DOI: 10.35681/1560-9189.2022.24.1.262673

УДК 621.391:519.216.3

Д. В. Ланде, В. В. Юзефович
Інститут проблем реєстрації інформації НАН України
вул. М. Шпака, 2, 03113 Київ, Україна

Лінгвістичний підхід до прогнозування часових рядів

Запропоновано методи прогнозування динамічних часових рядів (у тому числі й нестаціонарних), що базуються на лінгвістичному підході, а саме: дослідженні входжень і повторюваності так званих N-грам, які застосовуються в комп'ютерній лінгвістиці при створенні статистичних перекладачів, виявленні плагіату, дублікатів документів. На відміну від застосування в лінгвістиці, метод може бути розширений урахуванням кореляцій послідовностей сталих словосполучень, а також трендів. При цьому запропоновані методи не вимагають попереднього дослідження та визначення характеристик часових рядів і складного налаштування вхідних параметрів моделі прогнозування. Запропоновані методи дозволяють з високим рівнем автоматизації здійснювати короткострокові та середньострокові прогнози динамічних часових рядів, яким притаманні тренди та циклічність, зокрема, рядів динаміки публікацій у системах контент-моніторингу. Суттєвою перевагою підходу є відсутність вимог до стаціонарності часових рядів і мала кількість параметрів налаштування.

***Ключові слова*:** *часовий ряд, прогнозування, метод N-грам, квантування, тренд, модель, критерій подібності, кореляція, лінійна регресія.*

Вступ

Задача прогнозування різних часових рядів є актуальною та затребуваною у самих різних прикладних областях і при цьому не має універсальних вирішень, про що свідчить наявність, за деякими оцінками, понад 200 методів вирішення цієї задачі.

Існує багато задач прогнозування, де прогноз має отримуватись оперативно та з мінімальним залученням фахівців-аналітиків для організації і здійснення цього процесу. Наприклад, при вирішенні задачі паралельного прогнозування інтенсивності інформаційних потоків різних тематик [1], або прогнозування значень численних параметрів деякої великої складної системи в задачі моніторингу її стану [2]. У таких випадках поряд із вимогою точності прогнозу виникає додаткова вимога щодо мінімальної кількості параметрів моделі прогнозування, що підля-

гають налаштуванню та просте (швидке, без багатьох ітерацій) налаштування останніх. Особливої уваги заслуговують динамічні часові ряди (зміна значень яких переважно викликана динамікою процесу, що спостерігається, а не випадковими факторами), а у загальному випадку — нестаціонарні (на окремих ділянках яких математичне очікування, середнє, коваріація не є сталими). На сьогодні значне поширення отримав метод прогнозування часових рядів, що передбачає побудову авторегресійної інтегрованої моделі ковзного середнього ARIMA (Autoregressive Integrated Moving Average), сезонна модифікація моделі ARIMA (SARIMA) та ряд подібних, зокрема Generalized ESD (Extreme Studentized Deviant), Seasonal Hybrid ESD [3]. Також популярним є хвильовий метод Сорнетте [4], який потребує великої кількості параметрів налаштування та може застосовуватися для аналізу нестаціонарних часових рядів, навіть криз в економічній галузі [5].

Ці методи вважаються дуже гнучкими, надають можливість «тонкого» налаштування параметрів моделі часового ряду, однак потребують попереднього детального дослідження цього ряду та суттєвих знань і значного досвіду аналітика, що здійснює модельний прогноз (а отже додаткових часових і людських витрат), або ж використання потужних за функціональністю спеціалізованих програмних продуктів автоматизованого вирішення задачі прогнозування, в яких зазвичай слабкою стороною є підсистема пояснень отриманих результатів. При цьому аналітику важно зрозуміти, чому отримано саме такий, а не інший прогноз, що фактично робить його стороннім спостерігачем процесу, який має вірити отриманому результату практично без пояснень, «на слово». Крім того, такі програмні продукти непросто, а часом і неможливо вбудувати в інші автоматизовані системи, де задача прогнозування часових рядів є внутрішньою частковою задачею, підпорядкованою більш системному цільовому завданню.

