УДК 336.71:303.722.4

Дослідження/

Михайло Капулло
Mykhailo Kapullo

Начальник відділу аналізу ризиків позичальників Групи управління проектами міжнародних кредитних ліній при Національному банку України

Head of the Borrowers' Risk Analysis Division of the Group of International Credit Lines Project Management under the National Bank of Ukraine

Кластеризація часових рядів показників банків

Clusterization of bank's time series indicators

У статті наведено приклад кластерного аналізу динаміки показників діяльності сукупності банків у певному часовому інтервалі, що формують загальносистемний тренд. Розроблена методологія дає змогу виокремлювати характерні групи трендів – кластери трендів, на підставі приналежності банку до яких можна дійти висновку щодо характерних особливостей діяльності банку. У подальшому отримані результати можуть бути використані для побудови системи індикаторних показників для раннього реагування та оцінювання ризиків банку.

The article gives an example of the cluster analysis of the behavior of banks' indicators forming a systematic trend, during a certain time interval. The developed methodology makes it possible to distinguish particular groups of trends – trend clusters. On the ground of bank's belonging to the trend clusters one can draw conclusions concerning characteristic features of bank's activities. The obtained results can be used in construction of the system of early warning indicators as well as in bank's risk assessment.

Ключові слова: часовий ряд, добування даних, кластерний аналіз, оцінка ризиків банку.
Key words: time series, data mining, cluster analysis, bank's risk assessment.

МЕТА СТАТТІ

З метою визначення фінансового стану окремих банків та аналізу функціонування банківської системи в цілому важливим є дослідження динаміки (трендів) заданих показників діяльності банків упродовж певних часових інтервалів. Особливо актуальним це завдання є в періоди високої ринкової нестабільності, коли, з одного боку, можуть виникати загальносистемні тенденції, характерні для більшості банків (наприклад, відплив коштів), а з другого, – протилежні тенденції можуть компенсувати одна одну та спричиняти недостатню інформативність узагальненого тренду. Також важливо те, що загальносистемний тренд є середньозваженим результатом, на який найбільший вплив мають великі системні банки, що, в свою чергу, негативно впливає на репрезентативність загальносистемного тренду.

Таким чином, об'єктом аналізу має бути сукупність трендів та визначення ключових, схожих за своїми характеристиками, тенденцій та їхній вплив на загальносистемну тенденцію. Метою дослідження було провести структурний порівняльний аналіз великої сукупності часових рядів динаміки для:

- визначення наявності системних трендів та ступеня їхньої "загальності" й репрезентативності загальносистемного тренду;
- позиціонування окремого банку щодо інших банків;
- виявлення груп банків (кластерів банків), подібних за часовою динамікою заданих показників, зокрема, близьких за такою динамікою до проблемних банків.

ОБ'ЄКТ ДОСЛІДЖЕННЯ

У статті досліджувалися часові послідовності показників зобов'язань банків: щоденні залишки строкових коштів та коштів на вимогу, залучених від фізичних та юридичних осіб 46 найбільших банків, які становили 90% активів банківської системи України станом на 01.10.2014 р. (за період із 01.10.2014 р. до 01.12.2014 р.).

Для мінімізації впливу фактора обмінного курсу гривні щодо іноземних валют аналіз здійснено окремо за ресурсами, залученими банками у гривнях та в доларах США у номіналах валют.

Отже, загальна кількість досліджених показників, що характеризувалися типом осіб (юридична, фізична), строковістю (строкові, на вимогу) та валютою (гривня, долар США), становила 8.

Для порівняння часової поведінки абсолютних обсягів різних за розміром банків досліджувані показники було приведено у відносний вираз за описаною нижче методикою.

МАТЕМАТИЧНИЙ ОПИС

Позначимо z_n^t значення заданого показника Z на певний момент часу $t = 1,\ldots,T$ для заданого банку $n = 1,\ldots,N$, де T є загальною кількістю моментів часу досліджуваного періоду;

N – загальною кількістю банків.

