

**Поперешняк Світлана Володимирівна**

к.ф.-м.н, доцент, доцент кафедри інформатики та програмної інженерії,

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

Державний університет інформаційно-комунікаційних технологій, Київ, Україна

ORCID ID 0000-0002-0531-9809

spopereshnyak@gmail.com

**ІНТЕГРАЦІЯ ТОЧНОГО СТРУКТУРНО-ЙМОВІРІСНОГО ОЦІНЮВАННЯ ТА
СЕКВЕНЦІЙНОГО АГРЕГУВАННЯ КОРОТКИХ ВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ**

Анотація. У статті досліджено інтеграцію точного структурно-ймовірнісного оцінювання коротких випадкових послідовностей із секвенційним агрегуванням локальних оцінок для характеристики розвитку структурних змін у бінарних потоках даних. Обґрунтовано необхідність урахування не лише частоти появи символів, а й їх взаємного розташування та порядку виникнення локальних відхилень. Теоретичну основу підходу становлять точні спільні розподіли статистик входжень дво- і триланцюжків, які враховують залежність і перекривання структурних конфігурацій у коротких фрагментах. На їх основі визначаються оцінки локальної структурної нетиповості відносно базової ймовірнісної моделі. Запропоновано послідовне структурне агрегування оцінок одного фрагмента та секвенційне агрегування їх упорядкованої траєкторії. Сформоване представлення характеризує рівень, тренд, накопичення, поточну тривалість і мінливість структурних відхилень. Проведено відтворюваний обчислювальний експеримент на синтетичному бінарному потоці з контрольованою зміною режимів формування. Показано реакцію локальних оцінок на зміну залежності між сусідніми елементами за незмінної граничної ймовірності символів. Аналіз секвенційних компонентів виявив відмінності між наростанням, стійким збереженням і згасанням відхилень, які неможливо встановити лише за усередненим рівнем локальної нетиповості. Наукове значення результатів полягає в обґрунтуванні взаємодоповнювальності точного опису локальної структури та аналізу її послідовної динаміки. Інтегроване представлення створює основу для виділення потенційних ранніх структурних ознак небажаних змін у діагностичних потоках. Подальші дослідження передбачають перевірку запропонованого підходу на реальних даних, оцінювання чутливості й затримки виявлення за контрольованого рівня хибних спрацювань, а також визначення умов його практичного застосування.

Ключові слова: математичне моделювання; випадкові послідовності; структурний аналіз; спільні розподіли; поточкові дані; секвенційне агрегування; виявлення змін; моніторинг систем.

ВСТУП

Сучасні інформаційно-технічні системи характеризуються безперервним формуванням потоків діагностичних і технологічних даних, аналіз яких має забезпечувати своєчасне виявлення змін у функціонуванні контрольованих процесів. Особливої складності ця задача набуває за обмеженого обсягу доступних спостережень, коли накопичення тривалих статистичних вибірок є недоцільним або несумісним із вимогами оперативного моніторингу. За таких умов важливого значення набуває дослідження коротких фрагментів даних із урахуванням не лише їхніх частотних характеристик, а й внутрішньої структури та закономірностей її послідовної зміни.

Для бінарних потоків інформативними характеристиками локальної структури є кількості входжень заданих дво- і триланцюжків, які відображають взаємне розташування символів, переходи між ними, формування та переривання серій. Послідовності з однаковою кількістю нулів і одиниць можуть істотно відрізнятися за зазначеними характеристиками. Водночас можливість перекривання входжень і наявність структурних обмежень зумовлюють статистичну залежність відповідних показників, що потребує використання їх спільного ймовірнісного опису.



Застосування точних спільних розподілів набуває особливого значення для коротких послідовностей, оскільки дозволяє враховувати скінченну довжину фрагмента, множину допустимих конфігурацій і взаємозв'язки між структурними статистиками без заміни їх асимптотичними наближеннями. Проте оцінювання нетиповості окремого фрагмента не забезпечує повної характеристики розвитку структурних змін у потоці даних. Однакові або близькі локальні оцінки можуть відповідати різним ситуаціям: виникненню одиничного відхилення, його поступовому наростанню, тривалому збереженню або згасанню. Отже, актуальним є формування інтегрованого представлення, яке поєднує точне структурно-ймовірнісне оцінювання з аналізом упорядкованої траєкторії локальних характеристик.

Постановка проблеми

Використання точних спільних розподілів статистик входжень дво- і триланцюжків забезпечує можливість кількісного оцінювання структурної нетиповості коротких бінарних фрагментів відносно заданої базової ймовірнісної моделі. Однак така оцінка характеризує окремий фрагмент і без додаткового аналізу не дозволяє встановити спрямованість, накопичення та стійкість виявлених відхилень.

У разі оброблення потоку ковзним вікном формується послідовність взаємопов'язаних локальних оцінок, порядок виникнення яких містить додаткову інформацію про структурну динаміку досліджуваного процесу. Зведення цієї послідовності до одного усередненого показника не дозволяє розрізнити траєкторії з однаковим загальним рівнем нетиповості, але різними напрямками розвитку, тривалістю перевищень і характером коливань.

Таким чином, наукова проблема полягає в узгодженні точного ймовірнісного опису локальних структурних конфігурацій із багатокomпонентним аналізом динаміки сформованих оцінок. Її розв'язання передбачає встановлення взаємозв'язку між структурною нетиповістю окремих коротких фрагментів та характеристиками її розвитку в послідовності спостережень, що створює передумови для виділення потенційних ранніх структурних ознак зміни режиму функціонування контролюваного процесу.

Аналіз останніх досліджень і публікацій

Математичні основи дослідження локальних конфігурацій випадкових послідовностей сформовано в межах комбінаторно-ймовірнісного аналізу входжень заданих шаблонів. У класичній роботі [1] досліджено перекривання символічних ланцюжків і встановлено співвідношення, що враховують взаємозалежність появи окремих конфігурацій. Подальший розвиток точних методів представлено в сучасному дослідженні [2], присвяченому розподілам кількостей структурних шаблонів обмеженої довжини у скінченних бінарних послідовностях. Безпосередньо для коротких випадкових бітових послідовностей у роботі [3] отримано точні спільні розподіли структурних статистик, які характеризують взаємне розташування елементів послідовності. Зазначені результати обґрунтовують доцільність використання точного ймовірнісного опису локальних конфігурацій з урахуванням їх взаємозалежності та обмеженої довжини аналізованого фрагмента.

Окремий напрям становлять методи послідовного статистичного аналізу, які дозволяють оцінювати зміни контрольованих процесів у міру надходження нових спостережень. Фундаментальний підхід до накопичення інформації про відхилення реалізовано в методі CUSUM [4], тоді як у роботі [5] запропоновано байєсівський метод оперативного виявлення точок зміни, заснований на послідовному оновленні ймовірнісного опису тривалості поточного режиму. Сучасний розвиток цього напрямку відображено в дослідженні [6], у якому розглянуто співвідношення статистичної ефективності та обчислювальних витрат алгоритмів послідовного виявлення змін. У роботі [7] систематизовано підходи до аналізу динаміки змін у часових рядах, а в дослідженні [8] запропоновано метод оперативного виявлення змін у розподілених сенсорних мережах з урахуванням обмежень комунікаційних ресурсів. Зазначені праці підтверджують актуальність послідовного аналізу потокових даних, однак переважно орієнтовані на встановлення моменту зміни режиму або формування відповідного статистичного рішення.

Важливе місце в сучасних дослідженнях посідає виявлення аномалій у часових рядах. В огляді [9] систематизовано підходи до визначення аномальних спостережень з урахуванням особливостей послідовних даних, зокрема їх часової залежності та контексту. У роботі [10] виконано масштабне порівняльне оцінювання методів виявлення аномалій, результати якого підкреслюють необхідність узгоджених умов експериментального зіставлення алгоритмів. Подальший розвиток цього напрямку пов'язаний із методами глибокого навчання, сучасний стан яких узагальнено в дослідженнях [11, 12]. Перспективи застосування самокерованого навчання для виявлення аномалій часових рядів розглянуто в роботі [13]. Водночас для задач, у яких базова ймовірнісна модель короткої випадкової послідовності може бути задана аналітично, окремий інтерес становить формування інтерпретованих локальних оцінок структурної нетиповості без необхідності навчання їх базового розподілу на великих вибірках.