Разом із тим, існують і інші, більш прості методи прогнозування часових рядів, що не передбачають складних налаштувань, і при цьому дозволяють у багатьох випадках отримувати досить точні прогнози. До таких методів відносяться так звані експоненціальні методи, що передбачають, наприклад, побудову моделі простого ковзного середнього, моделей Брауна або Хольта (яку ще називають моделлю подвійного експоненціального згладжування) [6]. Модель Хольта, на відміну від простого експоненціального згладжування (моделі Брауна), чи ковзного середнього із заданим вікном, передбачає для отримання прогнозу згладжування не тільки значень часового ряду, але і його тренду. Модель Хольта з двома вхідними параметрами для часового ряду $X = x_1, x_2, ..., x_k, ..., x_K$ довжиною K має вигляд:

$$\begin{aligned} \hat{x}_k &= (1-\xi)x_k + \xi(\hat{x}_{k-1} + \hat{T}_{k-1}), \\ \hat{T}_k &= (1-\varphi)(\hat{x}_k - \hat{x}_{k-1}) + \varphi\hat{T}_{k-1}, \\ \hat{x}_{k+p} &= \hat{x}_k + P\hat{T}_k, \end{aligned} \qquad (1)$$

де $\hat{x}_k, \hat{x}_{k-1}$ — оцінки згладжених значень ряду на k-му та $(k-1)$ кроках згладжування;

ξ — коефіцієнт згладжування значень ряду (ξ змінюється від 0 до 1);

$\hat{T}_k, \hat{T}_{k-1}$ — оцінки тренду зміни значень часового ряду на k-му та $(k-1)$ кроках згладжування;

φ — коефіцієнт згладжування тренду ($\varphi \in [0,1]$);

P — кількість майбутніх значень ряду, на яку необхідно здійснити прогноз (час упередження).

Складність використання такої моделі полягає в необхідності підбору двох коефіцієнтів згладжування (ξ та φ), які спільно впливають на результат прогнозування, і при цьому такий підбір не може здійснюватися послідовно, оскільки рівняння згладжування значень ряду та тренду пов'язані одними й тими самими оцінками $\hat{x}$ (значення обох коефіцієнтів впивають на обидва рівняння).

Існує також удосконалений метод Хольта — метод Хольта-Вінтерса [7] з трьома вхідними параметрами, який додатково включає до моделі параметр сезонності часового ряду і застосовується, якщо ряд містить циклічні варіації. При цьому часовий ряд має попередньо додатково перевірятися на наявність такої сезонності, наприклад, шляхом аналізу його автокореляційної чи часткової автокореляційної функції.

Отже, застосування навіть простих методів прогнозування у випадку нестаціонарних часових рядів, потребують доволі активної ролі аналітика, що ускладнює процес їхнього автоматизованого використання для вирішення задачі отримання прогнозу.

Мета роботи

Метою роботи є розробка методу прогнозування, що, з одного боку, дозволяє здійснювати прогнозування динамічних часових рядів, у тому числі і нестаціонарних, а з іншого — не потребує складних налаштувань і містить мінімальну кількість вхідних параметрів.

У роботі запропоновано новий, так званий лінгвістичний підхід до прогнозування, на базі якого будується ряд методів. Підхід передбачає задавання лише двох параметрів для здійснення прогнозу часового ряду, однак ці параметри не є прямо пов'язаними між собою, задаються окремо та не потребують глибокого аналізу часового ряду. Запропонований підхід дозоляє здійснювати прогнозування як стаціонарних, так і нестаціонарних рядів, не передбачає аналізу часового ряду на сезонність, однак автоматично враховує її при формуванні прогнозних значень.