У разі абсолютної природи показника Z, аби мати можливість порівняння різних банків, сфор-

муємо відносний показник виду:

$$\bar{z}_i^t = \frac{z_i^t}{\bar{z}_i}, \qquad (1)$$

де $\bar{z}_i = \frac{\sum_{t=1}^{T} z_i^t}{T}$ – усереднене за часом значення показника i-го банку.

За змістом показник $\bar{z}_i^t$ відображає відносний рівень показника в кожний окремий момент часу щодо середнього рівня за заданий період та змінюється навколо 1. Для лаконічності у подальшому не використовуватимемо верхню риску.

Визначимо міру відстані між часовими послідовностями значень досліджуваного показника для банків i та j, або інакше, міру несхожості двох часових послідовностей як:

$$d(i,j) = \frac{1}{T}\sum_{t=1}^{T}\left(z_i^t - z_j^t\right)^2. \qquad (2)$$

Ця відстань пропорційна квадрату площі між кривими трендів (див. графік 1) та дорівнюватиме нулю лише в разі рівності трендів. Відстань буде тим більшою, чим більші відмінності трендів між собою.

Зауважимо, що запропонована метрика відстані жодним чином не використовує часової природи досліджуваних показників. Точнішим, але водночас і складнішим підходом для порівняння часових послідовностей є, наприклад, порівняння їхніх спектральних характеристик, що додатково дало б змогу ідентифікувати наявність циклічності показників.

Зобразити взаємне розташування сукупності банків з урахуванням визначених формулою (2) відстаней можемо у вигляді ненаправленого графа [1], [2]. Його вершина відповідатиме окремому банку, а ребра – відстані між банками (див. схему). Такий граф є повним – для кожного банку визначено відстань до кожного іншого банку. Тому для наочності варто залишити лише певну кількість найближчих банків (із найбільш схожими трендами). Оскільки для кожного банку визначено найближчі банки, такий граф точно відображатиме всі банки.

Графік 1. **Визначення відстані між часовими послідовностями показника різних банків як площі між кривими (позначено жовтим кольором)**

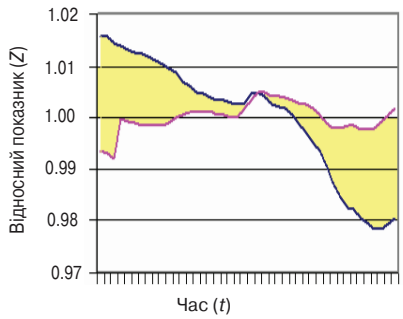
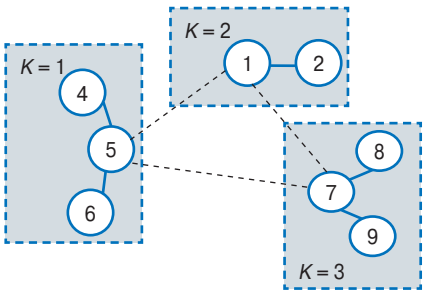


Схема. **Граф відстаней між банками (вершини графа 1–9 відповідають окремим банкам, ребра – відстаням між банками)**



Задача кластеризації банків у сенсі аналізу часової динаміки показника Z полягає в пошуку близьких у сенсі (2) трендів банків та визначення характерних однорідних груп банків – кластерів банків. При цьому відкритим залишається питання кількості кластерів, на які слід розподіляти досліджувані об'єкти. Кількість таких кластерів не може бути великою. Водночас кластери повинні у повному обсязі відображати характерну кількість відокремлених груп близьких об'єктів.

Математичне визначення оптимальної кількості кластерів має таку послідовність. Визначимо внутрішньокластерну відстань k-го кластера як суму всіх відстаней у межах цього кластера:

$$w_k = \sum_{i \subset k} d(i, r_k), \qquad (3)$$

де r_k – центр k-го кластера. На схемі три кластери $k = \overline{1,2,3}$ позначено прямокутниками. Їхні внутрішньокластерні відстані відповідають сумарній довжині ребер усередині цих кластерів.