Безпосередні передумови запропонованого підходу пов'язані з результатами досліджень [14, 15]. У роботі [14] запропоновано багатокomпонентну модель оцінювання коротких послідовностей, що враховує рівень, напрям зміни, накопичення, тривалість і мінливість структурної нетиповості. У дослідженні [15]



розглянуто метод секвенційно агрегованого оцінювання структурних змін у потоках даних. У поєднанні з точними спільними розподілами структурних статистик [3] зазначені результати формують математичні та методологічні передумови для побудови інтегрованого представлення, яке поєднує локальний імовірнісний опис бінарних конфігурацій з аналізом динаміки сформованих оцінок.

Проведений аналіз публікацій свідчить про активний розвиток точних методів дослідження структурних характеристик випадкових послідовностей, послідовного статистичного контролю та виявлення аномалій у часових рядах. Водночас окремого дослідження потребує узгоджене використання точних спільних розподілів взаємозалежних локальних статистик для формування структурно-імовірнісних оцінок із подальшим аналізом їх упорядкованої траєкторії. Зокрема, необхідно встановити, яку додаткову інформацію про характер розвитку структурних відхилень забезпечує багатокомпонентне секвенційне представлення порівняно з аналізом ізольованих локальних оцінок або їх неупорядкованого узагальнення.

Мета статті – обґрунтування та експериментальне дослідження інтегрованого підходу до оцінювання коротких випадкових послідовностей на основі точних спільних розподілів локальних структурних статистик і секвенційного агрегування сформованих оцінок для кількісного опису структурних змін з урахуванням характеру їх розвитку в потоці даних.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Точні спільні розподіли статистик локальної структури коротких випадкових послідовностей

Нехай $\Gamma_n = (\gamma_1, \dots, \gamma_n)$ – коротка бінарна випадкова послідовність, елементи якої за базовою моделлю незалежні та однаково розподілені:

$$P_r\{\gamma_i = 1\} = p, \quad P_r\{\gamma_i = 0\} = q = 1 - p, \quad 0 < p < 1. \quad (1)$$

Для фіксованого ланцюжка $w = (w_1, \dots, w_s)$ довжини $s \in \{2, 3\}$ кількість його входжень, включно з можливими перекривними, визначається статистикою

$$\eta_n(w) = \sum_{i=1}^{n-s+1} \mathbf{1}\{(\gamma_i, \dots, \gamma_{i+s-1}) = w\}. \quad (2)$$

На відміну від кількості одиниць або нулів, статистика (2) характеризує порядок розташування символів. Кількості входжень різних ланцюжків загалом залежні: поява однієї конфігурації може створювати або обмежувати можливість появи іншої, а сусідні входження можуть перекриватися. Тому ймовірності спільних конфігурацій слід визначати за спільними, а не за добутком маргінальних розподілів.

Математичну основу дослідження становить одержана раніше система точних дво- і тривимірних спільних розподілів кількостей входжень дво- та триланцюжків. Вона охоплює структурно відмінні конфігурації, пов'язані з однорідними серіями, переходами між символами, продовженням і перериванням серій, чергуванням та локально ізольованими бітами. Симетричні випадки пов'язуються інверсією символів (із відповідною заміною p та q), реверсом ланцюжків, перестановкою координат і маргіналізацією багатовимірних розподілів. Таке групування не отожднює залежні статистики, а усуває необхідність повторного запису еквівалентних співвідношень.

Як репрезентативний випадок розглянемо вектор статистик $\mathbf{K} = (\eta_n(10), \eta_n(110), \eta_n(100))$. Позначимо його можливе значення $\mathbf{k} = (k_1, k_2, k_3)$, а кількості одиниць і нулів у послідовності – відповідно m_1 та $m_0 = n - m_1$. Для цього випадку точний спільний розподіл має вигляд

$$P_r\{\mathbf{K} = (k_1, k_2, k_3)\} = \sum_{m_1=k_1}^{n-k_1} p^{m_1} q^{m_0} \binom{k_1}{k_2} \binom{k_1}{k_3} \times \binom{m_1-k_1}{k_2} \binom{m_0-k_1}{k_3}, \quad m_0 = n - m_1. \quad (3)$$

У формулі (3) біноміальний коефіцієнт дорівнює нулю, якщо його нижній аргумент виходить за межі від нуля до верхнього аргументу; для інших недопустимих значень параметрів імовірність також дорівнює нулю. Співвідношення враховує спільне виникнення переходу 10 та пов'язаних із ним триланцюжків 110 і 100, а не припускає незалежності трьох кількостей. Це конкретизація відповідного аналітичного результату дисертаційного дослідження для $t = 1$. Для інверсної конфігурації (01, 001, 011) застосовується взаємна заміна позначень нуля й одиниць та параметрів p і q .

Для довільної обраної системи структурних статистик j введемо випадковий вектор $\mathbf{K}_j$ та його область допустимих значень Ω_j . За базовою гіпотезою H_0 точкова ймовірність визначається як

$$p_{0j}(\mathbf{k}) = P_r\{\mathbf{K}_j = \mathbf{k}\}, \quad \mathbf{k} \in \Omega_j, \quad \sum_{\mathbf{k} \in \Omega_j} p_{0j}(\mathbf{k}) = 1. \quad (4)$$

Таким чином, точні розподіли задають імовірнісну шкалу, за якою можна зіставляти спостережувані структурні конфігурації коротких фрагментів. Формула (3) демонструє один елемент системи, а не замінює інші її співвідношення.

**Локальна структурно-ймовірнісна оцінка**

Нехай $\mathbf{k}_{rj}$ – спостережуване значення вектора статистик системи j у фрагменті з номером r . Для дискретного розподілу (4) визначимо множину конфігурацій, точкові ймовірності яких не перевищують ймовірності спостережуваної конфігурації:

$$\mathcal{T}_j(\mathbf{k}_{rj}) = \{\mathbf{k} \in \Omega_j : p_{0j}(\mathbf{k}) \leq p_{0j}(\mathbf{k}_{rj})\}. \quad (5)$$

Її сумарна ймовірність та локальна оцінка задаються співвідношеннями

$$q_{rj} = \sum_{\mathbf{k} \in \mathcal{T}_j(\mathbf{k}_{rj})} p_{0j}(\mathbf{k}), \quad e_{rj} = 1 - q_{rj}, \quad 0 \leq e_{rj} \leq 1. \quad (6)$$

У разі однакових точкових ймовірностей усі відповідні конфігурації включаються до множини (5). За прийнятою орієнтацією шкали більші значення e_{rj} відповідають менш типовим щодо H_0 конфігураціям. Водночас e_{rj} не є ані ймовірністю альтернативної гіпотези, ані ймовірністю небезпечного стану. Вибір оцінки через сумарну ймовірність дозволяє враховувати дискретність та неоднакову кратність точкових ймовірностей.

За використання M систем структурних статистик один фрагмент характеризується вектором

$$\mathbf{e}_r = (e_{r1}, e_{r2}, \dots, e_{rM}). \quad (7)$$

Для переходу до узагальненої локальної характеристики вводиться оператор структурного агрегування $\mathcal{G}$. Зокрема, зважена лінійна модель має вигляд

$$u_r = \mathcal{G}(\mathbf{e}_r) = \sum_{j=1}^M \alpha_j e_{rj}, \quad \alpha_j \geq 0, \quad \sum_{j=1}^M \alpha_j = 1. \quad (8)$$

Отже, $u_r \in [0,1]$ узагальнює декілька структурних характеристик **одного** фрагмента. Вибір систем статистик і ваг визначає зміст локального агрегування; сам оператор (8) ще не враховує порядок надходження фрагментів.