Лінгвістичний і лінгвокореляційний методи прогнозування динамічних часових рядів

В основу лінгвістичного методу прогнозування покладено метод N-грам, який використовується для обробки природної мови, зокрема для передбачення наступних слів у висловлюванні, якщо відомі усі попередні (відповідно, широко застосовується в комп'ютерному перекладі, системах виявлення дублікатів і плагіату) [8]. При цьому вважається, що умовна ймовірність появи наступного слова залежіть від попередніх слів і їхніх послідовностей. Для прогнозування часового ряду, як відомо, також використовується припущення, що ймовірність і його зна-

чення визначаються попередніми значеннями. Ідея застосування методу *N*-грам для прогнозування часових рядів базується на модельному припущені, що часовий ряд можна вважати деяким осмисленим текстом, а нове значення цього ряду — наступним словом, який треба визначити на основі аналізу попереднього «тексту».

Отже, для визначення невідомого «словосполучення» (прогнозу ряду) довжиною $P(\hat{X}_p)$ необхідно здійснити пошук у повному «тексті» (часовому ряді) X «словосполучення» X_N довжиною N ($X_N = x_{K+1}, x_{K+2}, ..., x_N$), що є тотожним (подібним) «словосполученню» X_K, яке безпосередньо передує моменту прогнозу ($X_K = x_{K-N}, x_{K-N+1}, ..., x_K$), і вважати, що P «слів», які знаходяться одразу після знайденого «словосполучення» X_N – $X_p = x_{N+1}, x_{N+2}, ..., x_{N+P}$ утворюють шукане «словосполучення» (прогнозні значення ряду), тобто $\hat{X}_p = X_p$.

Для порівняння словосполучень може використовуватися поелементна різниця значень ряду (різницевий критерій), яка у випадку їхнього повного збігу дорівнює 0. Разом із тим очевидно, що пошук у повному тексті словосполучення X_N, яке точно співпадає зі словосполученням X_K не обов'язково буде успішним. Отже, підвищення ймовірності знаходження у тексті тотожних словосполучень можна забезпечити шляхом зменшення кількості можливих різних слів у тексті — лексичного запасу мови, або зменшенням їхньої довжини. У іншому випадку, замість пошуку тотожних словосполучень необхідно шукати певним чином подібні, що очевидно може погіршити якість прогнозу. Критерієм подібності словосполучень може бути мінімальне значення поелементної різниці значень ряду, або максимум коефіцієнта кореляції (кореляційний критерій), який розраховується за виразом:

$$R_{X_N X_K} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{N}(y_k - \bar{y})^2}}, \qquad (2)$$

де x_k, y_k — елементи рядів X_N та X_K відповідно; $\bar{*}$ — усереднення (оцінка математичного очікування).

При застосуванні критерію (2) будемо говорити про лінгвокореляційний метод прогнозування.

Відповідно до зазначеного вище, алгоритм вирішення задачі отримання прогнозу часового ряду X на час упередження P буде наступним.

Із часового ряду X вибирається N значень підряд, що безпосередньо передують моменту прогнозу K. Значення N — це перший параметр моделі прогнозування, який необхідно задати аналітику (досліднику) до початку процесу прогнозування.

Замість параметра N на практиці буває зручно використовувати інший параметр M, дійсне число, яке задає кратність значення P. Цей параметр вибирається із розкладу $N = \lceil M \cdot P \rceil$, де M приймає тим більше значення, чим менше значення P, а $\lceil * \rceil$ — функція округлення до найближчого більшого цілого (*ceiling*).

Забігаючи наперед, як показали дослідження, M доцільно задавати, виходячи із наступних умов:

— якщо $P \geq 5$, то $1 \leq M \leq 2$, при цьому, чим більше P, тим менше M;
— якщо $P < 5$, то $2 < M \leq 5$, при цьому, чим менше P, тим більше M.

Далі, для збільшення ймовірності знаходження тотожних словосполучень часовий ряд, що розглядається, піддається процедурі квантування — розбиття діапазону можливих значень ряду на кінцеве число рівнів з округленням дійсних значень до найближчих до них рівнів. Метою зазначеної процедури є зведення усіх можливих значень часового ряду до їхнього обмеженого переліку — «слів з лексичного запасу тексту». Крок квантування має бути максимальним, але таким, що дозволяє зберегти основну динаміку зміни значень ряду, тобто кількість різних «слів» мови має бути мінімально необхідною.

