійну тенденцію (наприклад, параболу або експоненту), коефіцієнт автокореляції рівнів вихідного ряду може наближатися до нуля. Послідовність коефіцієнтів автокореляції першого, другого тощо порядків називають автокореляційною функцією часової послідовності. Графік залежності значень коефіцієнтів автокореляції від величини лагу (порядку коефіцієнта автокореляції) називають корелограмою. За допомогою аналізу автокореляційної функції і корелограми можна виявити структуру послідовності, тобто визначити присутність у ній тієї чи іншої компоненти.

Аналіз структури ряду можна проводити таким чином:

- якщо найбільш високим виявився коефіцієнт автокореляції першого порядку, то досліджуваний ряд містить тільки тенденцію;
- якщо найвищим виявився коефіцієнт автокореляції порядку τ, то ряд містить циклічні коливання з періодичністю в τ моментів часу;
- якщо жоден з коефіцієнтів автокореляції не є значущим, можна зробити одне з двох припущень щодо структури ряду:
    - ряд не містить тенденції і циклічних коливань, а включає тільки випадкову компоненту;
    - ряд містить сильну нелінійну тенденцію, для виявлення якої потрібно провести додатковий аналіз.

Для судження про наявність або відсутність автокореляції в досліджуваній послідовності значень, фактичне значення коефіцієнта автокореляції зіставляють з табличним для 5% або 1% рівня значимості (тобто за величиною ймовірності

допустити помилку при прийнятті гіпотези про незалежність рівнів ряду). Якщо обчислене значення менше ніж табличне, то гіпотеза про відсутність автокореляції приймається і, навпаки, в протилежному випадку, відкидається. Необхідно підкреслити, що лінійні коефіцієнти автокореляції характеризують тісноту тільки лінійного зв'язку поточного і попередніх рівнів ряду. Тому, за коефіцієнтами автокореляції можна судити тільки про наявність чи відсутність лінійної залежності (або близькою до лінійної). Для деяких часових рядів, які мають сильну нелінійну тенденцію (наприклад, параболу другого порядку або експоненту), коефіцієнт автокореляції рівнів вихідного ряду може наближатися до нуля. За знаком коефіцієнта автокореляції не можна робити висновок про зростаючу або спадну тенденції в рівнях ряду.

3.8. Розрахунок автокореляції

Загальною формулою для обчислення коефіцієнта автокореляції є така

$$r(\tau)=\frac{(n-\tau)\sum_{t=1}^{n-\tau}y_t y_{t+\tau}-\sum_{t=1}^{n-\tau}y_t\sum_{t=1}^{n-\tau}y_{t+\tau}}{\sqrt{[(n-\tau)\sum_{t=1}^{n-\tau}y_t^2-(\sum_{t=1}^{n-\tau}y_t)^2][(n-\tau)\sum_{t=1}^{n-\tau}y_{t+\tau}^2-(\sum_{t=1}^{n-\tau}y_{t+\tau})^2]}}. \qquad (26)$$

Порядок коефіцієнтів автокореляції визначає часовий лаг: першого порядку (при $\tau=1$), другого порядку (при $\tau=2$) тощо.

Послідовність коефіцієнтів автокореляції рівнів першого, другого, третього і т.д. порядків називають автокореляційною функцією. Значення автокореляційної функції можуть коливатися від -1 до +1, але зі стаціонарності випливає, що $r(\tau)=-r(\tau)$. Графік автокореляційної функції називається корелограмою.

Вибірковий коефіцієнт автокореляції обчислюється за формулою:

$$r(\tau)=\frac{\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(y_t-\bar{y})\cdot(y_{t+\tau}-\bar{y})}{\frac{1}{n-1}\sum_{t=1}^{n}(y_t-\bar{y})^2}. \qquad (27)$$

Коефіцієнт автокореляції рівнів ряду першого порядку, що вимірює залежність між сусідніми рівнями ряду y_t та y_{t-1}, тобто при лагу 1, розраховується за формулою:

$$r_1=\frac{\sum_{t=2}^{n}(y_t-\bar{y}_1)\cdot(y_{t-1}-\bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t-\bar{y}_1)^2\cdot\sum_{t=2}^{n}(y_{t-1}-\bar{y}_2)^2}}, \qquad (28)$$

де $\bar{y}_1=\frac{\sum_{t=2}^{n}y_t}{n-1}$; $\bar{y}_2=\frac{\sum_{t=2}^{n}y_{t-1}}{n-1}$.

Аналогічно визначаються коефіцієнти автокореляції другого і вищих порядків. Так, коефіцієнт автокореляції другого порядку характеризує тісноту зв'язку між рівнями y_t та y_{t-2} і визначається за формулою:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)\cdot(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}}, \qquad (29)$$

де $\bar{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}$; $\bar{y}_4 = \frac{\sum_{t=3}^{n} y_{t-2}}{n-2}$.

Для обчислення коефіцієнта автокореляції за формулою (26) в Excel можна скористатися функцією **КОРРЕЛ**. Припустимо, що базова змінна включає діапазон A1 : A34 . Тоді коефіцієнт автокореляції дорівнює:

= **КОРРЕЛ** (A1 : A33 ; A2 : A34) .

На практиці, як правило, при обчисленні автокореляції використовується формула (27). Аналіз автокореляційної функції і корелограми дозволяє визначити лаг, за якого автокореляція є найбільш високою, тобто за допомогою аналізу автокореляційної функції і корелограми можна виявити структуру ряду.

Тому коефіцієнт автокореляції рівнів і автокореляційну функцію доцільно використовувати для виявлення в часовому ряді наявності або відсутності трендової компоненти і сезонної компоненти.

Приклад. *Аналіз часового ряду валового внутрішнього продукту*

Валовий внутрішній продукт є на стадії виробництва сумою доданих вартостей галузей економіки, а на стадії використання – вартістю товарів і послуг, призначених для кінцевого споживання, накопичення й експорту.

Як вихідна інформація використовуються дані: номінальний обсяг валового внутрішнього продукту, млрд. крб. (з 1998 р млн. крб.) – квартальні дані з 1994:1 по 2003:1 (табл. КП,2).

Таблиця 2

ВВП

Дата	4кв.1994	1кв.1995	2кв.1995	3кв.1995	4кв.1995	1кв.1996	2кв.1996	3кв.1996	4кв.1996	1кв.1997
ВВП	225.00	235.00	325.00	421.00	448.00	425.00	469.00	549.00	565.00	513.00
№	1	2	3	4	5	6	7	8	9	10

Дата	2кв.1997	3кв.1997	4кв.1997	1кв.1998	2кв.1998	3кв.1998	4кв.1998	1кв.1999	2кв.1999	3кв.1999
ВВП	555.00	634.00	641.00	551.00	602.00	676.00	801.00	901.00	1102.00	1373.00
№	11	12	13	14	15	16	17	18	19	20

Дата	4кв.1999	1кв.2000	2кв.2000	3кв.2000	4кв.2000	1кв.2001	2кв.2001	3кв.2001	4кв.2001	1кв.2002
ВВП,	1447.00	1527.00	1697.00	2038.00	2044.00	1922.00	2120.00	2536.00	2461.00	2268.00
№	21	22	23	24	25	26	27	28	29	30

Дата	2кв.2002	3кв.2002	4кв.2002	1кв.2003
ВВП	2523.00	3074.00	2998.00	2893.10
№	31	32	33	34

Графік цього ряду наведено на рис. 14.