Секвенційне представлення структурної динаміки

Нехай послідовні фрагменти сформували впорядковану траєкторію локальних оцінок $U = (u_1, \dots, u_R)$. Для поточного положення $r \geq L$ розглянемо інтервал із L останніх значень:

$$\mathbf{u}_r^{(L)} = (u_{r-L+1}, u_{r-L+2}, \dots, u_r), \quad L \geq 2. \quad (9)$$

Секвенційне агрегування визначається відображенням $\mathcal{A}_L$, результатом якого є p -ятикомпонентний вектор

$$\mathbf{A}_r^{(L)} = \mathcal{A}_L(\mathbf{u}_r^{(L)}) = (A_{r,\text{lev}}^{(L)}, A_{r,\text{tr}}^{(L)}, A_{r,\text{acc}}^{(L)}, A_{r,\text{dur}}^{(L)}, A_{r,\text{var}}^{(L)}). \quad (10)$$

Компоненти характеризують відповідно рівень локальної нетиповості, напрям її зміни, накопичення перевищень контрольного порога, поточну тривалість неперервної серії перевищень та мінливість оцінок. Їх спільне використання зумовлене відмінністю інформаційних функцій: однаковий середній рівень може відповідати протилежним трендам, а близьке накопичення – різній неперервності відхилень. При цьому мінливість характеризує розкид значень, але сама по собі не визначає їх порядок. Конкретні математичні оператори компонентів і параметри їх обчислення встановлюються в методиці дослідження.

Інтегроване структурно-секвенційне представлення

Взаємозв'язок двох рівнів агрегування можна подати послідовністю відображень

$$X^{(r)} \rightarrow (\mathbf{k}_{r1}, \dots, \mathbf{k}_{rM}) \rightarrow \mathbf{e}_r \xrightarrow{\mathcal{G}} u_r \xrightarrow{\mathcal{A}_L} \mathbf{A}_r^{(L)}.$$

Тут перехід до $\mathbf{e}_r$ спирається на точні спільні розподіли обраних систем структурних статистик; оператор $\mathcal{G}$ об'єднує характеристики в межах одного фрагмента, тоді як $\mathcal{A}_L$ враховує впорядковану траєкторію декількох послідовних фрагментів. Таким чином, інтеграція поєднує ймовірнісно обґрунтовану локальну оцінку з характеристиками розвитку структурної нетиповості. Результатом є вектор ознак структурної динаміки, а не готове рішення про зміну режиму або прогноз небезпечного стану.

МЕТОДИКА ДОСЛІДЖЕННЯ

Методику побудовано як наскрізну процедуру перетворення потоку бінарних спостережень на багатокомпонентний опис структурної динаміки. На відміну від перевірки ізольованих локальних статистик, досліджується узгодженість трьох рівнів: точного спільного розподілу структурних статистик короткого фрагмента, структурного агрегування локальних оцінок одного фрагмента та секвенційного агрегування впорядкованої траєкторії. Останній рівень не є прогностичним рішенням: він формує ознаки для подальшого аналізу стану.

Нехай початковий потік має вигляд

$$X = (X_1, \dots, X_N), \quad X_i \in \{0,1\}.$$

Для довжини фрагмента n та кроку зміщення h поточний фрагмент визначається як

$$W_r = (X_{(r-1)h+1}, \dots, X_{(r-1)h+n}), \quad r = 1, \dots, R, \quad R = \left\lfloor \frac{N-n}{h} \right\rfloor + 1.$$



Якщо $h < n$, сусідні фрагменти перекриваються. Ця залежність зберігається під час оцінювання і враховується при калібруванні порогів на цілих траєкторіях.

Точне структурно-ймовірнісне оцінювання

Для кожної обраної системи j дво- і триланцюжків на фрагменті W_r обчислюється вектор кількостей їх входжень $\mathbf{k}_{rj}$. Входження визначаються в усіх допустимих позиціях, включно з перекривними. За базовою моделлю H_0 використовується точний спільний розподіл

$$p_{0j}(\mathbf{k}) = P_r\{\mathbf{K}_j = \mathbf{k}\}, \quad \mathbf{k} \in \Omega_j,$$

де Ω_j – область допустимих спільних значень. Використання саме спільного розподілу необхідне через структурну залежність кількостей входжень: добуток маргінальних імовірностей загалом не відтворює ймовірності їх одночасної реалізації.

Для зіставлення різних конфігурацій застосовується хвостова оцінка за порядком точкових імовірностей:

$$q_{rj} = \sum_{\mathbf{k} \in \Omega_j: p_{0j}(\mathbf{k}) \leq p_{0j}(\mathbf{k}_{rj})} p_{0j}(\mathbf{k}), \quad e_{rj} = 1 - q_{rj}.$$

При рівних точкових імовірностях у суму включаються всі відповідні конфігурації. Оцінка e_{rj} характеризує нетиповість локальної конфігурації щодо H_0 , а не ймовірність небезпечного стану. Точні розподіли отримуються з математичного апарату розділу 2 дисертації; незалежний повний перебір послідовностей використовується як перевірка обчислювальної реалізації. Зіставляються допустимі області, нормування і кожна точкова ймовірність; відхилення контролюється за максимальним модулем різниці між аналітичним та перебірним розподілами.

Структурне агрегування в межах фрагмента

Для поточного фрагмента формується вектор $\mathbf{e}_r = (e_{r1}, \dots, e_{rm})$, де m – кількість обраних систем структурних статистик. У базовій реалізації застосовується передбачена дисертаційною моделлю зважена лінійна операція:

$$u_r = \sum_{j=1}^m \alpha_j e_{rj}, \quad \alpha_j \geq 0, \quad \sum_{j=1}^m \alpha_j = 1.$$

Ваги та склад систем статистик фіксуються до аналізу підсумкових траєкторій. Зважена модель забезпечує інтерпретованість внеску кожної системи, але допускає компенсацію високої оцінки низькими значеннями інших. Тому у порівняльному аналізі зберігається і початковий вектор $\mathbf{e}_r$. Послідовність $U = (u_1, \dots, u_R)$ є результатом структурного, але ще не секвенційного агрегування.

Секвенційне агрегування та склад інтегрованої оцінки

Для поточного положення r розглядається впорядкований інтервал (9) довжини L . Відповідно до базової моделі дослідження формується вектор (10).

Для однозначності відтворення експерименту нижче наведено вибрані базові моделі компонентів; альтернативні оператори, що використовувалися в дослідженні, не змішуються з ними в межах одного порівняння. Позначимо $v_i = u_{r-L+i}$, $i = 1, \dots, L$. Зважений поточний рівень дорівнює

$$A_{r,\text{lev}}^{(L)} = \sum_{i=1}^L \beta_i v_i, \quad \beta_i \geq 0, \quad \sum_{i=1}^L \beta_i = 1.$$

Ваги β_i визначають пріоритет поточних або давніших оцінок. Для наряду використовується коефіцієнт лінійного тренду, обчислений за всіма точками інтервалу, та його нормована форма:

$$b_r = \frac{\sum_{i=1}^L (i - \bar{i})(v_i - \bar{v})}{\sum_{i=1}^L (i - \bar{i})^2}, \quad \bar{i} = \frac{L+1}{2}, \quad \bar{v} = \frac{1}{L} \sum_{i=1}^L v_i; \quad A_{r,\text{tr}}^{(L)} = \frac{b_r}{b_{\max}(L)}.$$

Тут $b_{\max}(L)$ – максимальний за модулем допустимий нахил для значень $v_i \in [0,1]$ за обраного L ; нормування забезпечує зіставність трендових оцінок. Для фіксованого контрольного рівня $\tau \in [0,1]$ нормоване накопичення визначається як

$$A_{r,\text{acc}}^{(L)} = \frac{1}{L(1-\tau)} \sum_{i=1}^L \max(0, v_i - \tau).$$

Для розрізнення ізольованих і неперервних перевищень вводиться індикатор $I_i = \mathbf{1}\{v_i > \tau\}$. Нехай d_r – кількість послідовних одиничних індикаторів наприкінці поточного інтервалу (нуль, якщо $I_L = 0$). Тоді

$$A_{r,\text{dur}}^{(L)} = \frac{d_r}{L}.$$

Нарешті, нормована мінливість визначається за дисперсійною характеристикою:



$$A_{r,var}^{(L)} = 4 \left[\frac{1}{L} \sum_{i=1}^L (v_i - \bar{v})^2 \right].$$

Компоненти мають різне функціональне призначення: рівень відображає загальну величину нетиповості, тренд – напрям, накопичення – сумарний обсяг перевищень, тривалість – неперервність поточного відхилення, мінливість – розкид оцінок. Нормована дисперсія не є самостійною характеристикою порядку; порядок зберігається в інтегрованому векторі насамперед через тренд і поточну тривалість. Для $L < 2$ тренд не визначається, тому секвенційний аналіз проводиться за $L \geq 2$.