Приклад здійснення такої процедури квантування показано на рис. 1.

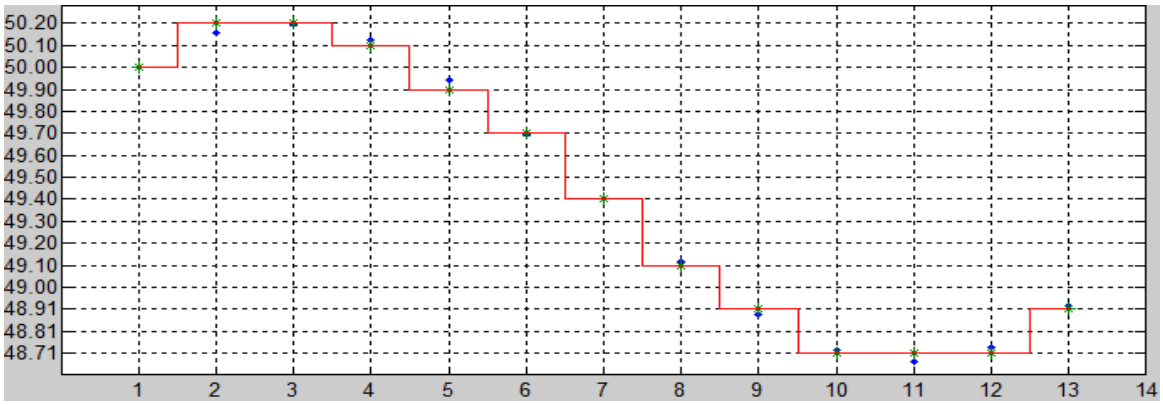


Рис. 1. Результат квантування часового ряду

Представлений на рис. 1 фрагмент часового ряду в результаті процедури квантування описано дев'ятьма різними «словами», що розташовані в певній послідовності.

Фактично крок квантування, або кількість «слів» (S) штучної мови, який описує часовий ряд є другим (і останнім) параметром, що має бути заданий аналітиком (дослідником) для здійснення процедури прогнозування. Особливістю даного методу прогнозування є те, що незначні зміни параметрів N та S не впливають суттєво на результат прогнозування. Їх легко підібрати шляхом використання кінцевої частини наявного часового ряду як тестової для перевірки якості прогнозу.

Як приклад застосування лінгвістичного методу розглянемо задачу прогнозування на час упередження $P = 20$ деякого детермінованого нестаціонарного часового ряду дійсних чисел довжиною $K = 100$, та з максимальним розмахом значень у 4 одиниці, який характеризується візуально явно вираженою сезонністю, однак не має міжсезонних змін тренду. Такий ряд може бути сформовано значеннями будь-якої періодичної функції, наприклад синусоїди. Відповідно до зазначеного вище правила задамо значення $M = 1$ (тоді $N = P$) та $S = 32$ (крок квантування — 4/30 = 0,125). Такий часовий ряд і результат прогнозування «лінгвістичним» методом показано на рис. 2.

Навіть без розрахунків похибок прогнозу видно, що запропонований метод у даному випадку дозволяє досить якісно вирішувати задачу прогнозування.

Використання квантування часового ряду також підвищує стійкість прогнозу до незначних випадкових збурень значень ряду без використання процедур згладжування. На рис. 3 показано результат прогнозування того самого ряду, але з випадковими збуреннями (закон розподілення збурень — рівномірний у діапазоні від –0,15 до 0,15 одиниць, з нульовим математичним очікуванням).