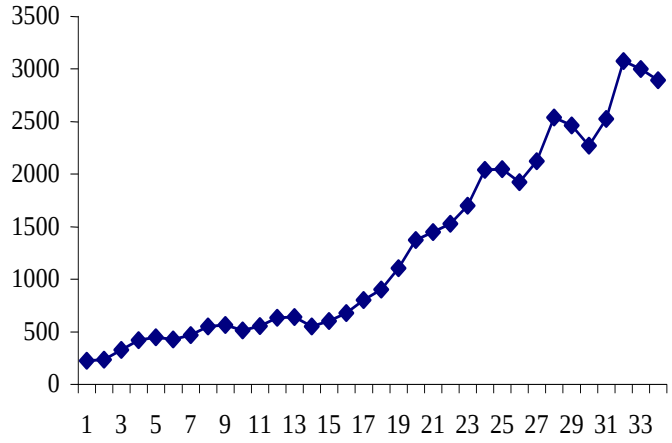


Рис. 14.

Бачимо, що дані мають зростаючий тренд. Таким чином, вже візуальний аналіз дозволяє зробити висновок про нестаціонарність вихідного часового ряду. Перевіримо це припущення, обчислимо коефіцієнти автокореляції (табл. 3) і побудуємо графік автокореляційної функції часового ряду ВВП, тобто його (корелограму) (рис. 15).

Таблиця 3

Автокореляційна функція								
Лаг	1	2	3	4	5	6	7	8
Коефіцієнти автокореляції	0.914	0.811	0.717	0.651	0.576	0.480	0.387	0.315

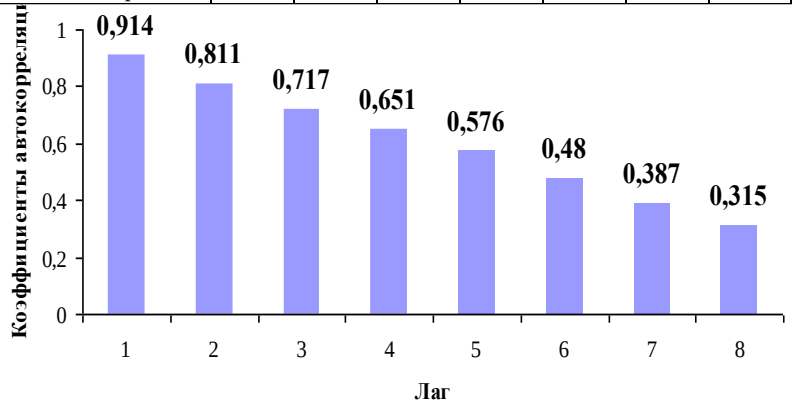


Рис. 15. Корелограма.

Корелограма автокореляційної функції у разі стаціонарного часового ряду повинна швидко спадати з ростом t після кількох перших значень. З рисунка видно, що досліджуваний ряд не є стаціонарним. Часовий ряд валового внутрішнього продукту містить трендовую компоненту

4. Ієрархічний агломеративний кластерний аналіз багатовимірних даних

В практичних дослідженнях різноманітних явищ, процесів, ситуацій, об'єктів дані спостережень використовують для отримання корисної і застосовної на практиці інформації переважно для прийняття рішень. Для цього, в першу чергу, здійснюють попереднє опрацювання існуючих даних, яке полягає у поділі даних на

однорідні групи. В результаті такого поділу пошук закономірностей в такій групі стає більш коректним. Основний метод, яким здійснюється такий поділ, є процедура кластерного аналізу. Ідея класифікації отриманих даних щодо деяких, подібних між собою об'єктів, за допомогою кластерного аналізу є за своїм змістом інтуїтивною, якщо припустити, що в n-вимірному просторі ознак існує деяка метрика, за якою ці об'єкти можна згрупувати в окремі групи – кластери.

Кластерний аналіз є одним із методів багатовимірного статистичного аналізу, тобто аналізу даних, коли кожне спостереження подається не одним деяким показником, а сукупністю значень різних показників. Він включає низку алгоритмів, за допомогою яких здійснюється і формування самих кластерів і розподіл об'єктів за кластерами. Кластерний аналіз, перш за все, розв'язує задачу внесення структури в дані, тобто їх групової однорідності, а також забезпечує виділення компактних, віддалених одна від одної груп об'єктів, тобто шукає «природне» розбиття сукупності на області скупчення об'єктів.

Методи багатовимірного аналізу – найбільш діючий кількісний інструмент дослідження процесів, які описані більшою кількістю характеристик. До них відносяться: кластерний аналіз, таксономія, розпізнавання образів.

Кластерний аналіз найбільше яскраво відбиває риси багатовимірного аналізу в сенсі класифікації конкретних об'єктів. Його процедури, а існує багато різновидностей кластерного аналізу, необхідні в тому випадку, коли структуру даних неможливо виявити ні візуально, ні за допомогою експертів. Велика перевага кластерного аналізу в тому, що він дозволяє здійснювати розбиття об'єктів не за одним параметром, а за цілим набором ознак. Крім того, на відміну від більшості математико-статистичних методів він не накладає будь-яких обмежень на вид об'єктів, що підлягають кластеризації.

Кластерний аналіз дозволяє розглядати досить значні обсяги даних, різко скорочувати і стискати їх, робити їх компактними та наочними. Важливе значення метод має стосовно сукупностей часових рядів, які характеризують динаміку розвитку. Іншими словами, стосовно часових рядів, він дозволяє виділяти періоди, в яких значення відповідних показників були достатньо близькими, а також визначати групи часових рядів з найбільш подібною динамікою. Іноді підхід кластерного аналізу називають у літературі чисельною таксономією, чисельною класифікацією, класифікацією із самонавчанням. Перше застосування кластерний аналіз знайшов у соціології. Назва кластерний аналіз походить від англійського слова cluster (гроно, скупчення). Вперше в 1939 році був визначений предмет кластерного аналізу та зроблено його опис дослідником Тріоном. Кластерний аналіз можна застосовувати до інтервальних даних, частот та бінарних даних.

Головне призначення кластерного аналізу – вирішення таких основних задач:

- розробка типології або класифікації;
- виявлення відповідної структури в даних;
- дослідження корисних концептуальних схем групування об'єктів;
- висунення гіпотез на основі дослідження даних;
- перевірка гіпотез або дослідження з метою визначити, чи дійсно типи (групи), виділені тим або іншим способом присутні в існуючих даних;
- розбиття множини досліджуваних об'єктів і ознак на однорідні, у відповідному розумінні, групи або кластери.

Незалежно від предмета дослідження застосування кластерного аналізу включає наступні етапи:

- відбір даних для кластеризації та подання їх у вигляді таблиці «об’єкт-властивість»;
- нормування даних, які подіх таблицею «об’єкт - властивість»;
- вибір та обґрунтування метрики для формування матриці близькостей;
- побудова матриці близькостей на основі нормованої таблиці «об’єкт - властивість»;
- вибір стратегії об’єднання для процедури кластерного аналізу;
- реалізація процедури кластерного аналізу стосовно матриці близькості;
- побудова дендрограми та виділення за відповідними критеріями потрібних кластерів.

Кластерний аналіз ставить такі вимоги до даних:

- показники не повинні корелювати між собою;
- показники властивостей мають бути безрозмірними;
- вплив будь-яких факторів на значення показників має бути виключений.

Методи кластерного аналізу можна застосовувати в різних випадках, у тому разі, коли мова йде про просте угруповання, в якому усе зводиться до утворення груп по кількісній подібності. В залежності від конкретної прикладної задачі мета кластерного аналізу може бути різна, наприклад:

– зрозуміти структуру множини об’єктів, розбивши їх на однорідні, в тому чи іншому сенсі, групи і, тим самим, спростити подальше опрацювання даних для прийняття рішень, працюючи з кожним кластером окремо;

– виділити нетипові об’єкти, які не належать до жодного з кластерів. Цю задачу називають однокласовою класифікацією виявлення нетиповості або новизни;

– зменшити, у випадку надвеликих вибірок X , обсяг даних для збереження, залишивши по одному, найбільш характерному представнику від кожного кластера;

– дослідити динаміку об’єктів в процесі їх експлуатації за зміною відстаней всередині класів та між класами.