Експериментальні режими та наскрізний розрахунок

Базовий режим H_0 задається незалежними бітами з однаковими ймовірностями нуля й одиниці. Для відокремлення структурної зміни від зміни граничної частоти символів використовується симетрична марковська альтернатива з параметром a :

$$P_r(X_{t+1} = X_t) = a, \quad P_r(X_{t+1} \neq X_t) = 1 - a.$$

За рівномірного початкового біта обидва символи зберігають однакові граничні ймовірності; параметр a змінює залежність між сусідніми елементами. Розглядаються також контрольовані сценарії короточасного, поступового та стійкого відхилення. Момент початку, інтенсивність і тривалість зміни задаються до аналізу результатів.

Основна демонстраційна конфігурація передбачає $n = 16$, $h = 1$ і $L = 8$; окремо досліджується чутливість результатів до цих параметрів. Для однієї контрольованої траєкторії здійснюється наскрізний розрахунок: вихідні фрагменти – вектори структурних статистик – точні ймовірності – локальні оцінки $e_{r,j}$ – узагальнені оцінки u_r – п'ять компонентів $A_r^{(L)}$. Табличне подання цих проміжних результатів забезпечує перевірюваність походження кожної точки подальших графіків.

Порівняльна перевірка та межі висновків

Порівнюються частотна характеристика фрагмента, окремі локальні структурні оцінки, невпорядковане узагальнення u_r та повний секвенційний вектор. Насамперед перевіряється інформаційна відмінність представлень: які контрольовані сценарії залишаються нерозрізненими за середнім рівнем, але розрізняються за трендом, накопиченням або тривалістю.

Для оцінювання ефективності виявлення необхідно окремо задати правило перетворення вектора ознак на сигнал. Пороги всіх порівнюваних правил визначаються на незалежних базових траєкторіях, після чого фіксуються та застосовуються до перевірних потоків. Порівняння здійснюється за узгодженим горизонтом спостереження та однаковим обмеженням на ймовірність принаймні одного хибного спрацювання протягом цього горизонту. Перекривні вікна не трактуються як незалежні випробування. Для альтернативних режимів оцінюються ймовірність виявлення та затримка від моменту введення зміни; результати багаторазового моделювання супроводжуються показниками статистичної невизначеності.

Така організація дозволяє відокремити три різні твердження: правильність точного локального розрахунку, додаткову інформативність секвенційного представлення та фактичну ефективність конкретного правила виявлення. Останнє не виводиться автоматично з перших двох і встановлюється лише за результатами відповідного порівняльного експерименту.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Експериментальне дослідження спрямоване на встановлення закономірностей формування інтегрованих структурно-ймовірнісних оцінок коротких випадкових послідовностей та визначення їхньої інформативності для характеристики динаміки структурних змін у потоці даних. Особливу увагу приділено дослідженню взаємозв'язку між локальною структурною нетиповістю окремих фрагментів і характером її розвитку в послідовності спостережень. Для цього проведено обчислювальний експеримент на контрольованому бінарному потоці з визначеними параметрами базового та альтернативних режимів його формування. Отримані результати дозволяють простежити зміну структурно-ймовірнісних характеристик, встановити особливості їх накопичення та дослідити можливість виділення структурно-динамічних ознак, що можуть передувати формуванню стійкого відхилення від базового режиму.

Базові розподіли та перевірка локального обчислення

За базовою моделлю H_0 усі $2^{16} = 65\,536$ бінарних фрагментів довжини $n = 16$ є рівномірними. Повним перебором кожного фрагмента, включно з перекривними входженнями ланцюжків, сформовано частотні таблиці для трьох систем структурних статистик: $S_1 = (\eta(10), \eta(110), \eta(100))$, $S_2 = (\eta(01), \eta(001), \eta(011))$ та $S_3 = (\eta(00), \eta(000), \eta(100))$. Третю систему включено для перевірки того, як змінюється оцінка за залучення конфігурацій повторення однакових символів. Це зафіксована експериментальна конфігурація, а не твердження про оптимальний склад систем.



У результаті перебору отримано відповідно 103, 103 та 86 допустимих спільних конфігурацій (табл. 1). Для кожної системи сума точкових ймовірностей дорівнює одиниці. Ця перевірка підтверджує внутрішню коректність побудованих перебірних таблиць; окрема числова збірка кожної точки з аналітичними формулами дисертації має виконуватися за їх програмною реалізацією і не підміняється результатом перебору.

Таблиця 1

Характеристики точних перебірних розподілів для $n = 16$

Система	Статистики входжень	Допустимих конфігурацій	Сума ймовірностей
S_1	10; 110; 100	103	1
S_2	01; 001; 011	103	1
S_3	00; 000; 100	86	1

На основі кожного розподілу обчислено локальну оцінку e_{Γ} . Для кожної допустимої конфігурації підсумовувалися ймовірності всіх конфігурацій із не більшою точковою ймовірністю; рівноймовірні стани враховано одночасно. Отже, e_{Γ} є перетворенням повного базового розподілу, а не довільним ранжуванням спостережуваних частот.

Наскрізне формування інтегрованої оцінки на контрольованому потоці

Для аналізу переходу від локального оцінювання до секвенційного представлення сформовано один контрольований потік із $N = 192$ бітів. Перші 80 бітів отримано за незалежною моделлю з $P(1) = 0,5$; наступні 64 біти – за симетричною марковською моделлю з ймовірністю збереження символу $a = 0,9$; останні 48 бітів знову сформовано незалежно. Початковий стан кожного сегмента задавався рівноймовірно. Генератор NumPy PCG64 ініціалізовано числом 20261014. Таким чином, параметри генерації та межі режимів задано до обчислення результатів.

За $n = 16$ та $h = 1$ утворено $R=177$ перекривних фрагментів. Фрагменти $r = 1-65$ цілком належать першому режиму, $r = 81-129$ – марковському, $r = 145-177$ – останньому; проміжні фрагменти на межах сегментів містять біти двох режимів і тому не інтерпретуються як однорідні. Для кожного фрагмента обчислено три вектори структурних статистик, оцінки e_1, e_2, e_3 та узагальнену характеристику u за рівними вагами $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$

$$u_{\Gamma} = (e_{\Gamma 1} + e_{\Gamma 2} + e_{\Gamma 3}) / 3.$$

Таблиця 2 містить вибрані контрольні точки наскрізного обчислення. Номер фрагмента однозначно визначає його вихідні 16 бітів у збереженій відтворюваній реалізації; усі значення в таблиці безпосередньо обчислено за ними. Таблиця призначена для перевірки походження графічних траєкторій, а не для вибіркового доведення ефективності методу.