Загальна внутрішньокластерна відстань є сумою відстаней, визначених у (3):

$$w(K) = \sum_{k=1}^{K} w_k, \qquad (4)$$

де K – загальна кількість кластерів.

Зі збільшенням K ця функція спадає, оскільки відстань до центрів кластерів зі збільшенням їхньої кількості зменшується. У граничному випадку, коли кількість кластерів дорівнює кількості банків (тобто кожен кластер складається з одного банку): $w(K = N) = 0$.

Визначимо міжкластерну відстань як суму відстаней між центрами кластерів:

$$b(K) = \sum_{k=1}^{K} d(r_k, r_0), \qquad (5)$$

де r_0 – центр системи. На схемі міжкластерна відстань відповідає сумі довжин ребер позначених штрих-пунктиром. Ця функція зростає зі збільшенням кількості кластерів та у граничному випадку, коли кількість кластерів дорівнює кількості банків $b(K = N) = w(K = 1)$.

Визначення оптимальної кількості кластерів K полягає в мінімізації загальної внутрішньокластерної та максимізації міжкластерної відстані, що досягається у граничному випадку. Отже, оптимальна кількість кластерів має бути обмежена з інших міркувань, наприклад, тією кількістю, що забезпечує найбільше зменшення загальної внутрішньокластерної відстані за умови однакового зростання міжкластерної відстані. В ролі такого критерію може бути використано мінімум функції:

$$\left|w(K) - b(K)\right|, \qquad (6)$$

який досягається при рівності $w(K)$, $b(K)$.

Остаточно питання щодо кількості необхідних кластерів треба вирішувати експертним шляхом, враховуючи при цьому специфіку конкретного застосування. Так, у випадку аналізу часової поведінки показників поділ на три категорії є інтуїтивним: висхідного, сталого (із певним визначенням сталості) та низхідного трендів. Перевірку правильності такої категоризації, визначення необхідності й можливості подальшого "подрібнення" категорій аналітик може здійснити зокрема за допомогою пропонованої методики.

Побічним результатом її застосування є загальна метрика несхожості всіх трендів банків між собою, а саме загальна внутрішньокластерна відстань $w(K = 1)$. Ця відстань для різних показників Z є виміром нестабільності показника протягом досліджуваного періоду. Великий вплив на величину $w(K = 1)$ мають тренди зі значними відхиленнями від загальносистемного тренду. Знівелювати цей вплив можна шляхом порівняння міжкластерної відстані для певної кількості сформованих кластерів, наприклад, $w(K = 5)$. У такому випадку для дуже віддалених трендів утворюватимуться окремі кластери, а отже, їхній вплив на загальну внутрішньокластерну відстань буде значно меншим. Велика різниця між $w(K = 1)$ та $w(K = 5)$ свідчить про те, що деякі банки мають значні відмінності трендів показника.

МЕТОДИКА ДОСЛІДЖЕННЯ

Щоденні балансові дані досліджуваних банків для уникнення

випадкових флуктуацій даних перед здійсненням подальшого аналізу були згладжені методом ковзного середнього. Розмір вікна обирався з міркувань найменшого викривлення даних, що не змінювало тижневі тренди та завдяки симетричності не викривляло дані.

Для згладжених часових послідовностей було побудовано сукупності відстаней між банками відповідно до (2).

Кластерізація здійснювалася за методом "найближчих сусідів" (K-means) [3] зі збільшенням кількості кластерів K з 1 до 10. Сутність цього методу кластеризації полягає у випадковому виборі початкових центрів кластерів та послідовному повторенні двох кроків:

– визначення приналежності банку до найближчого кластера;

– уточнення центрів кластерів як банків, що забезпечують найменшу внутрішньокластерну відстань.

Процедура припиняється після досягнення стійкого стану – незмінності центрів кластерів. Після цього остаточно вимірювалися загальна внутрішньокластерна та міжкластерна відстані, визначені відповідно в (4) та (5).