Переваги кластерного аналізу:

а) дозволяє робити розбиття об'єктів не по одному параметру, а за цілим набором ознак;

б) кластерний аналіз, на відміну від більшості математико-статистичних методів, не накладає ніяких обмежень на вид об'єктів, що розглядаються, і дозволяє досліджувати множину вихідних даних практично довільної природи;

в) дозволяє розглядати досить великий обсяг інформації й різко скорочувати, стискати більші масиви інформації, робити їх компактними й наочними;

г) кластерний аналіз можна використати циклічно: у цьому випадку дослідження проводиться доти, поки не будуть досягнуті необхідні результати. При цьому кожен цикл тут може надавати інформацію, що здатна сильно змінити спрямованість і підходи подальшого застосування кластерного аналізу. Цей процес можна представити системою зі зворотним зв'язком;

д) в задачах прогнозування досить перспективним є поєднання кластерного аналізу з іншими кількісними методами (наприклад, з регресійним аналізом).

Недоліки й обмеження кластерного аналізу:

а) склад і кількість кластерів залежить від обраних критеріїв розбиття;

б) при приведенні вихідного масиву даних до більш компактного вигляду можуть виникати певні перекручення, а також можуть губитися індивідуальні риси окремих об'єктів через заміну їхніми характеристиками узагальнених значень параметрів кластера;

в) при проведенні класифікації об'єктів часто ігнорується можливість відсутності в розглянутій сукупності яких-небудь значень кластерів.

У кластерному аналізі вважається, що:

а) обрані характеристики допускають, в принципі, бажане розбиття на кластери;

б) одиниці виміру (масштаб) обрані правильно.

Вибір масштабу відіграє велику роль. Як правило, дані нормалізують (обчисленням середнього й діленням на стандартне відхилення) так, що дисперсія виявляється рівною одиниці.

Методи кластерного аналізу можна застосовувати навіть тоді, коли необхідно здійснити звичайний поділ множини об’єктів на групи лише за кількісною подібністю. Алгоритми кластерного аналізу мають розроблену програмну реалізацію, що дозволяє розв’язувати задачі великої розмірності. Метод ієрархічного кластерного аналізу (інколи – чисельної таксономії) існуючих даних здійснює класифікацію, яка раніше не існувала, або створює нову, ігноруючи попередню, переглядаючи дані знову.

Вихідні дані. Для кластерного аналізу дані, що зібрані в процесі дослідження чи отримані експериментально подають у формі таблиці «об’єкт - властивість», в такий спосіб, що першим стовпчиком є назви об’єктів, які підлягають групуванню, а решта стовпчиків відповідають конкретним властивостям (ознакам, характеристикам, показникам тощо) і містять їх конкретні значення.

Задача. Полягає в тому, щоб на підставі даних представлених таблицею «об’єкт - властивість», розбити множину $G = \{ g_i : g \in G,\ i = 1,\ 2,\ \dots,\ m \}$ цих об'єктів на k (k – ціле) кластерів – неперетинних підмножин $Q_1, Q_2, \dots, Q_k$ так, щоб $\bigcup_{j=1}^{k} Q_j = G$, а $\bigcap_{j=1}^{k} Q_j = \varnothing$, тобто:

а) кожен об'єкт g_i має належати одній і тільки одній підмножині розбиття;

б) об'єкти, що належать одному і тому самому кластеру, були подібними;

в) об'єкти, що належать до різних кластерів були відмінними.

Результат. За отриманими даними має бути побудована дендрограма і приведена її інтерпретація.

***Виконання даної роботи полягає** у реалізації її двох частин, а саме, побудови матриці близькості на підставі таблиці об’єкт-властивість та проведення самого кластерного аналізу на основі побудованої матриці близькостей.*

Частина І. Побудова матриці близькості.

Ілюстрацію процедури агломеративного ієрархічного кластерного аналізу доцільно провести на конкретному прикладі, з відповідними поясненнями та обґрунтуваннями, використовуючи для розрахунків табличний процесор Ms Excel.

Як приклад використано результати тренажерної підготовки операторського персоналу систем опрацювання аерокосмічних зображень, які використовуються в задачах виявлення на них об’єктів заданого класу. В результаті опрацювання одного і того ж сценарію, який полягав у поданні на монітор послідовності тестових зображень, для кожного оператора отримані дані про час опрацювання кожного тесту, тобто пошуку і виявлення та ідентифікації об’єкта на тестовому прикладі цього сценарію. Значення часу є випадковими величинами з асиметричним одномодальним розподілом і поданими у вигляді індивідуальних часових рядів.

Формування таблиці «об’єкт- властивість». В результаті первинного опрацювання даних з використанням описової статистики в межах показників

визначених для цього в Ms Excel побудована таблиця, яку називають таблицею «об'єкт-властивість». В цьому випадку – це таблиця «оператор-індивідуальні показники» розміром $n \times m$, де $n = \overline{1, 13}$ – кількість операторів, а $m = \overline{1, 14}$ – кількість використаних показників описової статистики. Отже, для кластерного аналізу подається множина G, яка включає m об'єктів, кожен з яких характеризується n ознаками. Дані представлені нижче в таблиці 4.

Таблиця 4

Таблиця «оператор - показник» – значень показників об'єктів за описовою статистикою

	Показники													
Об'єкти	**Середнє**	**Стнд. пом.**	**Медіана**	**Мода**	**Стнд. відх.**	**Дисперсія**	**Ексцес**	**Асиметрія**	**Розмах**	**Мінімум**	**Максимум**	**Сума**	**К-сть даних**	**Надійність**
Bur	1886,2	26,68	1839,5	1697	322,4	103921	3,837	1,257	2290	1196	3486	275378	146	52,7
Cub	678,9	17,57	636,0	708	187,6	35189	0,790	0,995	976	376	1352	77395	114	34,8
Cup	1506,5	22,74	1484,0	1663	245,9	60482	-0,267	0,516	1130	1050	2180	176263	117	45,0
Hav	753,7	16,74	726,5	787	169,1	28600	1,714	1,208	892	483	1375	76878	102	33,2
Hod	709,5	16,62	674,0	714	171,1	29272	1,742	1,261	900	462	1362	75204	106	32,9
Kol	430,2	8,96	415,0	429	105,3	11089	6,887	2,169	722	241	963	59372	138	17,7
Lob	665,0	18,36	622,5	601	179,9	32354	0,828	1,037	845	355	1200	63840	96	36,4
Lot	608,2	15,81	593,0	623	178,2	31751	2,029	1,163	981	347	1328	77246	127	31,3
Oli	655,3	16,69	619,0	609	166,0	27563	3,005	1,503	890	396	1286	64872	99	33,1
Per	728,3	8,19	717,0	705	96,5	9313	7,528	1,662	769	504	1273	101229	139	16,2
Pon	675,9	14,43	651,5	518	165,8	27489	0,451	0,754	833	395	1228	89213	132	28,5
Sol	1633,1	28,83	1588,0	1446	311,9	97268	0,590	0,771	1524	1117	2641	191070	117	57,1
Syr	704,4	19,03	645,5	606	215,3	46362	1,952	1,325	1176	318	1494	90165	128	37,7

Наведені в табл. 4 показники за величиною і розмірністю є дуже неоднорідними, а це означає неможливість обґрунтованої інтерпретації отриманого результату кластерного аналізу. Це приводить до того, що величина, отриманих в результаті кластеризації відстаней між точками, що відображають положення об'єктів у просторі їхніх властивостей, визначатиметься довільно обраним масштабом. Щоб усунути неоднорідність виміру вихідних даних, всі їх значення попередньо нормуються, тобто виражаються через відношення цих значень до деякої величини, що відображає певні властивості цього показника. Важливо, щоб усі значення змінних змінювалися в порівняльних шкалах. Крім того, за наявності значного розкиду величин значень необхідно звести їх до інтервалу [0,1], тобто здійснити їх нормування.