Таблиця 2

Наскрісні результати локального структурно-ймовірнісного оцінювання (вибрані фрагменти)

r	$K_1 = (k_1, k_2, k_3)$	e_1	e_2	e_3	u_{Γ}
65	3, 2, 1	0,2249	0,3476	0,6230	0,3985
66	3, 2, 1	0,2249	0,3476	0,3268	0,2998
67	3, 2, 1	0,2249	0,2249	0,3268	0,2589
73	2, 1, 1	0,7151	0,2249	0,3268	0,4223
77	1, 1, 0	0,9772	0,9269	0,4248	0,7763
81	2, 2, 1	0,5467	0,5467	0,0000	0,3645
89	1, 1, 1	0,8540	0,8540	0,0000	0,5693
97	0, 0, 0	0,9979	0,9979	0,4248	0,8069
113	1, 1, 1	0,8540	0,9772	0,3268	0,7193
129	0, 0, 0	0,9979	0,9979	0,4248	0,8069
130	1, 1, 0	0,9772	0,9979	0,4248	0,8000
137	2, 1, 1	0,7151	0,7151	0,5362	0,6555
145	4, 2, 2	0,0000	0,3476	0,6895	0,3457
161	4, 3, 2	0,3476	0,0692	0,0954	0,1708
177	4, 2, 2	0,0000	0,3476	0,6678	0,3385

На рисунку 1 наведено всі 177 значень трьох локальних оцінок та їх структурного агрегату. Вертикальними пунктирними лініями позначено останні фрагменти, які повністю належать відповідно першому та другому режимам ($r = 65$ і $r = 129$); після цих меж виникають перехідні фрагменти, що

містять біти двох режимів. Значення u_r не слід інтерпретувати як імовірність марковської альтернативи або ризику: це оцінка нетиповості відносно обраної базової моделі.

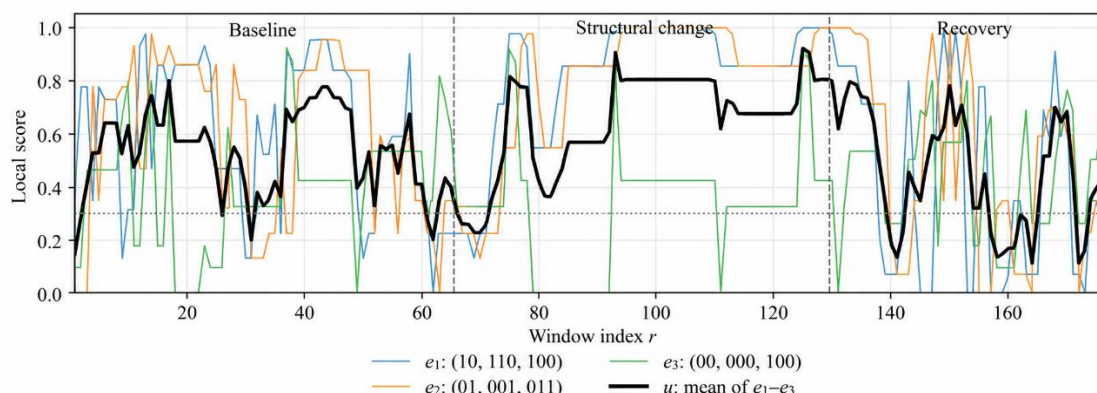


Рис. 1. Траєкторії трьох локальних структурних оцінок та їх узагальнення на контрольованому потоці

Середні значення узагальненої локальної структурно-ймовірнісної оцінки u_r для однорідних частин досліджуваного потоку становлять відповідно 0,5304 ($r = 1-65$), 0,7068 ($r = 81-129$) та 0,4116 ($r = 145-177$). Однак найбільший інтерес становлять не лише відмінності між середніми рівнями, а й характер зміни оцінок у послідовних фрагментах спостереження. Як видно з рис. 1, структурна нетиповість проявляється у вигляді локальних коливань, окремих підвищень та зміни інтенсивності їх виникнення. За таких умов поодинокі підвищення оцінки ще не є достатньою підставою для висновку про зміну стану досліджуваного процесу. Водночас послідовне виникнення підвищених значень, їх накопичення або формування спрямованої тенденції можуть розглядатися як потенційні ранні структурні ознаки розвитку відхилення, що передують його стійкому прояву.

Саме ця особливість визначає діагностичне значення локальних структурно-ймовірнісних оцінок: вони дозволяють виявляти зміни у взаємному розташуванні коротких бітових конфігурацій навіть за відсутності вираженої зміни загальної частоти появи окремих символів. Проте інтерпретація таких коливань як можливих передвісників розвитку небажаного стану потребує врахування їх часової організації. Тому подальше дослідження спрямоване на секвенційне агрегування отриманих оцінок, яке дозволяє перейти від фіксації окремих локальних відхилень до аналізу їхнього рівня, напряму зміни, накопичення, тривалості та мінливості.

Динаміка п'ятикомпонентного секвенційного представлення

Для кожного положення $r \geq 8$ використано $L = 8$ останніх значень u , рівні часові ваги $\beta_i = 1/8$ і контрольний рівень $\tau = 0,3$. Нормувальний коефіцієнт тренду для $L = 8$ дорівнює $b_{\text{ax}} = 4/21 \approx 0,190476$; значення дисперсійної компоненти множиться на 4, відповідно до прийнятої моделі.

Для відтворюваності в таблиці 3 наведено значення повного вектора на характерних положеннях. На відміну від таблиці 2, кожен її рядок характеризує вже не окремий фрагмент, а впорядковану вісімку локальних оцінок, яка закінчується в поточному положенні.

Таблиця 3

П'ятикомпонентна секвенційно агрегована оцінка на вибраних положеннях

r	Рівень	Тренд	Накопичення	Тривалість	Мінливість
65	0,3946	-0,1307	0,1565	0,3750	0,0671
73	0,2892	0,0848	0,0323	0,2500	0,0164
77	0,5234	0,5008	0,3394	0,7500	0,2097
81	0,6236	-0,2500	0,4623	1,0000	0,1223
89	0,5123	0,1553	0,3033	1,0000	0,0243
97	0,7347	0,2082	0,6210	1,0000	0,0600
113	0,7613	-0,0995	0,6590	1,0000	0,0168
129	0,7902	0,1106	0,7004	1,0000	0,0313
137	0,7328	-0,0235	0,6183	1,0000	0,0156
145	0,3235	0,0086	0,0970	0,3750	0,0589
161	0,2420	-0,1732	0,0334	0,0000	0,0425
169	0,4208	0,3759	0,2119	0,6250	0,1471
177	0,3655	-0,0940	0,1520	0,5000	0,1101



На рисунку 2 наведено повні траєкторії всіх п'яти компонентів. Спільна вісь абсцис дозволяє зіставляти їх реакції на ті самі зміни потоку. Рівень відображає поточне усереднене положення оцінок; тренд має знак і реагує на їх спрямовану зміну; накопичення враховує суму перевищень; тривалість фіксує саме неперервну серію наприкінці інтервалу; мінливість характеризує розкид незалежно від порядку.

Аналіз динаміки компонентів секвенційно агрегованої оцінки (рис. 2) свідчить про те, що зміна структурних властивостей потоку характеризується не лише підвищенням або зниженням локальної нетиповості, а й особливостями її розвитку в часі. Відмінності між траєкторіями компонентів відображають різні аспекти цього процесу: рівень характеризує поточну вираженість структурних відхилень, тренд – спрямованість їх зміни, накопичення – сумарну величину перевищень установленого порога, тривалість – неперервність поточного відхилення, а мінливість – ступінь коливання локальних оцінок у межах інтервалу спостереження.

Найбільш показовим є те, що зазначені характеристики змінюються не синхронно. Зростання локальної структурної нетиповості може супроводжуватися додатним трендом ще до формування тривалої серії перевищень, тоді як накопичення та тривалість відображають подальше закріплення відхилення. Водночас зниження поточного рівня не обов'язково означає негайне повернення процесу до базового режиму, оскільки накопичувальна компонента зберігає інформацію про попередні перевищення. Мінливість, зі свого боку, дозволяє відрізнити відносно стійке підвищення оцінок від їх інтенсивного чергування, яке може залишатися непомітним за аналізу лише усередненого рівня.