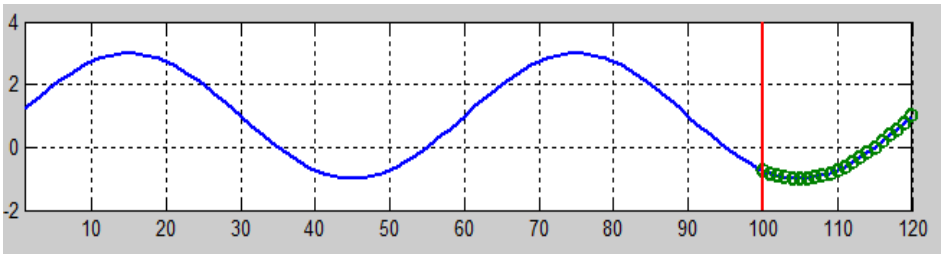


Рис. 2. Результат прогнозування часового ряду без тренду «лінгвістичним» методом

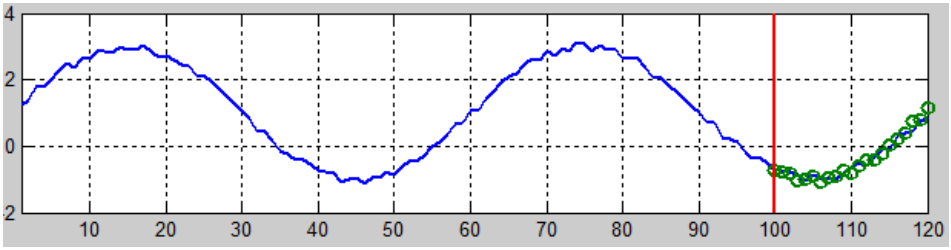


Рис. 3. Результат прогнозування часового ряду «лінгвістичним» методом у разі наявності випадкових збурень значень часового ряду

Розглянемо більш складну задачу прогнозування детермінованого нестаціонарного часового ряду дійсних чисел, який характеризується сезонністю та має лінійний тренд. Очевидно, що в такому випадку повний (непомилковий) збіг словосполучень X_N та X_K може не виконуватись. Отже, перед порівнянням словосполучень X_N та X_K необхідно вилучити їхні тренди. Тренд часового ряду можна розрахувати методом найменших квадратів. Для лінійного тренду, що описується рівнянням $y = Bx + A$ коефіцієнти B та A розраховуються за відомими виразами:

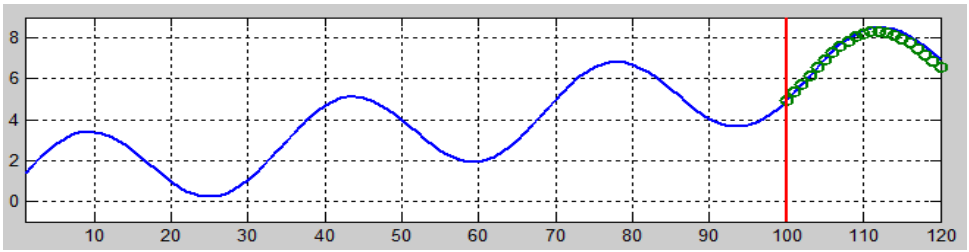
$$A = \overline{y} - B\overline{x},$$
$$B = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{x^2 - \left(\overline{x}\right)^2}, \qquad (3)$$

де x — порядковий номер елемента ряду; y — значення ряду.

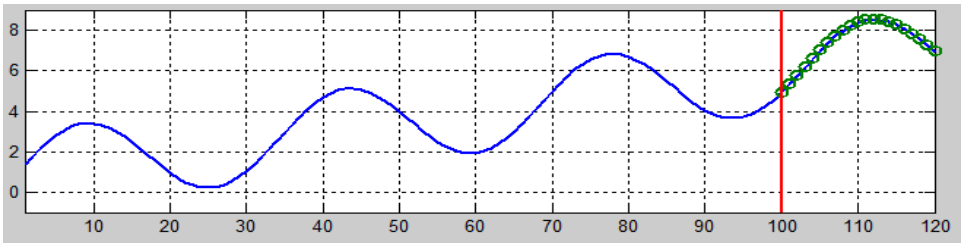
Далі, після знаходження «словосполучення» X_N та, відповідно X_P, до значень останнього необхідно додати тренд, що характеризує X_K.

На рис. 4 показано результати прогнозування детермінованого нестаціонарного часового ряду дійсних чисел ($K = 100$, $P = 20$, $M = 1$ ($N = P$) та $S = 30$), що характеризується сезонністю та має «глобальний» лінійний тренд.