Нормування таблиці «об'єкт-властивість». Основні труднощі полягають у виборі способу нормування, оскільки, необхідно максимально врахувати якісну специфіку показників. Нормування використовують в тому випадку, коли ознаки представлені різними шкалами і мають різний фізичний зміст (розмірність). Нормування – це перехід до деякого однакового опису для всіх ознак і введення нової умовної одиниці вимірювання, яка допускає формальне зіставлення об'єктів чи їх ознак. Нормування приводить до безрозмірних величин, зберігаючи за ними відповідність ознакам. Перехід від початкових значень x до нормованих z здійснюють за формулами:

$$z = \frac{x - x_{\text{сер}}}{s}, \qquad z = \frac{x}{x_{\text{етал}}}, \qquad z = \frac{x}{x_{\text{сер}}}, \tag{30}$$

$$z = \frac{x}{x_{\max}}, \qquad z = \frac{x - x_{\text{сер}}}{x_{\max} - x_{\min}}, \qquad z = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

де s – середньоквадратичне відхилення.

Найбільш поширеним нормуванням є лінійне перетворення, яке проводять за модифікованою останньою формулою.

Визнання рівноцінності різних показників не завжди є виправданим. Очевидно, що великі значення дисперсії, суми та їх варіації будуть суттєво впливати на результат кластерного аналізу.

В результаті нормування за цією формулою всі дані табл. 4 є приведені до одиничного інтервалу, тобто ця таблиця прийме вигляд табл. 5. Нормування здійснюється по кожному стовпчику окремо, тобто нормуються ознаки одного типу.

Таблиця 5.

Таблиця «оператор - показник» – після нормування

	Показники													
Оператори	**Середнє**	**Стнд. пом.**	**Медіана**	**Мода**	**Стнд. Відх.**	**Дисперсія**	**Ексцес**	**Асиметрія**	**Розмах**	**Мінімум**	**Максимум**	**Сумма**	**К-сть**	**Надійність**
Bur	1,000	0,896	1,000	1,000	1,000	1,000	0,527	0,447	1,000	1,000	1,000	1,000	1,000	0,893
Cub	0,171	0,454	0,155	0,220	0,403	0,274	0,136	0,288	0,162	0,141	0,154	0,083	0,360	0,455
Cup	0,739	0,705	0,750	0,973	0,662	0,541	0,000	-0,002	0,260	0,847	0,482	0,541	0,420	0,705
Hav	0,222	0,414	0,219	0,282	0,321	0,204	0,254	0,417	0,108	0,253	0,163	0,081	0,120	0,416
Hod	0,192	0,408	0,182	0,225	0,330	0,211	0,258	0,449	0,114	0,231	0,158	0,073	0,200	0,410
Kol	0,000	0,038	0,000	0,000	0,039	0,019	0,918	1,000	0,000	0,000	0,000	0,000	0,840	0,038
Lob	0,161	0,493	0,146	0,136	0,369	0,244	0,141	0,313	0,078	0,119	0,094	0,021	0,000	0,495
Lot	0,122	0,369	0,125	0,153	0,362	0,237	0,295	0,390	0,165	0,111	0,145	0,083	0,620	0,369
Oli	0,155	0,412	0,143	0,142	0,308	0,193	0,420	0,596	0,107	0,162	0,128	0,025	0,060	0,414
Per	0,205	0,000	0,212	0,218	0,000	0,000	1,000	0,692	0,030	0,275	0,123	0,194	0,860	0,000
Pon	0,169	0,302	0,166	0,070	0,307	0,192	0,092	0,142	0,071	0,161	0,105	0,138	0,720	0,302
Sol	0,826	1,000	0,823	0,802	0,954	0,930	0,110	0,152	0,511	0,917	0,665	0,610	0,420	1,000
Syr	0,188	0,525	0,162	0,140	0,526	0,392	0,285	0,488	0,290	0,081	0,210	0,143	0,640	0,525

Зауважимо, що нормування означає, що всі ознаки є рівноцінними з погляду з'ясування подібності розглянутих об'єктів. Інколи, поряд з нормуванням надають кожному з показників вагу і тим самим вказують на його значущість під час встановлення подібності і відмінностей між об'єктами.

Вибір метрики для побудови матриці близькості. Якщо відстань між об’єктами природно трактувати як міру відмінності об’єктів, то обернену величину $B = \frac{1}{D}$ можна розглядати як міру подібності (близькості) об’єктів.

1. Найчастіше відстань $D(\vec{x}_1, \vec{x}_2)$ між об’єктами вимірюють в евклідовій матриці, яка найбільш узгоджена з нашими інтуїтивними представленнями про близькість об’єктів і визначається за формулою

$$D_E(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_{i=1}^{N} (x_{1j} - x_{2j})^2} \quad , \tag{31}$$

де x_{1j} та x_{2j} – значення j-ї компоненти в описах 1-го та 2-го об'єктів, N – розмірність простору ознак, а в даному випадку $N = n$.

2. Широко використовується лінійна метрика (метрика міських кварталів або манхетенська метрика), яка задає відстань

$$D_M(\vec{x}_1\, \vec{x}_2) = \sum_{i=1}^{N} \left| \vec{x}_{1j} - \vec{x}_{2j} \right|, \tag{32}$$

а також sup-норма, яка визначає чебишевську відстань

$$D_Ч(\vec{x}_1, \vec{x}_2) = \max_j \left| x_{1j} - x_{2j} \right|. \tag{33}$$

Очевидно, що $D_Ч < D_E < D_M$ з ростом розмірності N манхетенська відстань D_M підкреслює, а чебишовська згладжує відмінності між об'єктами. Відстані $D_M, D_E, D_Ч$ визначаються для тих описів, в яких ознаки виражені кількісними шкалами і є частинними випадками узагальнення відстані Мінковського $D_p(\vec{x}_1, \vec{x}_2) = \left(\sum_{i=1}^{N} \left| x_{1j} - x_{2j} \right|^p \right)^{\frac{1}{p}}$, тобто відстані $D_M, D_E, D_Ч$ визначаються значенням степеня $p = 1, 2, , \infty$. Існує досить велика різноманітність мір, але на практиці ці три міри є найуживанішими.

Евклідова відстань ефективна при дослідженні слабко кореляційних сукупностей об'єктів (кулеподібні класи), а манхетенська тоді, коли об'єкти утворюють плоскі витягнені класи, ортогональні до будь-яких координатних осей простору ознак. Тому опрацювання однієї і тієї ж сукупності даних одним і тим же методом або алгоритмом, але з використанням різних метрик може дати різні, інколи кардинально протилежні, результати. Отже, до вибору метрики слід підходити дуже продумано і обережно, співставляючи результати використання різних метрик між собою і з цілями опрацювання даних, яке здійснюємо. Якщо ознаки представляються в якісних шкалах, зокрема в шкалах найменувань та порядку, використовують відстань Геммінга

$$D_H(\vec{x}_1, \vec{x}_2) = \frac{1}{N} \sum_{j=1}^{N} \left| x_{1j} - x_{2j} \right|, \tag{34}$$

для якої відмінності виражаються числом неспівпадінь властивостей порівнюваних об'єктів. У випадку якісних шкал ознаки розглядаються як бінарні, тобто такі, що можуть приймати лише два значення "0" та "1". Відстань Геммінга D_H є максимальною і рівна 1 для об'єктів з протилежними за значеннями описами, тобто елементи одного опису є протилежними до відповідних елементів опису другого об'єкту. Для об'єктів, всі ознаки яких (з числа включених в опис) співпадають, $D_H = 0$. При виборі виду міри близькості необхідно врахувати їх формальні властивості й порівняти їх зі змістовними особливостями задачі. В результаті застосування будь-якої з цих метрик до даних отримують матрицю близькості,

розмірність якої $M \times M$, а за своєю специфікою вона є симетричною відносно головної діагоналі.