Для уникнення можливого потрапляння в локальні мінімуми загальної внутрішньокластерної відстані описана в попередньому параграфі процедура повторювалася багато разів, а в ролі остаточного результату кластерізації обиралася конфігурація з найменшою загальною внутрішньокластерною відстанню.

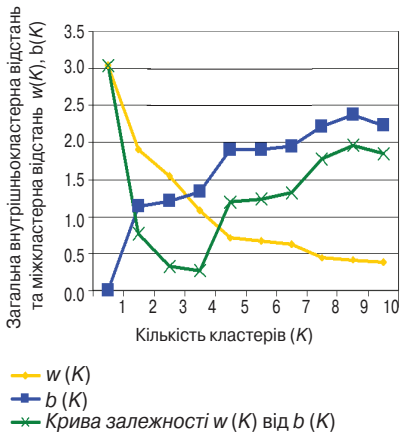
З метою перевірки правильності роботи алгоритмів було здійснено візуалізацію даних. Поряд із графіками часових послідовностей показників побудовано графи відстаней між трендами банків, аналогічні зображеному на схемі. При цьому здійснено два типи кольорового маркування вершин графів:

- різними кольорами позначено банки з різних кластерів;
- поруч позначено категоризацію трендів на три категорії: висхідного, сталого та низхідного трендів. При цьому під сталим трендом маємо на увазі зміну показника в межах 5%, а у висхідних та низхідних трендах було додано по дві підкатегорії: зміни на 5–10% та зміни на понад 10%.

Остаточна перевірка роботи алгоритму здійснювалася шляхом маркування часових послідовностей кольором кластера, до якого ці тренди належать.

Для автоматичної побудови графів використовувався продукт Графвіз (Graphviz) [4].

Графік 2. **Залежність сумарної внутрішньокластерної відстані $w(K)$ та міжкластерної відстані $b(K)$ від кількості кластерів**



ОСНОВНІ РЕЗУЛЬТАТИ

Розглянемо результати аналізу на прикладі показника залишків у банках гривневих строкових коштів юридичних осіб. Часову динаміку значених залишків за досліджуваний період різних банків наведено на графіку 3-а. Граф відстаней, аналогічний зображеному на схемі, на якому вершини відповідають окремим банкам, довжина ребер – відстаням між трендами банків, а колір маркування – зміні залишків за досліджуваний період, зображено на графіку 3-в. Цей граф свідчить про те, що загальносистемний сталий тренд (зображений великим сірим колом) став результатом двох протилежних тенденцій, що були притаманні банкам. При цьому домінували банки зі зростанням залишків (зображені червоними колами).

Для визначення оптимальної кількості кластерів було побудовано залежності загальної внутрішньокластерної відстані (4) та міжкластерної відстані (5) від кількості кластерів K (див. графік 2). Цілком очевидно, що характеристики виявлених кластерів відповідають теоретично очікуваним: спадною є загальна внутрішньокластерна відстань $w(K)$, натомість міжкластерна відстань $b(K)$ – зростає. Також заслуговує на увагу те, що швидкість спадання загальної внутрішньокластерної відстані зменшується. Оптимальна кількість кластерів відповідно до мінімуму функції (6) становить $K = 4$. Водночас із міркувань дотримання симетрії та ізоляції разюче відмінних від загальносистемного трендів кількість кластерів було збільшено до $K = 5$.

Результат кластеризації з кількістю кластерів $K = 5$ відображено на графіку 3-б, з якого видно, що існують такі кластери:

- два ізольовані кластери, кожен містить по одному банку, які мають значні відхилення від загальносистемного тренду (блакитне коло вгорі та бузкове внизу);
- група банків зі значними низхідними трендами (позначені зеленим знизу);
- група банків із висхідною динамікою залишків (позначені жовтим);
- найбільша група банків, динаміка залишків яких близька до середньої системної (позначені сірими колами).

Остаточна перевірка кластеризації здійснюється шляхом позначення окремих груп трендів на графіках часової динаміки кольорами відповідних кластерів (див. графік 3-а).

