

**ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА**

Механіко-математичний факультет
Кафедра математичної економіки, економетрії,
фінансової та страхової математики

Пояснювальна записка

до кваліфікаційної роботи магістра

на тему:

**"Аналіз та прогнозування залишків
на рахунках клієнтів банку"**

Виконав:

студент групи МТФМ-21с
спеціальності 111 Математика
Чорний Михайло Васильович

Керівник:

кандидат фіз.-мат. наук, доцент Прокопишин І. А.

Рецензент:

Львів – 2021

Зміст

ВСТУП	3
1. ПРОБЛЕМА УПРАВЛІННЯ ПАСИВАМИ БАНКУ	4
1.1. Ризик ліквідності і пасиви банку	4
1.2. Прогнозування залишків	6
2. ЧАСОВІ РЯДИ	8
2.1. Основні положення	8
2.1.1. Означення	8
2.1.2. Ергодичність.....	9
2.1.3. Приклади часових рядів	10
2.1.4. Розрахунок характеристик часового ряду.....	11
2.2. Моделювання та прогнозування часових рядів	12
2.2.1. Загальні міркування.....	12
2.3. Трендові моделі	15
2.3.1. Виділення тренду	15
2.3.2. Аналіз сезонності.....	16
3. ЛІНІЙНЕ МОДЕЛЮВАННЯ ЧАСОВИХ РЯДІВ.....	18
3.1. Модель ковзного середнього	18
3.2. Авторегресійна модель	19
3.3. Модель ARMA	21
3.4. Лінійні нестационарні моделі	24
3.4.1. Модель ARIMA	24
3.4.2. Модель SARIMA.....	25
4. ПРАКТИЧНА ІМПЛЕМЕНТАЦІЯ МОДЕЛІ SARIMA.....	26
4.1. Загальна схема аналізу та прогнозування.....	26
4.2. Аналіз та прогнозування залишків на рахунках клієнтів банку.....	27
ВИСНОВКИ.....	36
СПИСОК ЛІТЕРАТУРИ.....	37

ВСТУП

Метою роботи є проведення аналізу динаміки залишків на рахунках клієнтів банку та їх прогнозування з використанням авторегресійної інтегрованої моделі ковзного середнього з урахуванням сезонності – моделі SARIMA.

Перший розділ є вступним. У ньому розглядається проблема ризику ліквідності, зокрема проблема управління депозитним ризиком. Підкреслена важливість прогнозування залишків на депозитах та рахунках клієнтів банку. Приведено короткий огляд публікацій з цього питання.

У другому розділі детально описано основи теорії часових рядів, загальні проблеми їх моделювання та методи побудови прогнозу. Розглянуто методи оцінювання моделей, критерії, які використовують для порівняння моделей.

У третьому розділі вивчається авторегресійна інтегрована модель ковзного середнього з врахуванням сезонності (Seasonal Autoregressive Integrated Moving Average – SARIMA).

Практична імплементація моделі SARIMA з використанням електронних таблиць, мов програмування R та Python, розглянута у четвертому розділі. Проведено аналіз залишків на рахунках клієнтів банку та здійснено прогноз.

1. ПРОБЛЕМА УПРАВЛІННЯ ПАСИВАМИ БАНКУ

1.1. Ризик ліквідності і пасиви банку

Аналіз та моделювання фінансових ризиків є актуальною проблемою для діяльності будь-якої компанії, зокрема для фінансових компаній та банків.

Різні інституції дають свої класифікації фінансових ризиків. Basel Committee on Banking Supervision (BCBS) визначає 8 категорій ризиків, Нацбанк України встановлює 9 груп ризиків, а класифікатор GARP – 7 груп та 23 підгруп.

Узагальнено можна вказати такі основні види фінансових ризиків: ринкові ризики, кредитні ризики, операційні ризики, ризики ліквідності та ризики бізнес-подій.

Більшість з названих категорій ризиків є загально зрозумілими. Так ринкові ризики пов'язані з коливанням ринкових цін, валютних курсів, процентних ставок, тощо. Кредитні ризики пов'язані з контрагентами та виконання ними своїх зобов'язань. Операційні ризики зумовлені помилками персоналу та технічних пристроїв, діями зловмисників. Ризики бізнес-подій це – форс-мажорі обставини, викликані кліматичними процесами, військовими діями, політичними та правовими чинниками, втратою репутації тощо.

Окремо розглянемо ризики ліквідності, суть яких полягає у неможливості купити чи продати деякі активи за обмежений час, а також з виникненням ризику балансової ліквідності, наприклад дефіциту готівкових коштів.

Для зменшення ризику ліквідності важливим є правильне управління пасивами банку.

Пасиви банку – це кошти, які банк використовує в своїх операціях, тому часто пасиви банку називають «джерелами» ресурсів банку. Є два основних джерела банківських ресурсів: власники банку (так звані акціонери), а також вкладники та кредитори. Кошти власників банку називають власними коштами, а ресурси вкладників та кредиторів – залученими. Теж пасиви банку формують рахунок коштів статутного капіталу, дохід від розміщення

випущених цінних паперів, емісійні доходи, залучені кошти населення Усі операції структур банку (починаючи від видачі кредиту до закупівлею канцелярії для працівників банку), безпосередньо залежать від власних капіталів і засобів, залучених зі сторони.

Основними проблемами управління пасивами банку є:

- слабка капітальна база
- незацікавленість керівництва банку в залученні коштів клієнтів
- міжбанківська конкуренція
- рівень грошового та фінансового ринку України
- вплив держави і державних органів на банки

У банківництві ризики виникають не лише при здійсненні операцій з активами, а також при операціях з пасивами. В останньому випадку важливими є ризики депозитної діяльності. Депозитні вклади, а також залишки на рахунках клієнтів є одним з основних джерел ресурсів банків.

Тому проблема дослідження ризиків депозитної діяльності, дослідження їх причин та закономірностей, аналізу залишків на рахунках та їх прогнозування є важливою та актуальною [1, 2, 7, 10, 11].

Ліквідність – один з основних і найбільш складних факторів, що визначають фінансовий стан і надійність банку. Величезний вплив на банківську ліквідність має така категорія пасивів, як депозити й поточні рахунки клієнтів, про що свідчить занадто висока питома вага таких пасивів у балансах українських комерційних банків. Залишки на поточних та депозитних клієнтських рахунках є найдешевшим, а тому основним, джерелом коштів, використовуваних банками для здійснення кредитної політики, чим сприяють максимізації банківського прибутку шляхом значного зменшення процентних витрат банку. Однак дана категорія банківських ресурсів характеризується високою рухливістю коштів і пов'язана з ризиком ліквідності більше, ніж будь-яка інша категорія, тому що строки вилучення коштів з розглянутих рахунків є невизначеними за своєю природою, отже,

визначення очікуваного залишку коштів на рахунках допоможе банкам якомога раціональніше використовувати це досить прибуткове джерело залучених коштів, мінімізуючи при цьому ризик ліквідності банку, а також встановити напрями депозитної політики банку .