Побудова таблиці близькості. Для побудови матриці близькості за допомогою табличного процесора Ms Excel використовують дані табл. 5 в такий спосіб.

1. На робочому листі поміщають табл. 5 та її копію так, щоб номери стрічок в них були однакові (наприклад, якщо перший елемент табл. 5 – таблиці-оригіналу розміщений в комірці **B3**, то перший елемент таблиці-копії цієї таблиці є розміщений в комірці **R3)**.

2. Далі будують матрицю близькості. При цьому виникає задача вибору міри близькості. Найчастіше відстань $D(\vec{x}_1, \vec{x}_2)$ між об'єктами вимірюють в евклідовій метриці, її ще називають *евклідовою відстанню* і обчислюють за такою формулою:

$$d_E = \sqrt{\sum_{p=1}^{q} \left(x_{ip} - x_{jp}\right)^2} \ . \qquad (35)$$

Евклідова метрика, є найбільш узгодженою з нашими інтуїтивними представленнями про близькість об'єктів.

3. В результаті застосування до матриці даних цієї метрики отримують матрицю близькості, розмірність якої $M \times M$, де $p = 1, 2, \ldots, q$ – номер ознаки, i, j – індекси пари об'єктів, між якими визначають відстань. Матриця відстаней за своєю специфікою є симетричною відносно головної діагоналі.

В Excel матрицю близькості будують так. Визначають комірку для розміщення першого елемента матриці близькості, наприклад, **B18**, тобто нижче на кілька стрічок під таблицею, в якій розміщують (для даного випадку) таку формулу

=**КОРЕНЬ**(СУММ((B3-$R3)^2+($C$3-$S3)^2+(D3-$T3)^2+($E$3-$U3)^2+(F3-$V3)^2+($G$3-$W3)^2+(H3-$X3)^2+($I$3-$Y3)^2+(J3-$Z3)^2+($K$3-$AA3)^2+(L3-$AB3)^2+($M$3-$AC3)^2+(N3-$AD3)^2+($O$3-$AE3)^2))

і далі, шляхом автозаповнення в M клітинок стовпчика **B** будуть записані відповідні значення першого стовпчика матриці близькості. В результаті в комірці **B19** формула буде мати вигляд

=**КОРЕНЬ**(СУММ((B3-$R4)^2+($C$3-$S4)^2+(D3-$T4)^2+($E$3-$U4)^2+(F3-$V4)^2+($G$3-$W4)^2+(H3-$X4)^2+($I$3-$Y4)^2+(J3-$Z4)^2+($K$3-$AA4)^2+(L3-$AB4)^2+($M$3-$AC4)^2+(N3-$AD4)^2+($O$3-$AE4)^2)),

тобто, номер і значення стрічок таблиці-оригіналу залишаться без змін, а зміняться лише номери стрічок таблиці-копії. В результаті, буде визначено відстань між першим об'єктом, визначеним першою стрічкою таблиці-оригіналу, і кожним об'єктом, визначеним стрічкою таблиці-копії.

4. Для визначення відстані між другим об'єктом таблиці–оригіналу і всіма іншими об'єктами таблиці-копії копіюють формулу з комірки **B18** в комірку **C18**. Після активізації формули в цій комірці, змінюють адресу першої стрічки таблиці-оригіналу шляхом переміщення всіх кольорових рамок на одну комірку вниз. В результаті цього формула буде мати такий вигляд

=**КОРЕНЬ**(СУММ((B4-$R3)^2+($C$4-$S3)^2+(D4-$T3)^2+($E$4-$U3)^2+(F4-$V3)^2+($G$4-$W3)^2+(H4-$X3)^2+($I$4-$Y3)^2+(J4-$Z3)^2+($K$4-$AA3)^2+(L4-$AB3)^2+($M$4-$AC3)^2+(N4-$AD3)^2+($O$4-$AE3)^2)).

Реалізуємо цю формулу і шляхом автозаповнення формуємо другий стовпчик матриці близькості. Аналогічно роблять при визначенні решти стовпчиків.

Зауваження. Якщо, виходячи з розмірності власних задач, побудувати матрицю близькості для максимально можливих кількостей об'єктів і ознак, то використовуючи її як шаблон, можна знайти значення матриць близькості для будь-якої кількості об'єктів і ознак, замінюючи в таблиці-оригіналі і в таблиці-копії їхні значення і відкидаючи порожні комірки шаблону, отримуємо безпосередньо потрібну матрицю близькості. Іншими словами, маючи лист Excel з один раз зробленою процедурою побудови матриці близькості можна в таблиці ввести інші дані і, відкидаючи зайві або додаючи нові стовпчики і стрічки безпосередньо отримати матрицю близькості, але для конкретної метрики.

Після формування останнього стовпчика отримуємо матрицю близькості зображену у вигляді табл. 6.

Таблиця 6.

Матриця близькості

	Bur	**Cub**	**Cup**	**Hav**	**Hod**	**Kol**	**Lob**	**Lot**	**Oli**	**Per**	**Pon**	**Sol**	**Syr**
Bur	0,000	2,620	1,547	2,651	2,669	3,292	2,836	2,656	2,812	2,942	2,749	1,112	2,422
Cub	2,620	0,000	1,562	0,359	0,303	1,440	0,397	0,360	0,547	1,365	0,502	1,956	0,461
Cup	1,547	1,562	0,000	1,563	1,619	2,634	1,716	1,734	1,835	2,275	1,728	0,771	1,625
Hav	2,651	0,359	1,563	0,000	0,118	1,408	0,335	0,565	0,326	1,302	0,745	1,981	0,686
Hod	2,669	0,303	1,619	0,118	0,000	1,328	0,346	0,466	0,293	1,249	0,674	2,021	0,602
Kol	3,292	1,440	2,634	1,408	1,328	0,000	1,566	1,141	1,234	0,608	1,340	2,998	1,338
Lob	2,836	0,397	1,716	0,335	0,346	1,566	0,000	0,679	0,430	1,541	0,807	2,091	0,765
Lot	2,656	0,360	1,734	0,565	0,466	1,141	0,679	0,000	0,626	1,087	0,391	2,116	0,374
Oli	2,812	0,547	1,835	0,326	0,293	1,234	0,430	0,626	0,000	1,232	0,891	2,188	0,735
Per	2,942	1,365	2,275	1,302	1,249	0,608	1,541	1,087	1,232	0,000	1,226	2,695	1,307
Pon	2,749	0,502	1,728	0,745	0,674	1,340	0,807	0,391	0,891	1,226	0,000	2,174	0,649
Sol	1,112	1,956	0,771	1,981	2,021	2,998	2,091	2,116	2,188	2,695	2,174	0,000	1,890
Syr	2,422	0,461	1,625	0,686	0,602	1,338	0,765	0,374	0,735	1,307	0,649	1,890	0,000

Отримана матриця близькості, є симетричною діагональною матрицею, яка вказує на величину близькості між об'єктами. На основі такої матриці проводиться агломеративний ієрархічний кластерний аналіз.

Частина ІІ. Проведення агломеративного ієрархічного кластерного аналізу.