Аналогічні результати отримано для іншого показника – залишків строкових гривневих коштів фізичних осіб, їх наведено на графіку 4. Порівняння з попереднім результатом свідчить про те, що для фізичних осіб низхідні тренди є значно характернішими. А загалом тренди є вельми схо-

Порівняння параметрів кластеризації різних показників залишків коштів у банках України

Тип особи	Строковість	Валюта	Внутрішньокластерна відстань $w(K=1)$	Внутрішньокластерна відстань $w(K=5)$	Відносна змінна δ(%)	Оптимальна кількість банків у кластері K_{opt}
Юридичні особи	На вимогу	Долар США	3.39	1.72	49	6
	Строкові	Долар США	3.78	1.18	69	4
	Строкові	**Гривня**	**3.04**	**0.71**	**77**	**4**
	На вимогу	Гривня	0.56	0.29	47	6
Фізичні особи	На вимогу	Гривня	0.65	0.26	60	6
		Долар США	0.33	0.11	68	8
	Строкові	**Гривня**	**0.17**	**0.01**	**93**	**2**
		Долар США	0.03	0.01	65	3

Джерело: власні розрахунки автора.

жими між собою. Іншими словами, дії фізичних осіб з депозитними вкладами в різних банках значно однорідніші, ніж дії юридичних осіб.

Порівняння між собою властивостей кластеризації інших, досліджених у роботі показників, наведено в таблиці.

У таблиці наведено загальну внутрішньокластерну відстань $w(K = 1)$, $w(K = 5)$, її відносний спад $\partial = \frac{w(K=1) - w(K=5)}{w(K=1)}$ та оптимальну кількість кластерів K_{opt}, визначені відповідно до мінімуму функції (6). Таблиця відсортована у зворотному порядку за відстанню $w(K = 5)$, яка відображає ступінь несхожості між собою трендів показників різних банків – нестабільність показників. Напівжирним виділено рядки показників, результати аналізу яких наводилися вище.

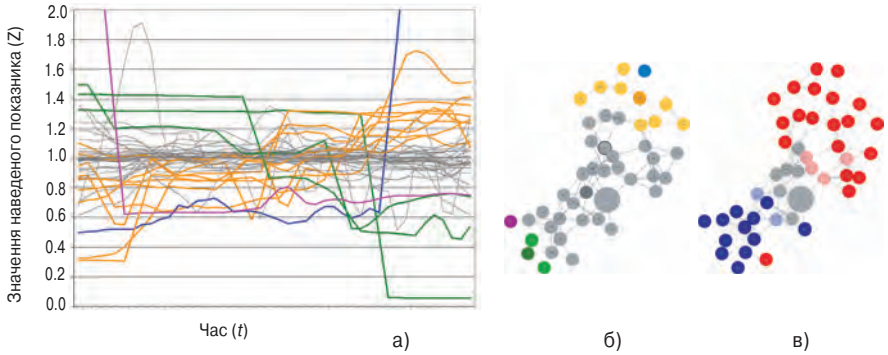
За даними таблиці бачимо, що найнестабільнішими впродовж досліджуваного періоду були залишки коштів юридичних осіб, адже їх динаміка для різних банків є найбільш різноманітною. Водночас більшу нестабільність мали кошти в доларах США. Динаміка строкових коштів, залучених від юридичних осіб, суттєво не відрізняється від динаміки відповідних коштів на вимогу.

Для фізичних осіб, на відміну від юридичних осіб, динаміка строкових коштів є вельми однорідною, серед яких найбільш однорідною була динаміка залишків коштів у доларах США. Цей факт можемо пояснити адміністративними заходами Національного банку, що діяли в цей період.

Відносна змінна ∂ відображає те, як швидко зменшується загальна внутрішньокластерна відстань при збільшенні кількості кластерів та свідчить про схильність до кластеризації трендів різних показників або про наявність помітно відмінних від загалу трендів. Як результат для великих значень ∂ відповідними є малі значення оптимальної кількості кластерів K_{opt}. Зокрема, за даними таблиці бачимо, що найбільшу схильність до групування мають залишки гривневих строкових коштів.