Таким чином, проблема прогнозування передбачуваного залишку коштів на поточних і депозитних рахунках займає особливе місце в процесі керування банківською ліквідністю, дає можливість, незважаючи на високу рухливість коштів на таких рахунках, визначити постійний гарантований їхній залишок і використати його як стабільний кредитний ресурс.

Проблеми аналізу та моделювання депозитних ризиків досліджені у статті [1]. У статті зроблено короткий огляд праць у досліджуваному напрямку. Проведено аналіз причин виникнення депозитного ризику. Для вимірювання ризику автори розглядають найпростіші міри ризику, такі як математичне сподівання залишків, їх середньоквадратичне відхилення та коефіцієнт варіації. Проведено аналіз реальних залишків на депозитах у першому кварталі 2018 року.

1.2. Прогнозування залишків

Для побудови моделі поведінки залишків коштів на депозитних рахунках необхідно відповідним чином описати її. Для збору відомостей про досліджуваний процес у першу чергу важливе розуміння сутності самого процесу. Очевидно, що ринок депозитів, а отже, й попит на них існує й розвивається не сам по собі, а під впливом різних факторів, як внутрішніх, так і зовнішніх. Залишок на рахунках, що аналізуються, у зв'язку з високою мобільністю коштів не постійний і часто вкрай мінливий. Числове значення залишку на всіх балансових рахунках даної групи являє собою середньозважену суму залишків на деякій кількості рахунків у відповідний момент часу і є випадковою величиною, безперервною в інтервалі $(0; +\infty)$.

Згідно із законом великих чисел при деяких досить широких умовах (попарна незалежність випадкових величин, рівність їхнього математичного сподівання, рівномірна обмеженість дисперсії) сумарне поведження значної кількості випадкових величин практично стає закономірністю.

Керуючись цими міркуваннями, з метою збільшення якості вихідних даних для прогнозування, а отже, і поліпшення якості самого прогнозу необхідно зробити класифікацію клієнтів комерційного банку залежно від параметрів часового ряду, що описує поведження залишків на кожному конкретному рахунку шляхом розбивки клієнтів на групи за величиною їхнього математичного сподівання, що характеризує середню суму коштів, що зберігаються на рахунку в аналізований період часу. Потім кожна група розбивається на підгрупи залежно від величини дисперсії. Прогнозування передбачається здійснювати по кожній групі окремо, представляючи загальний результат сумою отриманих прогнозів.

Проблема прогнозування залишків розглядалася багатьма авторами [2,7, 10, 11], які використовували різні моделі. Так, у роботі Ольги Білик та Анатолія Сидоренка [2] запропонована стохастична модель, яка враховує історію денних приростів, тижневих та місячних приростів. Проведені дослідження дозволили встановити горизонт достовірного прогнозування.

У статті Орлова Ю.К. та Кір'яна О.М. [10] розглядається проблема прогнозування коштів на поточних та депозитних рахунках клієнтів банків з урахуванням макроекономічних показників. Для дослідження використано апарат нейронних мереж, розглянуто десять факторів впливу. Розглядалися місячні середньо-зважені залишки на депозитних рахунках у гривні протягом березня 2001 року до травня 2006 року. Прогнозування проводилося для 6-ти місяців і показало високу відносну точність – в межах 5%.

2. ЧАСОВІ РЯДИ

2.1. Основні положення

2.1.1. Означення

Часовий ряд, або ряд динаміки, динамічний ряд, це – послідовність випадкових величин Y_t , які, зазвичай, задано в деякі моменти часу $t \in N$. Часовий ряд є випадковою функцією (випадковим процесом) з дискретним аргументом та неперервною або дискретною шкалою значень.

Нехай $F_t(y)$ – функція розподілу випадкової величини Y_t , а $f_t(y)$ – щільність розподілу, $t \in N$.

Конкретну i -ту реалізацію часового ряду зазвичай позначають $y_t^{(i)}$. У нашій роботі, для спрощення запису, випадкові величини Y_t та їх реалізації $y_t^{(i)}$ будемо позначати однаково – y_t , коли зміст цієї величини є зрозумілим з контексту.

Для кожного перерізу часового ряду (значення для конкретного часового моменту) можна ввести стандартні числові характеристики: математичне сподівання, дисперсію, коваріації тощо.

Математичним сподіванням та дисперсією часового ряду називають такі послідовності:

$$M_t = E(Y_t) = \int_{-\infty}^{\infty} y f_t(y) dy, \quad D(Y_t) = \sigma_t^2 = \int_{-\infty}^{\infty} (y - M_t)^2 f_t(y) dy, \quad t \in N \quad (2.1)$$

Подібним чином означають коваріційні характеристики. Автоковаріацією випадкових величин Y_t та Y_{t-k} називають їх коваріацію:

$$\gamma_{kt} = E[(Y_t - M_t)(Y_{t-k} - M_{t-k})]. \quad (2.2)$$

Зауважимо, що

$$\gamma_{0t} = \sigma_t^2. \quad (2.3)$$

Часовий ряд називають стаціонарним, якщо для будь-яких інтервалів часу h_1, h_2, \dots, h_k сукупний розподіл випадкового вектора $(Y_{t+h_1}, Y_{t+h_2}, \dots, Y_{t+h_k})$ не залежить від часу t . У цьому випадку усі характеристики випадкового процесу не залежать від часу.

Умову стаціонарності перевірити на практиці доволі складно. Тому широке практичне використання отримав більш широкий клас часових рядів, так звані коваріаційно або слабо стаціонарні ряди, для яких умова незалежності від часу виконується лише для першого та других моментів.

Слабо стаціонарним (коваріаційно стаціонарним, стаціонарним у широкому сенсі) називають часовий ряд у якого математичне сподівання та коваріації не залежать від часу

$$\forall t \in N \quad M_t = M, \quad \gamma_{kt} = \gamma_k. \quad (2.4)$$

Коваріаційно стаціонарні або слабо стаціонарні часові ряди у літературі можуть називати просто стаціонарними, а стаціонарні – строго стаціонарними. Зрозуміло, що усі строго стаціонарні часові ряди є коваріаційно стаціонарними, обернене твердження – не вірне.

2.1.2. Ергодичність

Важливим поняттям є поняття ергодичності. Воно дозволяє за окремою реалізацією часового ряду знайти його моментні характеристики.