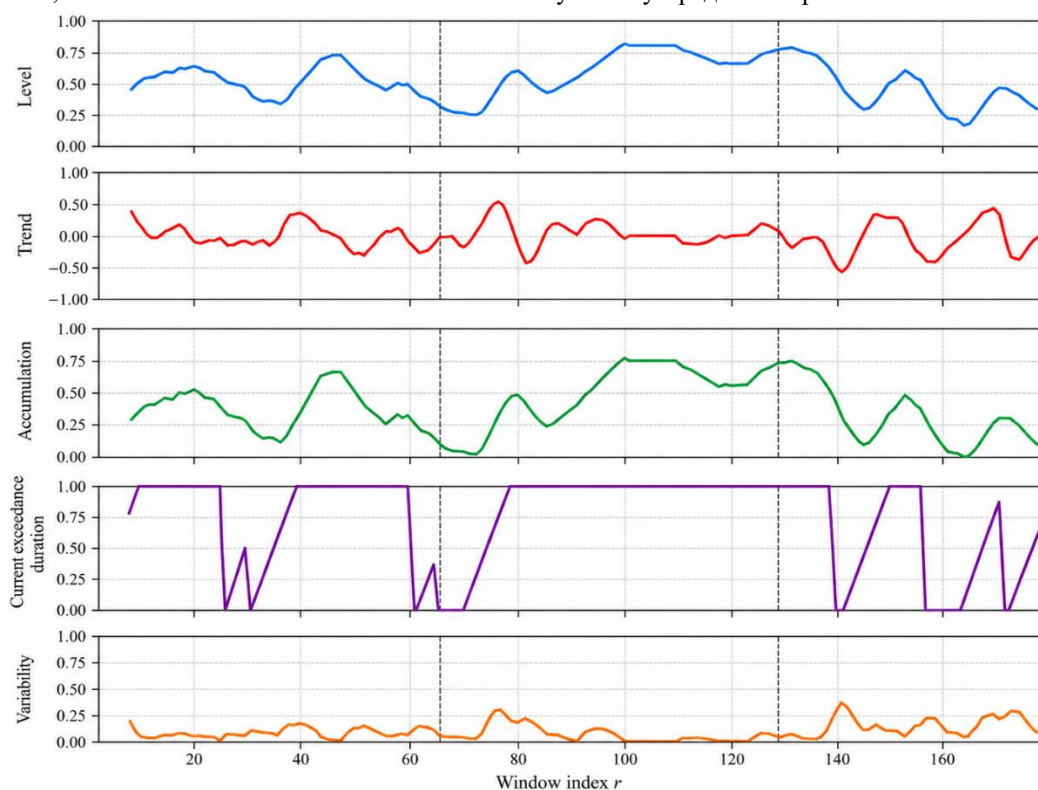


Рис. 2. Динаміка п'яти компонентів секвенційної оцінки ($L = 8$, $\tau = 0,3$)

Зокрема, у положенні $r = 97$, що відповідає марковському сегменту, підвищений рівень структурної нетиповості 0,7347 поєднується з додатним трендом 0,2082, значним накопиченням 0,6210 та максимальною нормованою тривалістю поточної серії перевищень 1,0000. Таке поєднання характеризує не одиничний локальний сплеск, а послідовність відхилень із вираженою тенденцією до зростання та їх неперервним збереженням протягом аналізованого інтервалу. Натомість у положенні $r = 161$, після повернення до незалежного режиму, спостерігаються нижчий рівень 0,2420, від'ємний тренд $-0,1732$, незначне накопичення 0,0334 та відсутність поточної серії перевищень. Сукупність цих характеристик відображає згасання структурного відхилення, яке неможливо настільки повно описати одним поточним значенням локальної оцінки.

Отже, необхідність п'ятикомпонентного представлення зумовлена різним інформаційним змістом його складових. Однаковий рівень структурної нетиповості може відповідати як початковому наростанню



відхилення, так і його стійкому збереженню або поступовому зменшенню. Аналогічно, близькі значення накопичення можуть формуватися внаслідок короточасних інтенсивних коливань або тривалої неперервної серії помірних перевищень. Використання лише однієї інтегральної характеристики призводить до втрати цих відмінностей. Спільний аналіз рівня, тренду, накопичення, тривалості та мінливості забезпечує багатокomпонентний опис структурної динаміки, необхідний для розрізнення короточасних флуктуацій, спрямованих змін і стійких відхилень, які можуть розглядатися як потенційні структурні передвісники розвитку небажаного стану.

Інформаційна відмінність впорядкованого та неупорядкованого узагальнення

Щоб відокремити інформацію про порядок від інформації про набір значень, розглянуто контрольний приклад: вісім фактично отриманих локальних оцінок u з положень $r = 30-37$ та ті самі значення у зворотному порядку. Оскільки набори чисел збігаються, їх середні, накопичення перевищень і дисперсійні характеристики мають бути тотожними; натомість тренд змінює знак, а поточна тривалість перевищення може змінитися (Табл.4).

Таблиця 4

Залежність компонентів від порядку однакового набору локальних оцінок

Траєкторія	Рівень	Тренд	Накопичення	Тривалість	Мінливість
Вихідний	0,3923	0,1914	0,1502	0,7500	0,0672
Зворотний	0,3923	-0,1914	0,1502	0,1250	0,0672

Таблиця 4 є конструктивною перевіркою інформаційної нееквівалентності двох представлень: скалярне неупорядковане усереднення не розрізняє ці траєкторії, тоді як повний вектор зберігає відмінність через компоненти, залежні від порядку. Водночас не всі п'ять компонентів претендують на чутливість до перестановок: рівень за рівних ваг, накопичення та дисперсійна мінливість закономірно залишаються незмінними.

Отриманий результат підтверджує, що наукове значення інтегрованого представлення визначається не лише поєднанням локальних структурних оцінок, а й урахуванням порядку їх виникнення, завдяки чому розрізняються траєкторії, нерозрізненні за неупорядкованими узагальненими характеристиками.

Узагальнення та межі встановлених результатів

Проведене дослідження встановлює відтворюваність повного обчислювального ланцюга та демонструє додаткові інформаційні властивості секвенційного представлення порівняно з окремою або усередненою локальною оцінкою. Результат стосується обраної базової моделі, трьох зафіксованих систем статистик, рівних ваг і конкретних параметрів $n = 16$, $h = 1$, $L = 8$, $\tau = 0,3$. Наведені траєкторії не дозволяють робити висновок про універсальну перевагу методу або оцінювати ймовірність раннього попередження про небезпечний стан.

Для встановлення ефективності конкретного правила виявлення необхідний наступний окремий експеримент: заздалегідь визначити спосіб перетворення п'ятикомпонентного вектора на сигнал, калібрувати його на незалежних цілих базових траєкторіях, порівняти з локальними та неупорядкованими альтернативами за однакової ймовірності хоча б одного хибного спрацювання протягом узгодженого горизонту та оцінити ймовірність і затримку виявлення з показниками статистичної невизначеності. Ця задача не підміняється візуальною відмінністю графіків і виходить за межі наведеного наскрізного демонстраційного експерименту.

ОБГОВОРЕННЯ РЕЗУЛЬТАТІВ

Отримані результати обґрунтовують доцільність інтеграції точного структурно-ймовірнісного оцінювання коротких випадкових послідовностей із секвенційним агрегуванням локальних оцінок. Використання точних спільних розподілів статистик дво- і триланцюжків дозволяє враховувати взаємозв'язки між локальними структурними характеристиками, які не відображаються окремими частотними показниками. Такий підхід узгоджується із сучасними дослідженнями точних розподілів структурних шаблонів у скінченних бінарних послідовностях [2]. Водночас перехід від ізольованих оцінок послідовних фрагментів до їх багатокomпонентного представлення забезпечує врахування не лише величини структурної нетиповості, а й характеру її розвитку в потоці спостережень.