Як видно з рис. 4, обидва методи дозволяють з високою точністю прогнозувати динамічні часові ряди із сезонністю та лінійним трендом. При цьому кореляційний критерій подібності, в даному випадку, показав дещо кращий результат. Слід зауважити, що проведені дослідження не показали стійкої переваги використання того, чи іншого критерію.



а)



б)

Рис. 4. Результати прогнозування часового ряду із лінійним трендом:
а) лінгвістичним методом; б) лінгвокореляційним методом

Найбільш загальним є варіант динамічного часового ряду з нелінійним трендом, але на базі наявного досвіду можна припустити, що використання лінійної регресії до послідовностей X_N та X_K, а не усього часового ряду X загалом дозволить відслідковувати не тільки лінійні «глобальні» тренди ряду X.

На рис. 5 показано результат прогнозування лінгвістичним методом із використанням лінійної регресії нестаціонарного часового ряду натуральних чисел, що має складній нелінійний тренд ($K = 100$, $P = 20$, $M = 1$, $S = 30$). Як і припускалося, результат прогнозування є досить точним.

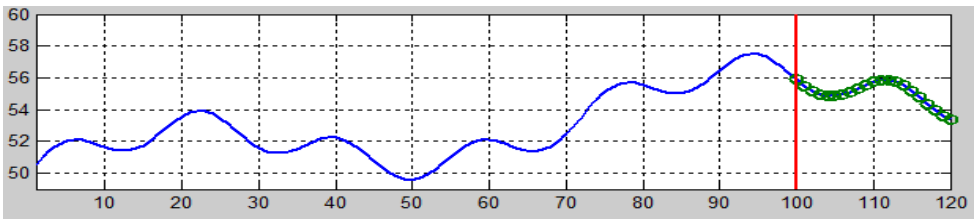


Рис. 5. результат прогнозування «лінгвістичним» методом із використанням лінійної регресії часового ряду з нелінійним трендом

На рис. 6 показано результат прогнозування запропонованим лінгвістичним методом схожого часового ряду до випадку, представленого на рис. 5, однак цей результат є помітно гіршим. Причиною погіршення точності прогнозу у випадку,

наведеному на рис 6, є відсутність у минулому «словосполучення» X_N тотожного X_K, через те, що нелінійний тренд не в повному обсязі «проявив» себе в межах часового ряду X. Тим не менше, короткостроковий прогноз (при $P \leq 10$) у випадку, що показаний на рис. 6, є також досить точним.

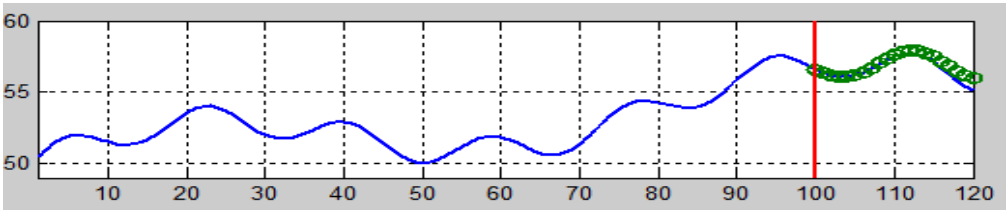
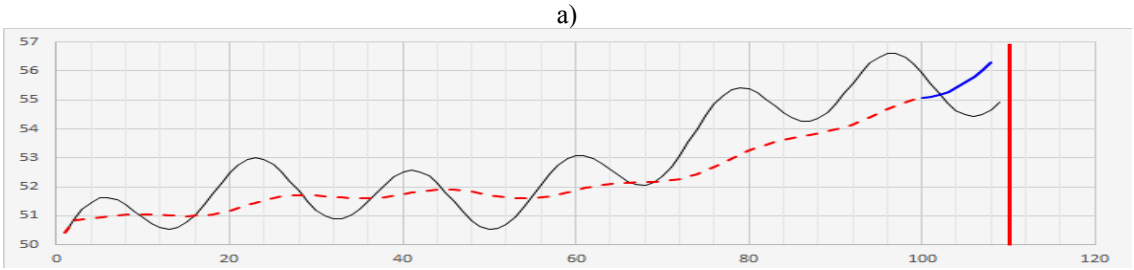