Насамперед означимо поняття часового середнього. Нехай $\{y_1, y_2, \dots, y_T\}$ – реалізація часового ряду на проміжку $[1, T]$. Часовим середнім називають таку величину:

$$\bar{y}_T = \frac{1}{T} \sum_{t=1}^T y_t. \quad (2.5)$$

Ергодичним за математичним сподіванням називають слабо стаціонарний часовий ряд, часове середнє якого прямує за ймовірністю при $T \rightarrow \infty$ до математичного сподівання:

$$\bar{y}_T \xrightarrow{P} M \Leftrightarrow \forall \varepsilon > 0 \lim_{T \rightarrow \infty} P(|\bar{y}_T - M| > \varepsilon) = 0. \quad (2.6)$$

Необхідні і достатні умови для виконання (2.6) визначає наступна теорема з праці [15].

Теорема 1.1. Для того щоб середнє по часу \bar{y}_T давало змістовну і незміщену оцінку математичного сподівання, необхідно і достатньо щоб виконувалася умова ергодичності:

$$\frac{1}{T} \sum_{k=0}^T \gamma_k \xrightarrow{T \rightarrow \infty} 0. \quad (2.7)$$

Легко встановити, що достатньою умовою виконання (2.7) є наступна умова прямування до нуля автоковаріації:

$$\lim_{k \rightarrow \infty} \gamma_k = 0. \quad (2.8)$$

2.1.3. Приклади часових рядів

Розглянемо деякі найпростіші часові ряди: дискретний білий шум, випадкове блукання та ряд з заданим трендом.

Часовий ряд, $t = 1, 2, \dots, T$ називають дискретним білим шумом, якщо випадкові величини ε_t незалежні, однаково розподілені і мають нульове математичне сподівання (ця умова не обов'язкова, оскільки її можна завжди досягнути центруванням випадкової величини ε_t):

$$E(\varepsilon_t) = 0, \quad \gamma_{kt} = \gamma_k = \delta_{k0} \sigma_\varepsilon^2, \quad \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \quad (2.9)$$

Якщо ε_t має нормальний розподіл, то такий шум називають дискретним білим гаусівським шумом.

Зрозуміло, що дискретний білий шум є стаціонарним випадковим процесом.

Розглянемо процес випадкового блукання, коли часовий ряд визначають так

$$Y_t = \varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_t, \quad (2.10)$$

де ε_t – дискретний білий шум (2.9).

Це нестационарний процес, оскільки дисперсія цього процесу не є сталою:

$$D(Y_t) = t\sigma^2.$$

Іншим прикладом нестационарного часового ряду є часовий ряд з трендом

$$y_t = a(t) + \varepsilon_t, \quad (2.11)$$

де $f(t) \neq const$ – детермінована функція часу, ε_t – дискретний білий шум.

Очевидно, що його математичне сподівання $E(y_t) = a(t)$ – залежить від часу, отже це – нестационарний процес.

2.1.4. Розрахунок характеристик часового ряду

Для статистичної оцінки характеристик часового ряду використовують спостереження за деякий час y_1, y_2, \dots, y_T .

За умови ергодичності оцінкою математичного сподівання є часове середнє, яке розраховують за формулою:

$$\hat{\mu}_T = \bar{y}_T = \frac{1}{T} \sum_{t=1}^T y_t. \quad (2.12)$$

Автоковаріації оцінюють так:

$$c_k = \frac{1}{T-k} \sum_{t=k+1}^T (y_t - \bar{y}_T)(y_{t-k} - \bar{y}_T). \quad (2.13)$$

На практиці найчастіше застосовують нормовані автоковаріації, які називають автокореляціями:

$$r_k = \frac{c_k}{c_0}. \quad (2.14)$$

Для аналізу автокореляцій r_k , $k = 1, 2, \dots$ будують корелограму – стовпцеву діаграму величин r_k . Корелограму використовують, зокрема, для візуальної перевірки достатньої умови ергодичності (2.8).

2.2 Моделювання та прогнозування часових рядів

2.2.1. Загальні міркування

Нехай задано часовий ряд y_t , $t = 1, 2, \dots$. Історією часового ряду називають множину значень y_t , відомих на момент часу τ :

$$H_\tau = \{y_\tau, y_{\tau-1}, \dots\}.$$

Моделюванням часового ряду є подання ряду у вигляді:

$$y_t = \Phi_t + E_t. \quad (2.15)$$

Складова Φ_t є власне моделлю, а E_t – залишком. Основна частина представлення (2.15) $\Phi_t = \Phi(t, \bar{\theta}, H_{t-1}, \dots)$ – це часовий ряд, який визначений аналітично чи алгоритмічно через детерміновані та випадкові величини, може залежати від часу t , деяких параметрів $\bar{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$, історії.

Основна мета моделювання – знайти найкращу у деякому сенсі модель Φ_t , яка, зокрема, дає "найменший" залишок E_t .

Залишок E_t повинен бути близьким до стаціонарного процесу і мати дисперсію значно меншу дисперсії досліджуваного процесу.

Ідеальним є наближення, коли залишок близький до дискретного білого гаусівського шуму $E_t \approx \varepsilon_t$ і наближено задовольняє наступні умови:

- 1) умову некорельованості;
- 2) має близьке до нуля середнє значення;
- 3) має сталу дисперсію, яка значно менша дисперсії вихідного ряду;
- 4) має нормальний розподіл.

В останньому випадку отримаємо подання

$$y_t = \Phi_t + \varepsilon_t. \quad (2.16)$$

Оцінку параметрів $\hat{\theta}$ знаходять стандартними статистичними методами найменших квадратів, максимальної правдоподібності, моментів та їх комбінаціями.

Отримане наближення часового ряду матиме вигляд:

$$y_t = \hat{y}_t + \varepsilon_t, \quad \hat{y}_t = \Phi(t, \hat{\theta}, H_{t-1}, \dots). \quad (2.17)$$

Перевірку умов 1).-3). легко провести для випадку нормального розподілу залишку. Для цього потрібно провести аналіз гістограми залишку.

Некорельованість залишків встановлюють по корелограмі. Якщо задатися рівнем значущості α , то автокореляції повинні бути у відрізьку [4]:

$$\left[-\frac{z_{1-\alpha/2}}{\sqrt{T}}, \frac{z_{1-\alpha/2}}{\sqrt{T}} \right], \quad (2.37)$$

де $z_{1-\alpha/2}$ – квантиль рівня $1-\alpha/2$ функції стандартного нормального розподілу.

Для тестуванню гіпотез про сталість математичного сподівання та дисперсії можна використати критерії Стюдента і Фішера, які застосовують за припущення про нормальність розподілу значень часового ряду.