Додаткового дослідження потребує обґрунтування вибору тривалості досліджуваного періоду (Т). Спостереження за зміною параметрів кластеризації при зміні досліджуваного періоду в майбутньому допоможе виявити зміну динаміки показників – появу нових чи зникнення існуючих кластерних груп.

Графік 3. **Кластерний аналіз часової динаміки залишків гривневих строкових коштів *юридичних осіб* 46 найбільших банків України**

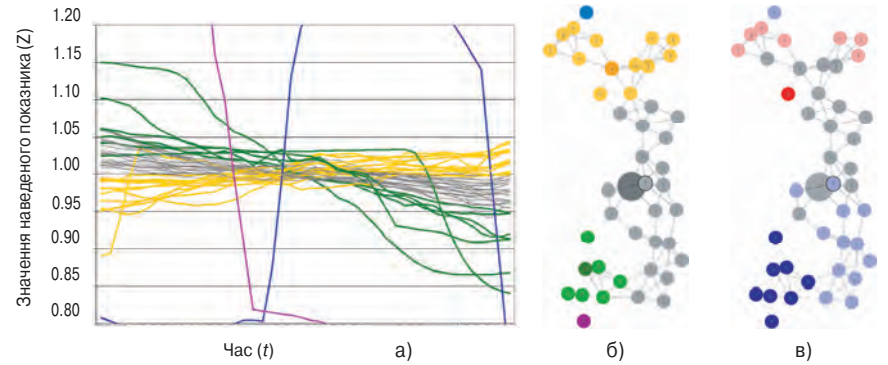


а) часова динаміка залишків коштів. Кожна окрема крива відповідає залишкам кожного окремого банку. Кольорами позначено виявлені кластери близьких трендів

б) кластеризація трендів банків при $K = 5$. Кольорами позначено виявлені кластери схожих трендів

в) позначення змін показника за досліджуваний період: синім кольором позначено зменшення; сірим – сталий тренд; червоним – збільшення. Великим колом зображено загальносистемний тренд

Графік 4. **Кластерний аналіз часової динаміки залишків гривневих строкових коштів *фізичних осіб* 46 найбільших банків України**



а) часова динаміка залишків коштів. Кожна окрема крива відповідає залишкам кожного окремого банку. Кольорами позначено виявлені кластери близьких трендів

б) кластеризація трендів банків при $K = 5$. Кольорами позначено виявлені кластери схожих трендів

в) позначення змін показника за досліджуваний період: синім кольором позначено зменшення; сірим – сталий тренд; червоним – збільшення. Великим колом зображено загальносистемний тренд

ВИСНОВОК

Запропонована методика застосування кластерного аналізу для групування часової динаміки показників діяльності банків дає змогу:

– аналізувати спосіб утворення загальносистемної тенденції;

– виявляти характерні кластерні групи за змінами показників та позиціонувати банки в ці групи.

Зокрема, особливої уваги потребують банки зі значними відхиленнями від загальносистемної тенденції динаміки показників. Застосування запропонованої методики в практиці банківського нагляду в поєднанні з традиційними інтерполяційними методами аналізу часових рядів доповнить показники раннього реагування та забезпечить якісне посилення ефективності нагляду. ❑

Список використаних джерел

1. Харари Ф. Теория графов. – М.: Мир, 1973. – 311 с. – (Russian source).

2. Кристофидес Н. Теория графов. Алгоритмический подход. – М.: Мир, 1978. – 429 с. – (Russian source).

3. David J. Hand, ISBN-13: 978-0262082907, Principles of Data Mining (Adaptive Computation and Machine Learning), A Bradford Book, 2001. – 584 p.

4. Graphviz – Graph Visualization Software. – [Електронний ресурс]. – Режим доступу: http:// www.graphviz.org.