Рис. 6. Результат прогнозування часового ряду із нелінійним трендом «лінгвістичним» методом із використанням лінійної регресії

Для порівняння з іншими розповсюдженими методами прогнозування на рис. 7 показано результат прогнозування часового ряду, тотожного тому, який показано на рис. 6 із використанням моделі Хольта. Як видно з рис. 7, у такому випадку модель Хольта можна використовувати лише для отримання короткострокового прогнозу — $P \leq 5$ (рис. 7,*а*), або ж для формування «глобального» тренду для грубих довгострокових прогнозів (рис. 7,*б*).

а)



б)

Рис. 7. Результат прогнозування часового ряду з нелінійним трендом із застосуванням моделі Хольта

Залишилося перевірити, чи запропонований пошук лінійних трендів у вибірках ряду X_N та X_K методом лінійної регресії не призведе до внесення помилок у прогноз, коли насправді глобального тренду не існує. Якщо такий вплив є суттєвим, до лінгвістичного та лінгвокореляційного методів необхідно додати додатковий етап аналізу глобального тренду часового ряду до початку прогнозування. Це,

у свою чергу, може нівелювати переваги запропонованих методів та зменшити їх універсальність. Однак проведені дослідження спростували це перестереження.

Покажемо це на прикладі. Розглянемо випадок, аналогічний наведеному на рис. 2, де часовий ряд не містить глобального тренду. Як видно з рис. 8, пошук і врахування трендів вибірок X_N та X_K за допомогою побудови лінійної регресії за відсутності глобального тренду не призвів до помітного погіршення прогнозу.

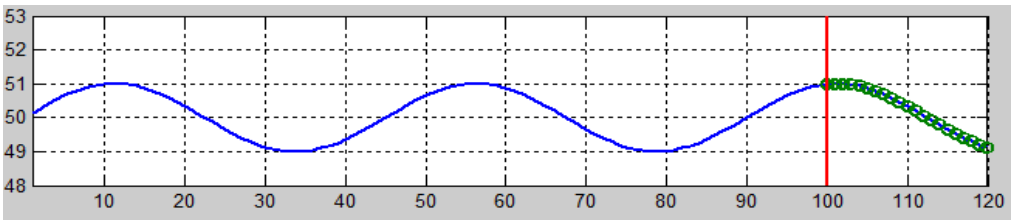


Рис. 8. Результат прогнозування часового ряду без тренду лінгвокореляціним методом із застосуванням лінійної регресії

На рис. 9 показано результат практичного застосування лінгвокореляційного методу прогнозування часових рядів для отримання прогнозних значень кількості публікацій у російськомовних засобах масової інформації зі згадуванням прізвища президента США, на основі відповідного часового ряду (з кроком в одну добу), отриманого системою контент-моніторингу [1, 9].

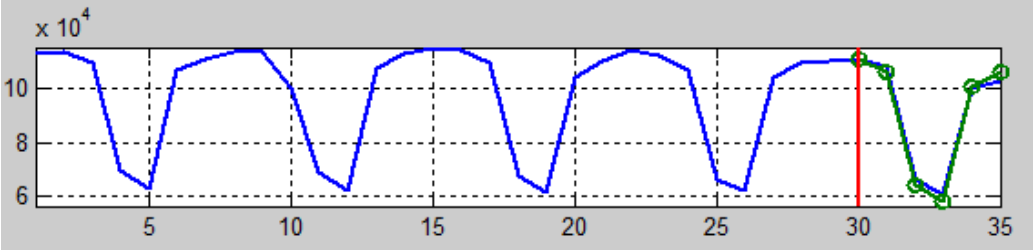


Рис. 9. Результат прогнозування кількості публікацій у російськомовних ЗМІ зі згадуванням прізвища президента США запропонованим лінгвістичним методом

Як видно з рис. 9, прогнозні значення ряду досить точно відображають динаміку зміни частоти публікацій на задану тему і в цілому відповідають контрольним дійсним значенням.