Результати обчислювального експерименту показали, що зміна залежностей між сусідніми елементами бінарного потоку за незмінної граничної ймовірності появи окремих символів супроводжується зміною локальних структурно-ймовірнісних оцінок. При цьому окремі підвищення та коливання таких оцінок не забезпечують однозначної характеристики поточного режиму, оскільки подібні локальні значення можуть виникати за різних умов формування послідовності. Важливого значення



набуває їх часова організація: спрямованість змін, накопичення перевищень, формування неперервних серій та характер коливань. Необхідність урахування часових залежностей і контексту аномальних спостережень також відображено в сучасних дослідженнях аналізу часових рядів [9, 12]. У межах запропонованого підходу зазначені властивості розглядаються як потенційні структурно-динамічні ознаки розвитку відхилень, діагностична значущість яких потребує подальшої кількісної перевірки.

Дослідження п'ятикомпонентного секвенційного представлення підтвердило функціональну відмінність його складових. Компонента рівня характеризує поточну вираженість структурної нетиповості, тренд відображає напрям її зміни, накопичення враховує сумарну величину перевищень, тривалість характеризує неперервність поточного відхилення, а мінливість – розкид локальних оцінок. Несинхронність зміни цих характеристик свідчить про обмеженість опису структурної динаміки лише однією усередненою оцінкою. Зокрема, близькі значення поточного рівня можуть відповідати як наростанню відхилення, так і його стійкому збереженню або поступовому згасанню. Конструктивне порівняння траєкторій, утворених однаковими локальними значеннями в різному порядку, додатково підтвердило інформаційну відмінність секвенційного представлення: за незмінних рівня, накопичення та мінливості змінюються тренд і поточна тривалість перевищень. Таким чином, основний результат інтеграції полягає у формуванні взаємодоповнювального структурно-динамічного представлення, здатного розрізняти сценарії, які залишаються нерозрізненими за невпорядкованими узагальненими характеристиками.

У контексті відомих підходів до аналізу часових послідовностей запропоноване представлення має інше функціональне призначення, ніж класичні процедури виявлення змін. Метод CUSUM [4] ґрунтується на послідовному накопиченні інформації про відхилення контрольованої характеристики, тоді як у досліджуваному підході попередньо формується оцінка нетиповості спільної структурної конфігурації короткого фрагмента, після чого аналізуються взаємодоповнювальні характеристики її динаміки. Байєсівський підхід до оперативного виявлення точок зміни [5] передбачає послідовне оновлення ймовірного опису тривалості поточного режиму. Сучасні дослідження послідовного виявлення змін [6–8] додатково акцентують увагу на статистичній ефективності, обчислювальних обмеженнях і можливості оперативного аналізу потокових даних. Натомість запропонована інтеграція орієнтована насамперед на формування інтерпретованого структурно-динамічного вектора ознак, який надалі може використовуватися як вхідне представлення для процедур виявлення змін.

На відміну від методів глибокого та самокерованого навчання, сучасний стан яких узагальнено в дослідженнях [11–13], локальна ймовірнісна основа запропонованого підходу визначається аналітично або точним перебором за заданою базовою моделлю та не потребує навчання її розподілу на великій вибірці. Водночас це не усуває необхідності обґрунтованого вибору базової моделі, систем структурних статистик, параметрів агрегування та подальшого калібрування правил прийняття рішень. Як свідчать результати порівняльного дослідження методів виявлення аномалій [10], висновки щодо ефективності алгоритмів мають спиратися на узгоджені експериментальні умови. Тому встановлені методологічні відмінності не означають доведеної переваги запропонованого представлення за точністю або швидкістю виявлення змін.

На відміну від попередніх досліджень, присвячених багатокомпонентній моделі оцінювання [14] та методу секвенційно агрегованого аналізу структурних змін [15], у цій роботі основну увагу зосереджено на їх узгодженому застосуванні із системою точних спільних розподілів статистик коротких випадкових послідовностей [3]. Це забезпечує наскрізний перехід від ймовірного опису локальних бітових конфігурацій до характеристики розвитку структурних відхилень у потоці даних. Отримане представлення становить інтерес для моніторингу бінарних діагностичних подій інформаційно-технічних систем, зокрема за умов, коли зміна взаємного розташування подій може бути інформативною навіть без істотної зміни загальної частоти їх виникнення.

Водночас результати дослідження характеризують насамперед інформаційні властивості інтегрованого представлення та не є безпосереднім підтвердженням підвищення ефективності прогнозування небезпечних станів. Експеримент виконано на контрольованому синтетичному потоці, що дозволило дослідити реакцію компонентів на визначені структурні зміни, однак не охоплює різноманіття залежностей, характерних для реальних діагностичних даних. Подальше кількісне оцінювання ефективності використання отриманих характеристик потребує багаторазових порівняльних експериментів із незалежним калібруванням порогів на рівні цілих траєкторій, урахуванням статистичної залежності перекривних фрагментів та забезпеченням однакового рівня хибних спрацювань. Отже, проведене дослідження формує математично обґрунтовану основу для подальшого розроблення процедур раннього виявлення структурних змін і прогнозування розвитку небажаних станів.



ВИСНОВКИ ТА ПЕРСПЕКТИВИ ДОСЛІДЖЕННЯ

У роботі запропоновано інтегрований підхід до аналізу коротких випадкових послідовностей, який поєднує точне структурно-ймовірнісне оцінювання локальних бітових конфігурацій із секвенційним агрегуванням отриманих оцінок. Його принциповою особливістю є перехід від оцінювання нетиповості окремого фрагмента до багатокомпонентної характеристики розвитку структурних відхилень у потоці даних.

Використання точних спільних розподілів статистик дво- і триланцюжків забезпечує врахування взаємозв'язків між структурними характеристиками коротких фрагментів. Проведений обчислювальний експеримент показав, що сформовані на їх основі локальні оцінки реагують на зміну залежності між сусідніми елементами бінарного потоку навіть за незмінної граничної ймовірності появи окремих символів. Водночас встановлено, що поодинокі підвищення локальної структурної нетиповості не дозволяють повною мірою характеризувати розвиток відхилення, оскільки подібні значення можуть спостерігатися за різних режимів формування послідовності.

Секвенційне агрегування локальних оцінок дозволило сформувати п'ятикомпонентне представлення, яке одночасно характеризує рівень, тренд, накопичення, поточну тривалість і мінливість структурних відхилень. Аналіз отриманих траєкторій засвідчив функціональну відмінність зазначених компонентів та їхню взаємодоповнювальність. Зокрема, спрямоване наростання структурної нетиповості, її стійке збереження та поступове згасання можуть супроводжуватися близькими значеннями окремих локальних характеристик, але відрізняються за сукупністю секвенційних показників. Це обґрунтовує доцільність багатокомпонентного опису замість зведення послідовності спостережень до одного усередненого значення.

Основний науковий результат дослідження полягає в обґрунтуванні взаємодоповнювальності точного ймовірнісного опису локальних структурних конфігурацій і секвенційного аналізу їхньої динаміки в межах єдиного інтегрованого представлення. Такий підхід створює математичну основу для виділення потенційних ранніх структурних ознак розвитку відхилень у потоках бінарних діагностичних подій, зокрема у випадках, коли зміна їх взаємного розташування не супроводжується істотною зміною загальної частоти виникнення.