Вибір стратегій об'єднання. Процедура кластерного аналізу базується на перерахунку значень матриці близькості і, в результаті, кожного такого кроку обчислень об'єднуються об'єкти, об'єкт з групою або дві групи. Після кожного такого об'єднання розмірність матриці зменшується на одиницю, а кількість кластерів або кількість об'єктів в конкретному кластері збільшується на одиницю. Проте такі об'єднання відбуваються не будь як довільно, а в рамках конкретно вибраної стратегії, яка діє протягом усієї процедури. Зміст такої стратегії полягає в тому, що кожен новий кластер визначався значеннями ознак, отриманими в результаті перерахунку відповідних значень ознак об'єктів і кластерів, які об'єднуються в цей новий кластер. Іншими словами, процедуру об'єднання об'єктів в кластери можна подати так.

Суть стратегії групування полягає в наступному. У випадку n об'єктів обчислюються всі $\frac{n(n-1)}{2}$ мір відмінностей і пара об'єктів з найменшою мірою об'єднується в одну групу. На наступному кроці визначають відповідну міру відмінності (нове значення близькості) між цією групою і рештою $n-2$ об'єктами, а на більш пізніх стадіях треба буде визначати цю міру між об'єктом і групою будь-якого обсягу, а також між будь-якими двома групами. На кожному кроці класифікації виконується те об'єднання (між двома об'єктами, між об'єктом і групою або між двома групами), для яких міра відмінності мінімальна серед всіх існуючих на даному кроці. Міра повинна бути такою, щоб об'єкт можна було розглядати як групу з одного елемента, Стратегія об'єднання визначається саме мірою відмінності між групами.

Нехай є дві групи i та j з n_i і n_j елементами відповідно; міру відмінності між цими групами позначимо d_{ij}. Припустимо, що d_{ij} – найменша міра з усіх, що залишились, так що i та j об'єднуються і утворюють нову групу k з $n_k = n_i + n_j$ елементами. Розглянемо деяку іншу групу h з n_h елементами. Перед об'єднанням відомі значення мір d_{hi}, d_{hj}, d_{ij} та об'ємів n_h, n_i, n_j. Значення розглянутих мір обчислюють за формулою

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \left| d_{hi} - d_{hj} \right|, \qquad (36)$$

де параметри α_i, α_j, β і γ визначають суть стратегії.

Найчастіше використовують наступні стратегії.

1. *Стратегія найближчого сусіда.* Відстань між двома групами визначається як відстань між двома найближчими елементами з цих груп. Ця стратегія монотонна і сильно стискає простір ознак, а її параметрами є $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = -0.5$.
2. *Стратегія найдальшого сусіда.* Тут відстань між двома групами визначається як відстань між двома найбільш віддаленими представниками (елементами) цих груп. Вона монотонна і сильно розтягує простір. Її параметри мають значення $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = 0.5$.
3. *Стратегія групового середнього* (середнього зв'язку). Якщо одна група складається з m_1, а друга з m_2 елементів, то відстань між цими групами в цій стратегії визначається як середнє арифметичне відстаней між елементами з цих груп. Ця стратегія є монотонною і зберігає метрику простору. Параметри стратегії рівні: $\alpha_i = \frac{n_i}{n_k}$, $\alpha_j = \frac{n_j}{n_k}$, $\beta = \gamma = 0$.
4. *Гнучка стратегія.* Може бути застосована для будь-якої міри близькості і визначається наступними обмеженнями $\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta < 1$, $\gamma = 0$. Стратегія монотонна, а її властивості повністю залежать від β. Якщо $\beta = 0$, то стратегія зберігає метрику простору. Якщо $\beta > 0$ то стратегія стискає простір, а

якщо $\beta < 0$, то розтягує. Для практичного використання для параметрів приймають наступні значення $\alpha_i = \alpha_j = 0.625$, $\beta = -0.25$.

Об'єктами класифікації можуть бути практично будь-які об'єкти. Причому стратегії класифікації, тобто чисельні методи не залежать від природи об'єктів, що класифікуються, але різні стратегії, як правило, приводять до різних результатів. Тому вибір стратегії є досить складною задачею і вимагає високої кваліфікації від фахівця. Дані, які підлягають класифікації, утворюють множину елементів, кожен з яких визначається набором ознак, які відповідають узагальненому поняттю змінної. Такі дані або множину даних вважають неоднорідними, тобто множину даних розглядають як сукупність підмножин, таких, що всередині підмножини її елементи між собою є більш подібними, ніж з будь-яким іншим елементом будь-якою іншої підмножини. В цьому розумінні виділяють два підходи до аналізу. Перший є знаходженням міри впевненості, що при використанні цієї чисельної процедури можна вважати, що існують такі підмножини (тобто якщо їх немає, то і не повинно бути їх знайдено); другий – допускає, що істотних відмінностей підмножини не мають, проте для полегшення аналізу дані все таки треба розбити штучно.

В математичному плані задача класифікації даних тобто елементів формулюється як задача побудови розбиття елементів множини даних на деяке наперед задане чи знайдене в ході аналізу число непорожніх попарно неперетинних підмножин (класів) елементів.

Проведення кластерного аналізу. Процедура, яка складає суть ієрархічної класифікації полягає в тому, що в матриці близькості з двох об’єктів, між якими найменша відстань, формують перший кластер, значення якого перераховують у відповідності з вибраною стратегією. Другий об’єкт з більшим номером стовпчика і стрічки викидається, а замість першого об’єкта (з меншим номером стовпчика і стрічки) вставляється утворений з цих об’єктів кластер з перерахованими значеннями. В результаті розмірність матриці зменшується на одиницю. На наступному кроці знову знаходять найменшу відстань між її елементами і чинять аналогічно. Коли матриця близькості матиме розмірність 2×2 процедура кластеризації припиняється. На основі отриманої на кожному кроці інформації про об’єднання кластерів і знайдені значення мінімальних відстаней будується дендрограма і подається її інтерпретація.

Суть цієї процедури полягає в тому, що: в матриці близькості, в якій стовпці і стрічки є векторами пронумерованих об’єктів, шукають найменше значення, визначають, які об’єкти відповідають цьому значенню і об’єднують їх в одну групу. Далі перераховують значення векторів цих об’єктів і подають цю групу як новий об’єкт зі своїм вектором значень.

Ця процедура реалізується за таким алгоритмом.

1. Знаходять в матриці близькості найменше значення і об’єднують об’єкти, яким воно відповідає, в одну групу.

2. Вилучають стовпчики, що належать цим об’єктам, і розміщують їх, довільно і поряд, під матрицею близькості. Залишаємо порожнім місце стовпчика першого (лівого) об’єкта і зсуваємо вліво всі стовпчики, що лежали справа від стовпчика другого об’єкта, тобто ліквідуємо порожне місце. Зсувом вгору ліквідуємо вектор стрічку другого об’єкта.

3. Вибираємо стратегію об’єднання об’єктів. Відповідно до обраної стратегії значення вилучених стовпчиків перераховуємо за відповідною формулою. В даному прикладі використано гнучку стратегію з параметрами $\alpha_i = \alpha_j = 0.625$, $\beta = -0.25$,

оскільки ця стратегія за даного значення β дещо розтягує простір, а отже, і віддаляє між собою кластери, підкреслюючи їхні індивідуальності і однорідність.

4. В отриманому в результаті перерахунку стовпчику шукають два найменших значення, які знаходяться в комірках, що відповідають коміркам з нулями стовпчиків, які перераховуються, і верхнє мінімальне значення замінюємо на нуль, а нижнє ліквідуємо зсувом вгору всіх нижче розташованих комірок в цих стовпчиках.

5. Перерахований стовпчик вставляємо на місце (порожне) першого вилученого стовпчика і перевіряємо чи його нуль лежить на головній діагоналі.

6. Копіюємо всі значення цього стовпчика, транспонуємо їх у стрічку і замінюємо нею стрічку першого вилученого стовпчика.