Таким чином, запропоновані методи лінгвістичного (без урахування глобального тренду) та лінгвокореляційного (з пошуком тренду) прогнозування динамічних часових рядів продемонстрували свою дієвість і можливість застосування для здійснення прогнозу різних динамічних нестаціонарних часових рядів.

Висновки

Запропоновані методи лінгвістичного та лінгвокореляційного прогнозування динамічних часових рядів, які базуються на лінгвістичному підході, використанні методу N-грам і лінійної регресії для аналізу локальних трендів часових рядів, відносяться до класу авторегресійних методів, що легко автоматизуються, оскіль-

ки мають лише два вхідні параметри, які не потребують точного налаштування. У роботі наведено приклади застосування методів, показано їхні особливості.

Методи дозволяють здійснювати прогнозування на короткостроковий і середньостроковий періоди динамічних нестаціонарних часових рядів як з лінійними, так і нелінійними трендами. Також вони дозволяють з високим рівнем автоматизації здійснювати прогнози часових рядів, яким притаманна циклічність, зокрема, ряди динаміки публікацій у системах контент-моніторингу, без додаткового аналізу їхньої сезонності.

Суттєвою перевагою підходу є відсутність вимог до стаціонарності часових рядів і мала кількість параметрів налаштування.

Подальші дослідження можуть бути спрямовані на дослідження різних критеріїв подібності фрагментів часового ряду, використання нелінійних критеріїв подібності (наприклад, хвильових), пошуку способів автоматичного визначення раціонального кроку квантування часового ряду.

1. Lande D., Subach I., Puchkov A. A System for Analysis of Big Data from Social Media. *Information & Security: An International Journal*. 2020. **47**. No. 1. P. 44–61. DOI: https://doi.org/10.11610/isij.4703

2. Юзефович В.В., Цибульська Є.О. Підхід до формування інформаційного ресурсу єдиного інформаційного простору системи організаційного управління. Збірник доповідей XXI Міжнародної науково-практичної конференції «Інформаційні технології та безпека». 9 грудня 2021. С. 185–192 (ІТБ-2021).

3. Nielsen A. Practical Time Series Analysis: Prediction with Statistics and Machine Learning. O'Reilly Media, Inc., 2020. 504 p. ISBN 978-1-492-04165-8.

4. Sornette D. Why Stock Markets Crash: Critical Events in Complex Financial Systems. Princeton University Press, 2004. DOI: https://doi.org/10.23943/princeton/ 9780691175959.001.0001.

5. Sornette D., Becke S. Financial Market and Systemic Risks. in Market Risk and Financial Markets Modeling. Berlin, Heidelberg: Springer-Verlag, 2012. DOI: https://doi.org/10.1007/978-3-642-27931-7.

6. Li Xiaochen. Comparison and Analysis Between Holt Exponential Smoothing and Brown Exponential Smoothing Used for Freight Turnover Forecasts. In Conference Proc.: Intelligent System Design and Engineering Applications (ISDEA), 2013. Third International Conference. 2013. DOI: 10.1109/ISDEA.2012.112

7. Yohana James Mgale, Yunxian Yan, Shauri Timothy. A Comparative Study of ARIMA and Holt-Winters Exponential Smoothing Models for Rice Price Forecasting in Tanzania. *Open Access Library Journal*. 2021. No. 8. P. 1–9. DOI: https://doi.org/10.4236/oalib.1107381.

8. Rahman M.M., Kabir M.F., Huda M.N. A corpus based n-gram hybrid approach of bengali to english machine translation. In 2018 21st International Conference of Computer and Information Technology (ICCIT). IEEE, 2018. P. 1–6. DOI: https://doi.org/10.1109/ICCITECHN.2018.8631938.

9. Lande D., Shnurko-Tabakova E. OSINT as a part of cyber defense system // Theoretical and Applied Cybersecurity, 2019. No. 1. P. 103–108. DOI: https://doi.org/10.20535/tacs.2664-29132019.1.169091.

Надійшла до редакції 25.04.2022