Перспективи подальших досліджень пов'язані з кількісним оцінюванням ефективності інтегрованого підходу на множині незалежних реалізацій і реальних діагностичних даних, визначенням його чутливості та затримки виявлення структурних змін за контрольованого рівня хибних спрацювань, а також розробленням і перевіркою моделей прогнозування розвитку небажаних станів інформаційно-технічних систем на основі сформованих багатокомпонентних характеристик.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Guibas, L. J., & Odlyzko, A. M. (1981). String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2), 183–208. [https://doi.org/10.1016/0097-3165\(81\)90005-4](https://doi.org/10.1016/0097-3165(81)90005-4)
2. Makri, F. S., & Psillakis, Z. M. (2023). Distribution of patterns of constrained length in binary sequences. *Methodology and Computing in Applied Probability*, 25, Article 90. <https://doi.org/10.1007/s11009-023-10068-5>
3. Masol, V. I., & Popereshnyak, S. V. (2021). Joint distribution of some statistics of random bit sequences. *Cybernetics and Systems Analysis*, 57(1), 139–145. <https://doi.org/10.1007/s10559-021-00337-x>
4. Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1–2), 100–115. <https://doi.org/10.1093/biomet/41.1-2.100>
5. Adams, R. P., & MacKay, D. J. C. (2007). *Bayesian online changepoint detection*. arXiv. <https://doi.org/10.48550/arXiv.0710.3742>
6. Wang, H., & Xie, Y. (2024). Sequential change-point detection: Computation versus statistical performance. *WIREs Computational Statistics*, 16(1), Article e1628. <https://doi.org/10.1002/wics.1628>
7. Gupta, M., Wadhvani, R., & Rasool, A. (2024). Comprehensive analysis of change-point dynamics detection in time series data: A review. *Expert Systems with Applications*, 248, Article 123342. <https://doi.org/10.1016/j.eswa.2024.123342>
8. Yang, Z., Eckley, I. A., & Fearnhead, P. (2024). A communication-efficient, online changepoint detection method for monitoring distributed sensor networks. *Statistics and Computing*, 34, Article 115. <https://doi.org/10.1007/s11222-024-10428-2>
9. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3), Article 56. <https://doi.org/10.1145/3444690>



10. Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9), 1779–1797. <https://doi.org/10.14778/3538598.3538602>
11. Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C. C., & Salehi, M. (2025). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1), Article 15. <https://doi.org/10.1145/3691338>
12. Jia, X., Xun, P., Peng, W., Zhao, B., Li, H., & Shen, C. (2025). Deep anomaly detection for time series: A survey. *Computer Science Review*, 58, Article 100787. <https://doi.org/10.1016/j.cosrev.2025.100787>
13. Sánchez-Ferrera, A., Calvo, B., & Lozano, J. A. (2026). A review on self-supervised learning in time series anomaly detection: Recent advances and open challenges. *ACM Computing Surveys*, 58(5), Article 119. <https://doi.org/10.1145/3770575>
14. Popereshnyak, S. V. (2026). Multicomponent model for sequential aggregated assessment of short sequences for system state change detection. *Problems in Programming*, (3), 69–82. <https://doi.org/10.15407/pp2026.03.069>
15. Popereshnyak, S. V. (2026). Method for sequentially aggregated assessment of structural changes in data streams for detecting degradation of technical systems. *Telecommunication and Information Technologies*, 3(92), 190–200. <https://doi.org/10.31673/2412-4338.2026.031919>

**Svitlana V. Popereshnyak**

Candidate of Physics and Mathematics, Associate of Professor,

Associate Professor of the Department of Informatics and Software Engineering,

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

State University of Information and Communication Technologies, Kyiv, Ukraine

ORCID ID 0000-0002-0531-9809

spopereshnyak@gmail.com

INTEGRATION OF EXACT STRUCTURAL-PROBABILISTIC ASSESSMENT AND SEQUENTIAL AGGREGATION OF SHORT RANDOM SEQUENCES

Abstract. This paper investigates the integration of exact structural-probabilistic assessment of short random sequences with the sequential aggregation of local scores to characterize the evolution of structural changes in binary data streams. The need to account not only for symbol occurrence frequencies but also for their relative arrangement and the temporal ordering of local deviations is substantiated. The theoretical foundation of the proposed approach is provided by exact joint distributions of occurrence-count statistics for two- and three-symbol patterns, which account for dependencies and overlaps among structural configurations in short sequence fragments. These distributions are used to derive local structural atypicality scores relative to a baseline probabilistic model. A two-stage aggregation procedure is proposed, comprising structural aggregation of the scores within individual fragments, followed by sequential aggregation of their ordered trajectory. The resulting representation characterizes the level, trend, accumulation, current duration, and variability of structural deviations. A reproducible computational experiment was conducted using a synthetic binary data stream with controlled changes in its generating regimes. The results demonstrate the response of local scores to changes in dependence between adjacent elements while the marginal symbol probabilities remain unchanged. Analysis of the sequential components revealed differences between increasing, persistent, and diminishing deviations that cannot be identified solely from the average level of local atypicality. The scientific significance of these findings lies in establishing the complementary nature of exact local structural characterization and the analysis of its sequential dynamics. The integrated representation provides a foundation for identifying potential early structural indicators of undesirable changes in diagnostic data streams. Future research will focus on validation using real-world data, evaluation of detection sensitivity and delay under a controlled false-alarm rate, and identification of the conditions under which the proposed approach can be effectively applied in practice.

Keywords: mathematical modeling; random sequences; structural analysis; joint distributions; streaming data; sequential aggregation; change detection; system monitoring.

REFERENCES

1. Guibas, L. J., & Odlyzko, A. M. (1981). String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, Series A*, 30(2), 183–208. [https://doi.org/10.1016/0097-3165\(81\)90005-4](https://doi.org/10.1016/0097-3165(81)90005-4)
2. Makri, F. S., & Psillakis, Z. M. (2023). Distribution of patterns of constrained length in binary sequences. *Methodology and Computing in Applied Probability*, 25, Article 90. <https://doi.org/10.1007/s11009-023-10068-5>
3. Masol, V. I., & Popereshnyak, S. V. (2021). Joint distribution of some statistics of random bit sequences. *Cybernetics and Systems Analysis*, 57(1), 139–145. <https://doi.org/10.1007/s10559-021-00337-x>
4. Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1–2), 100–115. <https://doi.org/10.1093/biomet/41.1-2.100>
5. Adams, R. P., & MacKay, D. J. C. (2007). *Bayesian online changepoint detection*. arXiv. <https://doi.org/10.48550/arXiv.0710.3742>
6. Wang, H., & Xie, Y. (2024). Sequential change-point detection: Computation versus statistical performance. *WIREs Computational Statistics*, 16(1), Article e1628. <https://doi.org/10.1002/wics.1628>
7. Gupta, M., Wadhvani, R., & Rasool, A. (2024). Comprehensive analysis of change-point dynamics detection in time series data: A review. *Expert Systems with Applications*, 248, Article 123342. <https://doi.org/10.1016/j.eswa.2024.123342>



8. Yang, Z., Eckley, I. A., & Fearnhead, P. (2024). A communication-efficient, online changepoint detection method for monitoring distributed sensor networks. *Statistics and Computing*, 34, Article 115. <https://doi.org/10.1007/s11222-024-10428-2>
9. Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3), Article 56. <https://doi.org/10.1145/3444690>
10. Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9), 1779–1797. <https://doi.org/10.14778/3538598.3538602>
11. Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C. C., & Salehi, M. (2025). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1), Article 15. <https://doi.org/10.1145/3691338>
12. Jia, X., Xun, P., Peng, W., Zhao, B., Li, H., & Shen, C. (2025). Deep anomaly detection for time series: A survey. *Computer Science Review*, 58, Article 100787. <https://doi.org/10.1016/j.cosrev.2025.100787>
13. Sánchez-Ferrera, A., Calvo, B., & Lozano, J. A. (2026). A review on self-supervised learning in time series anomaly detection: Recent advances and open challenges. *ACM Computing Surveys*, 58(5), Article 119. <https://doi.org/10.1145/3770575>
14. Popereshnyak, S. V. (2026). Multicomponent model for sequential aggregated assessment of short sequences for system state change detection. *Problems in Programming*, (3), 69–82. <https://doi.org/10.15407/pp2026.03.069>
15. Popereshnyak, S. V. (2026). Method for sequentially aggregated assessment of structural changes in data streams for detecting degradation of technical systems. *Telecommunication and Information Technologies*, 3(92), 190–200. <https://doi.org/10.31673/2412-4338.2026.031919>

Отримано редакцією журналу / Received: 16.08.26

Прорецензовано / Revised: 30.08.26

Схвалено до друку / Accepted: 26.10.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.