7. Присвоюємо новому об'єкту, утвореному в результаті об'єднання вилучених об'єктів, наступний за порядком номер. В результаті цієї операції матриця близькості зменшується на один стовпчик і одну стрічку, і в ній з'являється новий об'єкт.

8. Цю процедуру повторюють до тих пір поки матриця не зменшиться до розміру 2×2. Результати кластерного аналізу наведені в табл. 7.

Таблиця 7.

Результати кластерного аналізу операторського персоналу

	Процедура об'єднання кластерів		
Кроки	**Об'єднання**	**Вузол**	**Метрика**
1	4+5	**d14**	0,118
2	14+9	**d15**	0,357
3	2+8	**d16**	0,36
4	15+7	**d17**	0,427
5	16+13	**d18**	0,432
6	18+11	**d19**	0,59
7	6+10	**d20**	0,608
8	3+12	**d21**	0,771
9	19+17	**d22**	1,038
10	1+21	**d23**	1,469
11	22+20	**d24**	2,368
12	23+24	**d25**	5,059

Побудова дендрограми. Візуалізація результатів кластерного аналізу здійснюється з допомогою дендрограми, тобто графічного зображення результатів процесу послідовної кластеризації, яку проводять в термінах матриці близькості. За допомогою дендрограми можна графічно або геометрично зобразити процедуру кластеризації за умови, що ця процедура оперує тільки з елементами матриці відстаней або подібності. Вид дендрограми залежить від вибору міри подібності або відстані між об'єктом і кластером та методу кластеризації. Найбільш важливим моментом є вибір міри подібності або міри відстані між об'єктом і кластером. Не дивлячись на те, що існує досить велика кількість різноманітних програмних засобів проведення кластерного аналізу, наприклад, ППП «Statistica» та SPSS, а також різних додатків до статистичних пакетів програм опрацювання даних, їх використання часто є дуже складним в сенсі знаходження найбільш відповідного, його придбання, узгодження з використовуваними засобами тощо.

Проте побудова дендрограми кластеризації декількох десятків об'єктів легко може бути здійснена вручну безпосередньо або за дві три ітерації в чорновому варіанті і остаточно сформована в, практично, будь якому графічному середовищі. Процедура побудови дендрограми вручну є нескладною і потребує лише уважності,

використовує результати кластерного аналізу приведені в табл. 7, складається з побудови ескізу, на підставі якого здійснюється її графічне редагування та масштабування відстаней між між об'єктами і вузлами та включає такі кроки.

1. Побудова ескізу дендрограми. Дендрограму будують починаючи з «кореня дерева», тобто з вузла останнього об'єднання, відобразивши його точкою, відміченою його номером.

2. Оскільки кластерний аналіз є дихотомічною процедурою, то з кореневого вузла проводять дві гілки, які закінчуються точками наступних вузлів. Помічаються номери вузлів.

3. Якщо гілка закінчується на об'єкті, то її кінець помічають номером відповідного об'єкта, якщо ж вузлом, то вказують відповідний йому номер.

4. Після того, як визначені всі гілки і помічені усі об'єкти, здійснюють графічну корекцію і масштабування. Для цього в декартовій системі координат по осі абсцис відкладають значення відстаней об'єднання об'єктів, об'єктів з групами та груп, а гілки будують прямими лініями, довжини яких відповідають значенням відстаней. В результаті отримують дендрограми зображені на рис. 15.

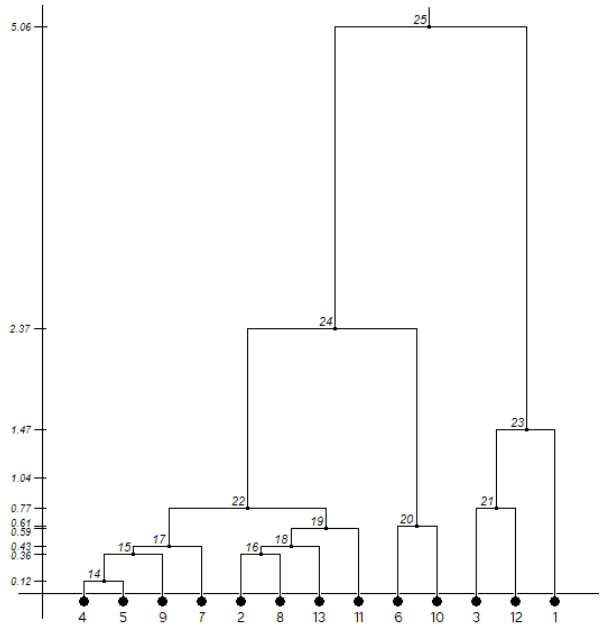


Рис. .15. Дендрограма.

Інтерпретація дендрограми. Проведення горизонтальних ліній в площині дендрограми на заданій висоті, в даному випадку, дозволяє виділити окремі кластери. Вже з першого погляду на дендрограму можна вказати, що на рівні 0.75 маємо в принципі шість досить чітко відображених кластерів, в які входять такі об'єкти.

1 кластер – об'єкти 4, 5, 9, 7;
2 кластер – об'єкти 2, 8, 13, 11;
3 кластер – об'єкти 6, 10;
4 кластер – об'єкт 3;
5 кластер – об'єкт 12;
6 кластер – об'єкт 1;

На рівні 1.0 маємо 4 кластери.

1 кластер – об'єкти 4, 5, 9, 7, 2, 8, 13, 11;
2 кластер – об'єкти 6, 10;

3 кластер – об’єкти 3, 12;
4 кластер – об’єкт 1;

На рівні 1.5 отримуємо три кластери, в які входять такі об’єкти:

1 кластер – об’єкти 4, 5, 9, 7, 2, 8, 13, 11;
2 кластер – об’єкти 6, 10;
3 кластер – об’єкти 3, 12;

5. Порядок роботи

Мета роботи – ознайомлення з основними методами візуалізації – графічного відображення та первинного статистичного опрацювання числових даних, які представлені вибірковою сукупністю або часовим рядом. Проведення процедури агломеративного ієрархічного кластерного аналізу.

Для досягнення поставленої мети в роботі вирішені такі завдання.

1. Проведені експериментальні дослідження, а зібрані дані, сформовані у файл *.dat (або *.txt).

2. Для отриманих даних побудована звітна таблиця. (Підписи, написи, шрифт, форматування тощо).

3. Для наочного подання отримані дані візуалізовані в декартовій та полярній системах координат (підписи, написи, шрифт, форматування тощо).

4. Кількісні значення характеристик отриманих даних приведені у таблицях описової статистики.

5. Інформація про розподіл даних як випадкових величин подається характером гістограми.

6. Наочне подання графіка функції розподілу даних здійснено за допомогою емпіричної кумуляти.

7. Вихідні дані, подані у вигляді таблиці «об’єкт-властивість».

8. Повна формула для відстані в математичній символіці.

9. Таблиця близькості з виділеним найменшим значенням.

10. Назва, параметри та формула стратегії об’єднання.

11. Таблиця «об’єднання – вузол – метрика».

12. Побудована дендрограма.

13. Проведена інтерпретація результатів.

14. Висновок.

Умови для завдань і проведення агломеративного ієрархічного кластерного аналізу здійснюється за вказаними: формулою нормування, метрикою та стратегією, що наведені в табл. 8.

Таблиця 8.