На основі моделі (2.17) можна зробити прогноз часового ряду на момент часу $t+h$, як умовний часовий ряд:

$$y_{t+h|t} = \Phi(t+h, \hat{\theta}, H_t, \dots) + \varepsilon_{t+h}, \quad h \in N, h \geq 1. \quad (2.18)$$

Для практичних потреб важливими є точковий та інтервальний прогнози. За точковий прогноз приймають умовне математичне сподівання

$$\hat{y}_{t+h|t} \equiv E(y_{t+h} | H_t), \quad (2.19)$$

оскільки це забезпечує мінімальну похибку серед усіх можливих прогнозів [12].

Для побудови інтервального прогнозу, зокрема, розраховують умовне стандартне відхилення:

$$\hat{\sigma}_{t+h|t} = E(y_{t+h|t} - \hat{y}_{t+h|t})^2. \quad (2.20)$$

На його основі можна оцінити інтервал, якому належить прогнозована величина з високою ймовірністю.

Для оцінки точності точкового прогнозу використовують метод прогонки – порівняння прогнозованих та реальних даних. Як міру найчастіше використовують середню відносну похибку – Mean absolute percentage error (MAPE):

$$\bar{\delta} = 100\% \frac{1}{N} \sum_{t=1}^N \left| \frac{e_t}{y_t} \right|, \quad (2.21)$$

де y_t – фактичне значення показника, $e_t = y_t - \hat{y}_t$ – похибка прогнозу, – тривалість тесту.

Вважають, що точність моделі є високою, коли $\bar{\delta} < 10\%$, доброю – при $\bar{\delta} = 10-20\%$ і задовільною – при $\bar{\delta} = 10-50\%$.

Модифікацією цього показника є симетрична середня відносна абсолютна похибка – Symmetric mean absolute percentage error (sMAPE):

$$sMAPE = 100\% \frac{1}{n} \sum_{t=1}^N \frac{2|e_t|}{y_t + \hat{y}_t}. \quad (2.22)$$

Для вибору кращої моделі з деякого класу моделей зазвичай використовують інформаційні критерії.

Розглянемо інформаційний критерій Akaike (AIC) .

Нехай k – кількість параметрів моделі, n – загальна кількість спостережень. Припустимо, що похибки моделі $e_t = y_t - \hat{y}_t$ – незалежні та нормально розподілені. Значення критерію розраховують за формулою

$$AIC = 2k + n [\ln(2\pi RSS / N) + 1]$$

$RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2$ – сума квадратів похибок моделі.

Критерій враховує середньоквадратичну похибку наближення, а також кількість параметрів наближення. За кращу приймають модель з найменшим

значенням критерію АІС. Абсолютне значення АІС не має сенсу – він вказує тільки на відносний порядок порівнюваних моделей.

Процес моделювання та прогнозування часових рядів є складною математичною проблемою, який включає такі основні етапи [11]:

- 1) аналіз вихідного ряду та його початкова обробка;
- 2) вибір класу моделей для наближення ряду;
- 3) оцінка параметрів;
- 4) здійснення точкового та інтервального прогнозування;
- 5) перевірка адекватності моделі шляхом порівняння прогнозу та реальних даних;
- 6) вибір найкращої моделі.

На сьогодні, відомо багато різних методів моделювання та прогнозування часових рядів, серед яких виділимо наступні:

- 1) трендові моделі;
- 2) моделі на основі авторегресії;
- 3) моделі експонентного згладжування.

2.3. Трендові моделі

2.3.1. Виділення тренду

Точного математичного визначення тренду немає. Найчастіше тренд визначають як деяку детерміновану функцію, яка відображає основну тенденцію у поведінці часового ряду.

Для побудови тренду використовують два основні класи методів – параметричні та непараметричні. В першому випадку тренд знаходять у формі деякої параметризованої функції, параметри якої знаходять, наприклад, методом найменших квадратів. Непараметричні методи гуртуються на методах згладжування – ковзного середнього та зваженого ковзного середнього, експонентного згладжування, які також є загальними методами моделювання часових рядів.

Розглянемо найпростішу адитивну модель часового ряду з урахуванням тренду та сезонності [14]:

$$y_t = \tau_t + s_t + e_t, \quad (2.23)$$

де τ_t – тренд, s_t – сезонна складова, e_t – залишок.

Розрізняють три основні види трендів: тренд середнього, тренд дисперсії та тренд автоковаріації. Перший тренд можна візуально розпізнати, якщо часовий ряд має вигляд повільно змінюваної функції з адитивним шумом. Тренд дисперсії проявляється, якщо ряд має змінювану амплітуду коливань. Тренд автоковаріації спостерігати по графічному поданню ряду важко.

Для опису тренду найчастіше використовують поліноміальну, логарифмічну та експонентну апроксимації. В першому випадку:

$$\tau_t = a_0 + a_1 t + \dots + a_p t^p. \quad (2.24)$$

Для виділення періодичного тренду використовують наближення гармонійною функцією:

$$\tau_t = a \sin(\omega t) + b \cos(\omega t). \quad (2.25)$$

Параметр частоти ω часто можна оцінити емпірично, оцінку інших параметрів знаходять методом найменших квадратів.

Оцінку точності трендової моделі проводять за коефіцієнтом детермінації R^2 (R – квадрат):

$$R^2 = \frac{\delta_\tau^2}{\delta_y^2}, \quad \delta_\tau^2 = \frac{1}{T} \sum_{t=1}^T (\tau_t - \bar{y})^2, \quad \delta_y^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \bar{y})^2. \quad (2.26)$$

Коефіцієнт приймає значення від 0 до 1. Вважається, що чим ближче коефіцієнт до 1, тим кращою є модель.

2.3.2. Аналіз сезонності

Сезонність рідко є гармонійним циклічним процесом. Для її візуальної оцінки можна використати так звані сезонні графіки, коли сезонні ділянки часових рядів будують в одних координатах. На прояв сезонності також вказують періодичні піки корелограми.

Для аналітичного опису сезонності найчастіше використовують лінійне наближення.

Припустимо, що є h сезонів – $\{H_i, i=1,2,\dots,h\}$. Наприклад для помісячних даних параметр сезонності $h=12$ – кількість місяців, для щоденних даних з тижневою сезонністю – $h=7$.

Лінійне наближення сезонності матиме вигляд:

$$s_t = \lambda_1 \delta_{1t} + \lambda_2 \delta_{2t} + \dots + \lambda_h \delta_{ht}. \quad (2.27)$$

Величини δ_{it} відповідають h сезонам і визначаються так:

$$\delta_{it} = \begin{cases} 1, & t \in H_i \\ 0, & t \notin H_i \end{cases}. \quad (2.28)$$

Зазначимо, що виконується співвідношення:

$$\sum_{i=1}^h \delta_{it} = 1. \quad (2.29)$$

Це може привести до неоднозначного визначення параметрів $\lambda_1, \lambda_2, \dots, \lambda_h$.