№	Формула нормування	Вид відстані (метрика)	Стратегія
1	$z = \dfrac{x - x_{\text{сер}}}{s}$	$D_E(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_{i=1}^{N} (x_{1j} - x_{2j})^2}$	*Стратегія найближчого сусіда.* $\alpha_i = \alpha_j = 0.5,$ $\beta = 0,\ \gamma = -0.5.$
2	$z = \dfrac{x}{x_{\text{етал}}}$	$D_M(\vec{x}_1\ \vec{x}_2) = \sum_{i=1}^{N} \left\| \vec{x}_{1j} - \vec{x}_{2j} \right\|$	*Стратегія найдальшого сусіда.* $\alpha_i = \alpha_j = 0.5,$

№	Формула нормування	Вид відстані (метрика)	Стратегія
			$\beta = 0,\ \gamma = 0.5.$
3	$z = \frac{x}{x_{\text{сер}}}$	$D_{ч}(\vec{x}_1, \vec{x}_2) = \max_j \left\| x_{1j} - x_{2j} \right\|$	*Стратегія групового середнього* $\alpha_i = \frac{n_i}{n_k},\ \alpha_j = \frac{n_j}{n_k},$ $\beta = \gamma = 0.$
4	$z = \frac{x}{x_{\max}}$		*Гнучка стратегія.* $\alpha_i + \alpha_j + \beta = 1,$ $\alpha_i = \alpha_j,\ \beta < 1,$ $\gamma = 0.$
5	$z = \frac{x - x_{\text{сер}}}{x_{\max} - x_{\min}}$		
6	$z = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$		

6. Хід роботи

1. Побудова звітної таблиці

1. Ввести дані з *.dat або *.txt. файлу.
2. Сформувати звітну таблицю з мінімальною кількістю порожніх комірок.
3. Побудувати графіки даних в декартовій та полярній системі координат.
4. Визначити кількісні характеристики описової статистики.
5. Подати дані гістограмою.
6. Подати дані кумулятами: за даними гістограми та за інтегральним процентом.
7. Подати результати роботи у відповідній формі звітності.

2. Для досягнення поставленої мети, а саме: ознайомлення з основними методами висвітлення тенденції поведінки досліджуваного показника, яка представлена характером його тренду, за допомогою методів згладжування часових рядів та візуалізації отриманих даних необхідно провести такі дослідження.

1. Згладжування за формулами з Кендела – простого ковзного середнього.

Особливість цих формул полягає в тому, що з їх допомогою просто розраховувати втрачені рівні на початку та вкінці згладженого ряду.

а) згладжуємо дані, використовуючи розміри інтервалу згладжування w = 3, 5, 7, 9, 11, 13, 15. Маємо отримати підряд сім стовпчиків;

б) згладжуємо дані, використовуючи розмір інтервалу згладжування w = 3, далі, отримані згладжені дані згладжуємо знову, але використовуємо розмір інтервалу згладжування w = 5. Продовжуємо згладжування отриманих даних інтервалом згладжування w = 7 і так до w = 15. Маємо отримати підряд сім стовпчиків.

в) в обох випадках знайти для кожного згладжування кількість поворотних точок та коефіцієнти кореляції між оригінальними значеннями та згладженими.

2. Згладжування за формулами з Полларда. Тут, в залежності від розміру інтервалу згладжування, вага для центрального рівня змінюється. Згладжування провести аналогічно до попереднього пункту.
3. Експоненціальне згладжування. Основним параметром експоненціального згладжування є параметр α, який приймає значення в інтервалі $0.1 \le \alpha \le 0.3$. Необхідно, здійснити згладжування того самого ряду зі значеннями параметра α = 0.1, 0.15, 0.2, 0.25, 0.3 і у всіх цих випадках знайти для кожного згладжування кількість поворотних точок та коефіцієнти кореляції між оригінальними значеннями та згладженими.
4. Медіанне згладжування. В цьому випадку використати ті ж самі розміри інтервалу згладжування та операції як в пункті 2.
5. Отримані результати подати у формі зведених графіків, діаграм та таблиць.

3. Кореляція даних

1. Побудова кореляційного поля. Як вхідні дані використати звітну таблицю з файлу *.dat (або *.txt) (дані повинні бути впорядковані певним чином).
2. Визначення значення коефіцієнта кореляції.
3. Обчислити кореляційне відношення.
4. Побудувати графіки автокореляційних функцій.
5. Розбити одну з послідовностей на три рівні частини.
6. Побудувати для них кореляційну матрицю.
7. Знайти коефіцієнти множинної кореляції.
8. Автокореляція.

4. Провести кластер ний аналіз даних

2. Сформувати з даних таблицю «об’єкт-властивість».

3. Утворити з них близько розташовані «таблицю-оригінал» та «таблицю-копію».

4. Вибрати формулу для розрахунку близькості між об’єктами і місце для неї на **Листі**.

5. Побудувати матрицю близькості.

7. Вибрати стратегію об’єднання об’єктів у групи

8. Виконати процедуру кластерного аналізу.

9. Побудувати таблицю «об’єднання – вузол – метрика».

10. Побудувати дендрограму.

11. Виконати інтерпретацію результату кластерного аналізу.

12. Оформити звіт.

7. Форма звітності

1. Титулка з назвою роботи.
2. Мета роботи.
3. Короткі теоретичні відомості
4. Форми і методи подання та попереднє статистичне орацювання числових даних часових послідовностей: вхідні дані, хід роботи, результати

4.1. Попереднє опрацювання даних та подання результатів:

4.1.1. Формування файлу даних у формі таблиць.

4.1.2. Графічне подання даних.

4.2. Описова статистика – кількісні характеристики даних.

4.2.1. Побудова гістограми

4.2.2. Побудова кумуляти.

5. Виявлення тенденції часового ряду методами згладжування

5.1. Методи згладжування часових рядів.

5.2. Метод ковзної середньої.

5.3. Метод зваженої ковзної середньої.

5.4. Властивості ковзного середнього.

5.4.1. Лінійне згладжування для $\mathbf{w} = \mathbf{3}$.

5.4.2. Лінійне згладжування для $\mathbf{w} = \mathbf{5}$.

5.4.3. Нелінійне згладжування $\mathbf{w} = \mathbf{7}$.

5.5. Медіанна фільтрація.

5.6. Нормування часових послідовностей.

5.7. Критерії ефективності згладжування часових рядів.

5.8. Формули для зваженого ковзного середнього.

6. Згладжування формулами з Кендела:

- подати узагальнений графік результатів згладжування для однієї реалізації даних, лише один ряд);

- побудувати кореляційну таблицю для всіх інтервалів згладжування, включаючи і ряд оригінальних значень;

- побудувати діаграму поворотних точок для всіх інтервалів згладжування.

6.1. Кореляційний аналіз часових послідовностей.

6.2. Кореляційне поле.

6.3. Коефіцієнт кореляції.

6.4. Кореляційне відношення.

6.5. Властивості кореляційного відношення.

6.6. Кореляційна матриця.

6.7. Автокореляція.

6.8. Автокореляція в часових рядах.

6.9. Розрахунок автокореляції.

7. Згладжування за формулами з Полларда: - аналогічно до п. 6.

8. Експоненціальне згладжування: - аналогічно до п. 6.

9. Медіанне згладжування: - аналогічно з п. 6.

10. Ієрархічний агломеративний кластерний аналіз багатомірних даних: вхідні дані, хід роботи, результати

10.1. Постановка задачі. Виконання цієї роботи полягає у реалізації її двох частин, а саме, побудови матриці близькості на підставі таблиці об’єкт-властивість та проведення самого кластерного аналізу на основі побудованої матриці близькості.

10.2. Побудова матриці близькості.

10.2.1. Формування таблиці «об’єкт- властивість».

10.2.2. Нормування таблиці «об’єкт-властивість».

10.2.3. Вибір метрики для побудови матриці близькості.

10.2.4. Побудова таблиці близькості.

10.3. Проведення агломеративного ієрархічного кластерного аналізу.

10.3.1. Вибір стратегій об’єднання.

10.3.2. Проведення кластерного аналізу.

10.3.3. Побудова дендрограми.

11. Висновок розгорнутий щодо отриманих результатів не менше сторінки (а не одне речення, що ніби чомусь навчилися).