Тому вводять обмеження:

$$\lambda_1 + \lambda_2 + \dots + \lambda_h = 0. \quad (2.30)$$

Якщо тренд наближати многочленом порядку m , то отримаємо представлення у формі лінійної регресії:

$$y_t = a_0 + a_1 t + \dots + a_p t^m + \lambda_1 \delta_{1t} + \lambda_2 \delta_{2t} + \dots + \lambda_h \delta_{ht} + \varepsilon_t, \quad (2.31)$$

з додатковою умовою (2.30).

Оцінку невідомих параметрів a_i та λ_j проводять методом найменших квадратів.

3. ЛІНІЙНЕ МОДЕЛЮВАННЯ ЧАСОВИХ РЯДІВ

3.1 Модель ковзного середнього

Нехай, як і раніше, ε_t – дискретний білий шум. Лінійним називають часовий ряд y_t , якщо його можна подати у вигляді лінійної комбінації з деякими вагами ($\psi_0 = 1, \psi_1, \psi_2, \dots$) через дискретний білий шум у попередні моменти часу:

$$y_t = \mu + \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}. \quad (3.1)$$

За умови абсолютної збіжності ряду ваг

$$\sum_{i=0}^{\infty} |\psi_i| < \infty, \quad (3.2)$$

часовий ряд (3.1) буде коваріаційно стаціонарним та ергодичним за математичним сподіванням [12].

За цих умов, легко знайти характеристики ряду:

$$\begin{aligned} E(y_t) &= \mu, \quad \sigma^2(y_t) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \psi_i^2, \\ \gamma_l &= E \left[\sum_{i,j=0}^{\infty} \psi_i \psi_j y_{t-i} y_{t-l-j} \right] = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+l}, \\ \rho_l &= \frac{\gamma_l}{\gamma_0} = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+l}}{\sum_{i=0}^{\infty} \psi_i^2}. \end{aligned} \quad (3.3)$$

Частковим випадком (3.1) є моделі ковзного середнього (mean average).

Модель $MA(q)$ визначається через значення білого шуму у минулому так

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}. \quad (3.4)$$

Нехай B – оператор зсуву на одиницю часу: $B y_t = y_{t-1}$. Тоді, ввівши оператор

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q, \quad (3.5)$$

подання моделі (3.4) можна переписати так:

$$y_t = \mu + \theta(B)\varepsilon_t. \quad (3.6)$$

Модель $MA(q)$ є скінченною лінійною комбінацією білого шуму, тому вона завжди стаціонарна. Знайдемо її характеристики:

$$E(y_t) = \mu, \quad \gamma_0 = (1 + \theta_1^2 + \dots + \theta_q^2)\sigma_\varepsilon^2,$$

$$\gamma_l = \left(-\theta_l + \sum_{i=1}^{q-l} \theta_i \theta_{l+i} \right) \sigma_\varepsilon^2, \quad 0 < l < q, \quad \lambda_l = 0, \quad l > q,$$

$$\rho_0 = 1, \quad \rho_l = \frac{-\theta_l + \sum_{i=1}^{q-l} \theta_i \theta_{l+i}}{1 + \theta_1^2 + \dots + \theta_q^2}, \quad 0 < l < q, \quad \rho_l = 0, \quad l > q$$

Зазначимо, якщо лаг більший порядку моделі ($l > q$), то відповідні автокореляції рівні нулю. Це важливо для практичного визначення порядку моделі.

3.2. Авторегресійна модель

Авторегресійна модель порядку $AR(p)$ записується так:

$$y_t = \varphi_0 + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t. \quad (3.7)$$

У цій моделі значення часового ряду у момент часу t виражаються через значення у попередні моменти.

Розглянемо модель авторегресії першого порядку

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varepsilon_t. \quad (3.8)$$

Для математичного сподівання отримаємо

$$E(y_t) = \varphi_0 + \varphi_1 E(y_{t-1}),$$

далі з умови стаціонарності $E(y_t) = E(y_{t-1}) = \mu$, отримаємо вираз для математичного сподівання:

$$\mu = \frac{\varphi_0}{1 - \varphi_1}, \quad \varphi_1 \neq 1.$$

Використовуючи рекурентно рівняння (3.6) отримаємо:

$$y_t - \mu = \varphi_1 (y_{t-1} - \mu) + \varepsilon_t = \varphi_1 (\varphi_1 (y_{t-2} - \mu) + \varepsilon_{t-1}) + \varepsilon_t = \sum_{i=0}^{\infty} \varphi_1^i \varepsilon_{t-i}.$$

На основі останньої рівності отримаємо:

$$\text{Cov}[y_t, \varepsilon_{t+1}] = E[(y_t - \mu)\varepsilon_{t+1}] = E\left[\left(\sum_{i=0}^{\infty} \varphi_1^i \varepsilon_{t-i}\right)\varepsilon_{t+1}\right] = 0,$$

$$\text{Var}(y_t) = \varphi_1^2 \text{Var}(y_{t-1}) + \sigma_\varepsilon^2, \quad \gamma_0 = \text{Var}(y_t) = \frac{\sigma_\varepsilon^2}{1 - \varphi_1^2}.$$

З останньої формули встановлюємо нерівність $|\varphi_1| < 1$, яка є необхідною і достатньою умовою стаціонарності.

Проведемо дослідження загальної моделі $AR(p)$.

Для цієї моделі просто знайти:

$$\mu = \frac{\varphi_0}{1 - \varphi_1 - \varphi_2 - \dots - \varphi_p}, \quad \text{де } \varphi_1 + \varphi_2 + \dots + \varphi_p \neq 1.$$

$$\gamma_l = \varphi_1 \gamma_{l-1} + \dots + \varphi_p \gamma_{l-p},$$

$$\gamma_0 = \varphi_1 \gamma_1 + \dots + \varphi_p \gamma_p + \sigma_\varepsilon^2,$$

$$\rho_l = \varphi_1 \rho_{l-1} + \dots + \varphi_p \rho_{l-p}.$$

З використанням оператора зсуву і введенням оператора

$$\varphi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p \tag{3.9}$$

загальну модель $AR(p)$ можна записати так:

$$\varphi(B) y_t = \omega_t, \tag{3.10}$$

де $\omega_t = \varphi_0 + \varepsilon_t$.

Нехай λ_i , $i = 1, \dots, p$ є коренями рівняння:

$$\lambda^p - \varphi_1 \lambda^{p-1} - \dots - \varphi_{p-1} \lambda - \varphi_p = 0, \quad \text{тоді оператор (3.9) можна записати так}$$

$$\varphi(B) = (1 - \lambda_1 B) \dots (1 - \lambda_p B).$$

Якщо, $|\lambda_i| < 1$, $i = 1, \dots, p$ то отримаємо стаціонарне подання:

$$y_t = (1 - \lambda_1 B)^{-1} \dots (1 - \lambda_p B)^{-1} \omega_t,$$

яке можна переписати так

$$y_t = \sum_{i=0}^{\infty} (a_1 \lambda_1^i + \dots + a_p \lambda_p^i) \omega_{t-i}, \quad (3.11)$$

де

$$a_i = \frac{\lambda_i^{p-1}}{\prod_{\substack{1 \leq j \leq p \\ j \neq i}} (\lambda_i - \lambda_j)}.$$

Враховуючи, що $\omega_t = \varphi_0 + \varepsilon_t$, з остання рівність є поданням $AR(p)$ у формі $MA(\infty)$.

3.3. Модель ARMA

Модель $ARMA(p, q)$ записують у вигляді:

$$y_t = \varphi_0 + \sum_{i=1}^p \varphi_i y_{t-i} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j}. \quad (3.12)$$

Вона поєднує моделі авторегресії та ковзного середнього, тому є стаціонарною при виконанні необхідних умов.

Математичне сподівання легко визначити з (3.12):

$$\mu = \frac{\varphi_0}{1 - \varphi_1 - \varphi_2 - \dots - \varphi_p}.$$

Через введені раніше оператори (3.5) та (3.9) модель запишеться так:

$$\varphi(B) y_t = \varphi_0 + \theta(B) \varepsilon_t. \quad (3.13)$$

Зауважимо, дія згадані оператори є лінійними, а їх дія на сталу дає сталу, наприклад:

$$\varphi(B)c = (1 - \varphi_1 - \dots - \varphi_p)c = \varphi(1)c.$$

Позначимо:

$$\frac{\varphi(B)}{\theta(B)} = 1 - \pi_1 B - \pi_2 B^2 - \dots = \pi(B),$$

$$\frac{\theta(B)}{\varphi(B)} = 1 + \psi_1 B + \psi_2 B^2 + \dots = \psi(B).$$

Тоді легко записати представлення $ARMA(p, q)$ у формі моделі $AR(\infty)$ або $MA(\infty)$:

$$y_t = \frac{\varphi_0}{1 - \theta_1 - \dots - \theta_q} + \pi_1 y_{t-1} + \pi_2 y_{t-2} + \dots + \varepsilon_t, \quad (3.14)$$

$$y_t = \mu + \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \dots = \mu + \psi(B) \varepsilon_t. \quad (3.15)$$

З останнього подання легко бачити:

$$Var(y_t) = (1 + \psi_1^2 + \psi_2^2 + \dots) \sigma_\varepsilon^2.$$

Характеристичні корені моделі $ARMA(p, q)$ відповідають кореням $AR(p)$. Тому, умовою стаціонарності $ARMA(p, q)$ є аналогічні нерівності $|\lambda_i| < 1, i = 1, \dots, p$.

Порядок моделі (p, q) встановлюють шляхом аналізу вибіркової автокореляційної функцію (ACF) та часткової автокореляційної функції (PACF). Коефіцієнти ACF оцінюють за формулою (2.14). Вони повинні бути спадними після r_p . Коефіцієнти PACF r_{kk} є статистичними оцінками МНК коефіцієнту φ_k для моделі $AR(k)$ $r_{kk} = \hat{\varphi}_k$. Вони повинні спадати після r_{qq} .

Після встановлення порядку моделі оцінювання параметрів проводять методом максимальної правдоподібності [12].

Перейдемо до розгляду питання про прогнозування за моделлю $ARMA(p, q)$.

Нехай, на момент часу T відома історія $H_T = \{y_T, y_{T-1}, \dots\}$. Як вже зазначалося у п. 2.2.1 найкращий прогноз на момент $t = T + h$ дає умовне математичне сподівання:

$$\hat{y}_{T+h|T} = E(y_{T+h} | H_T).$$

Тому, для прогнозу на один період отримаємо:

$$\hat{y}_{T+1|T} = \hat{y}_{T+1} = E(y_{T+1} | H_T) = \varphi_0 + \sum_{i=1}^p \varphi_i y_{T+1-i}.$$

$$y_{T+1} = \varphi_0 + \sum_{i=1}^p \varphi_i y_{T+1-i} - \sum_{j=1}^q \theta_j \varepsilon_{T+1-j}$$

Похибка буде такою:

$$e_{T+1} = y_{T+1} - \hat{y}_{T+1|T} = \varepsilon_{T+1}, \text{Var}(e_{T+1}) = \text{Var}(\varepsilon_t) = \sigma^2(\varepsilon_t) = \sigma_\varepsilon^2.$$

Прогноз на h кроків вперед проводять шляхом зсуву на відповідну кількість кроків:

$$\hat{y}_{T+h|T} = E(y_{T+h} | H_T) = \varphi_0 + \sum_{i=1}^p \varphi_i \bar{y}_{T+h-i} - \sum_{j=1}^q \theta_j \bar{\varepsilon}_{T+h-j}, \quad (3.16)$$

$$\text{де позначено } \bar{y}_{T+k} = \begin{cases} y_{T+k}, & k \leq 0 \\ \hat{y}_{T+k|T}, & k > 0 \end{cases}, \text{ і } \bar{\varepsilon}_{T+k} = \begin{cases} \varepsilon_{T+k}, & k \leq 0 \\ 0, & k > 0 \end{cases}.$$

Похибку отримують з подання $MA(\infty)$ для моделі (3.15):

$$\hat{y}_{T+h|T} = \mu + \varepsilon_{T+h} + \psi_1 \varepsilon_{T+h-1} + \psi_2 \varepsilon_{T+h-2} + \dots +, \quad (3.17)$$

Тому для похибки матимемо:

$$e_{T+h|T} = \varepsilon_{T+h} + \psi_1 \varepsilon_{T+h-1} + \dots + \psi_{h-1} \varepsilon_{T+1}, \quad (3.18)$$

$$\text{Var}(e_{T+h|T}) = \sigma_{T+h|T}^2 = \left(1 + \psi_1^2 + \dots + \psi_{h-1}^2\right) \sigma^2(\varepsilon_t).$$

Припускаючи нормальний розподіл похибок, дають інтервальний прогноз:

$$\left[\hat{y}_{T+h|T} - z_\alpha \hat{\sigma}_{T+h|T}, \hat{y}_{T+h|T} + z_\alpha \hat{\sigma}_{T+h|T} \right], \quad z_\alpha = N^{-1}(\alpha). \quad (3.19)$$

3.4 Лінійні нестационарні моделі

3.4.1 Модель ARIMA

Як вже зазначалося, достатньою умовою стаціонарності моделі $ARMA(p, q)$ є умова коренів:

$$|\lambda_i| < 1, \quad i = 1, \dots, p.$$

Припустимо, що ця умова виконується лише для деяких коренів, а інші d рівні по модулю один. Тоді можна записати:

$$\varphi(B) = \varphi^*(B)(1-B)^d = \varphi^*(B)\Delta^d, \quad (3.20)$$

де, $\varphi^*(B)$ – стаціонарний оператор авторегресії, $\Delta = 1 - B$.

Вихідну нестационарну модель, припускаючи для спрощення $\varphi_0 = 0$, можна подати так:

$$\varphi^*(B)\Delta^d y_t = \theta(B)\varepsilon_t, \quad (3.21)$$

або через допоміжний процес:

$$\begin{aligned} \varphi^*(B)x_t &= \theta(B)\varepsilon_t, \\ x_t &= \Delta^d y_t. \end{aligned}$$

Покажемо, що процес $x_t = \Delta^d y_t$ є оборотним. Нехай S є таким оператором:

$$S = (1 + B + B^2 + \dots) = (1 - B)^{-1} = \Delta^{-1}.$$

Тому можна записати формулу відновлення

$$y_t = S^d x^t.$$

Тому, нестационарний процес (3.21) можна отримати через стаціонарний процес $\varphi^*(B)x_t = \theta(B)\varepsilon_t$.

Побудоване таким способом узагальнення моделі $ARMA(p, q)$ називають авторегресійною інтегрованою моделлю ковзного середнього $ARIMA(p, d, q)$.

На практиці відновлення ряду за його приростами здійснюють на основі наступних рекурентних співвідношень:

$$\Delta^{i-1} y_{t+k} = \Delta^{i-1} y_{t+k-1} + \Delta^i y_{t+k}, \quad k = 1, 2, \dots, h, \quad i = d, d-1, \dots, 1. \quad (3.22)$$

Використання скінчених різниць дозволяє видалити поліноміальний тренд, оскільки для поліному порядку m скінченні різниці порядку $d = m$ є сталими, а вищих порядків – нульовими. На основі цього на практиці знаходять порядок моделі d – з умови стабілізації відповідних скінчених різниць.

3.4.2 Модель SARIMA

За допомогою скінчених різниць можна також врахувати сезонність. Для цього означимо різницевий оператор з лагом s :

$$\Delta_s y_t = (1 - B^s) y_t = y_t - y_{t-s}, \quad B^s y_t = y_{t-s}. \quad (3.23)$$

Модель з мультиплікативним врахування сезонності без нульового члена формулюють в операторній формі так [14] :

$$\varphi(B) \Phi(B^s) (1 - B)^d (1 - B^s)^D y_t = \theta(B) \Theta(B^s) \varepsilon_t, \quad (3.24)$$

де використано оператор зсуву на параметр сезонності (3.23):

$$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} \dots - \Phi_P B^{Ps},$$

$$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} \dots - \Theta_Q B^{Qs}.$$

Отже, ця модель описується параметрами:

$$d; \varphi_1, \varphi_2, \dots, \varphi_p; \theta_1, \theta_2, \dots, \theta_q;$$

$$D; \Phi_1, \Phi_2, \dots, \Phi_P; \Theta_1, \Theta_2, \dots, \Theta_Q.$$

Тому модель SARIMA формально записують ARIMA(p, d, q)(P, D, Q)[s].

4. ПРАКТИЧНА ІМПЛЕМЕНТАЦІЯ МОДЕЛІ SARIMA

4.1 Загальна схема аналізу та прогнозування

Загальна схема дослідження часового ряду з використанням моделі SARIMA аналогічна дослідженню часових рядів ARIMA за схемою Бокса-Дженкінса. Коротко її можна подати так:

- 1) Побудуйте ряд і шукайте можливі викиди.
- 2) Стабілізуйте дисперсію шляхом перетворення даних (Box-Cox).
- 3) Проаналізуйте стаціонарність перетвореного ряду.
- 4) Якщо ряд не стаціонарний, то ми використовуємо диференціювання.
- 5) Визначте сезонну модель, проаналізувавши сезонні коефіцієнти ACF та PACF
- 6) Визначте регулярний компонент, досліджуючи ACF та PACF залишків сезонної моделі.
- 7) Перевірте значущість коефіцієнтів.
- 8) Проаналізуйте залишки:
- 9) Порівняйте різні моделі за допомогою інформаційних критеріїв, наприклад AIC.

В мережі Інтернет можна знайти цілий ряд джерел з викладом приведеної схеми і прикладами її реалізації на мовах Python та R. Вкажемо на деякі з них:

<https://russianblogs.com/article/60511336750/>

<https://www.8host.com/blog/prognozirovanie-vremennyx-ryadov-s-pomoshhyu-arima-v-python-3/>

<https://ru.coursera.org/lecture/data-analysis-applications/primier-postroeniia-proghnoza-rnCnt>

<https://youtu.be/bJ8PUaQ-WAg>

<https://github.com/mariocastro73/ML2020-2021/blob/master/scripts/SARIMA-examples.R>

<https://otexts.com/fpp2/>

4.2 Аналіз та прогнозування залишків на рахунках клієнтів банку

Досліджували залишки на рахунках клієнтів (ЗРК) банку АТ "Кредобанк". Розглядали дані з 04.01.2016 по 13.12.2019. На рис. 4.1 показана динаміка ЗРК по робочих днях відносно початкового моменту

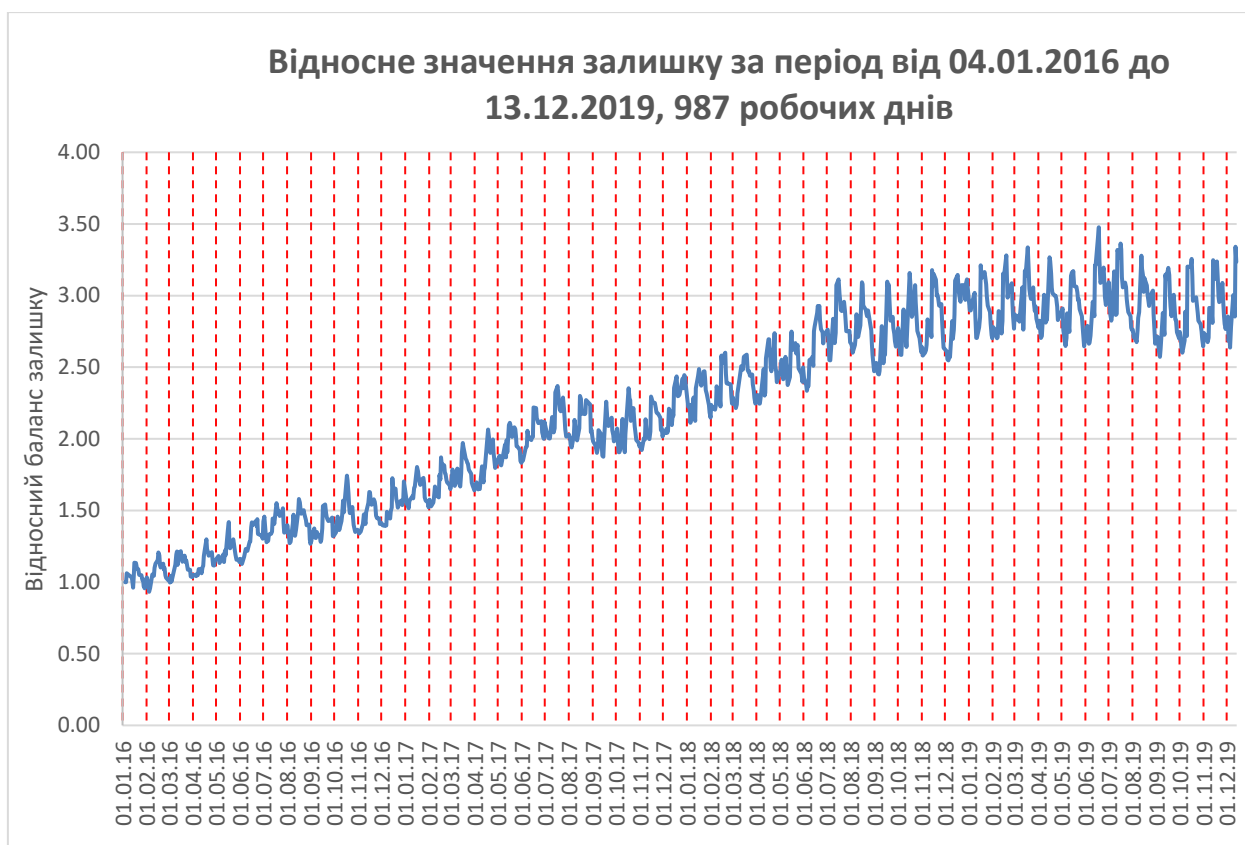


Рис.4.1. Динаміка ЗРК з 04.01.2016 по 13.12.2019.

Вивчаючи графік легко встановити, що:

1) спостерігається місячна сезонність;

2) часовий ряд має дві яскраво виражені ділянки – ділянка 1 зростання з лінійним трендом (04.01.2016 – 30.09.2018) та ділянка 2 стабільної поведінки без тренду (01.10.2018 – 13.12.2019).

Спочатку розглянемо ділянку 2. Відносний баланс ЗРК показаний на Рис.4.2.

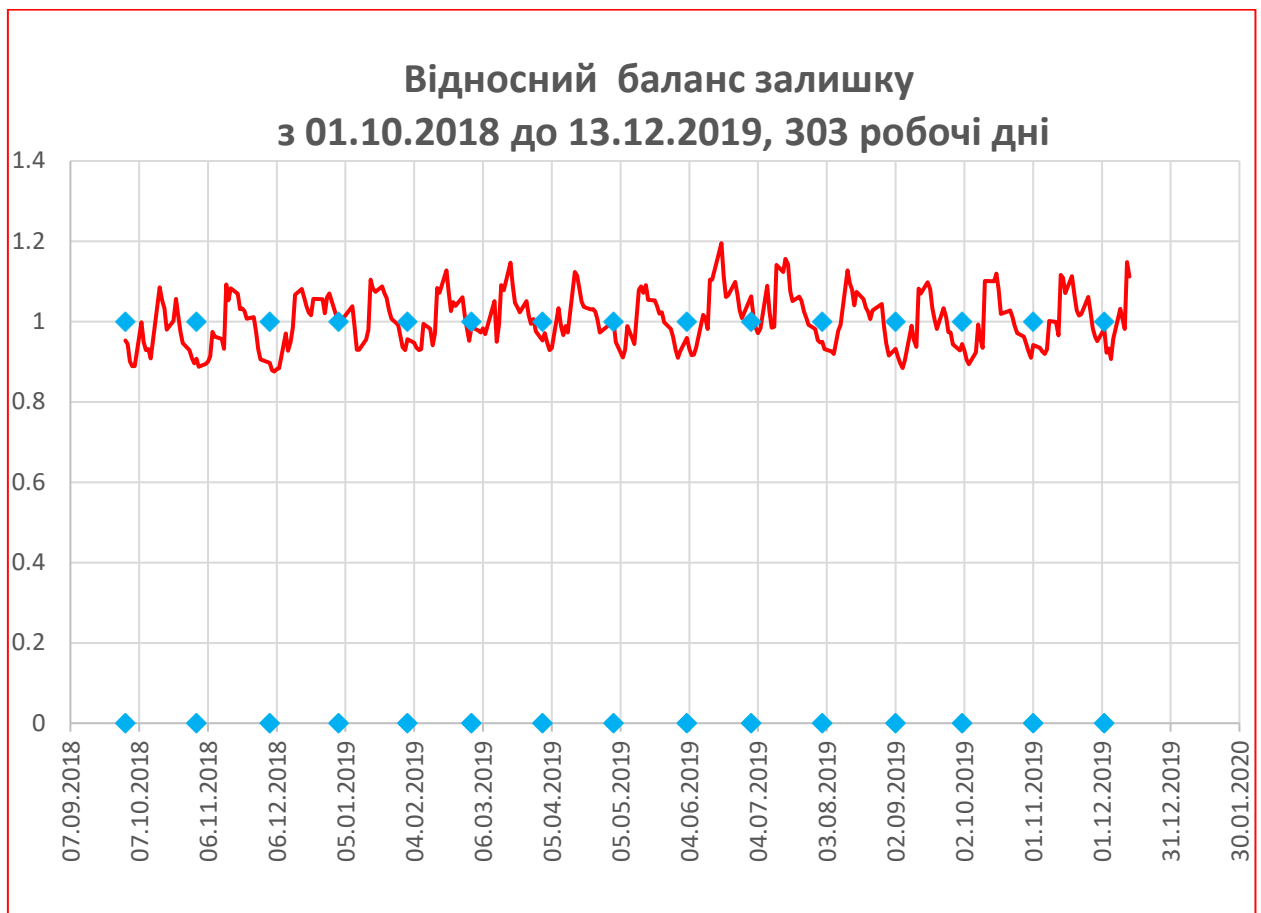


Рис.4.2. Динаміка ЗРК, 01.10.2018 – 13.12.2019.

Аналіз підтверджує по-місячну сезонність. Однак, кількість робочих днів у кожному місяці є різною, а середня за рік – 21 день. Для чіткого опису сезонності було проведено коригування даних до 21 робочих днів на місяць, шляхом усереднення за сусідніми днями.

Сезонність підтверджує STL декомпозиція – набір евристик для виділення тренду та сезонності (рис. 4.3).

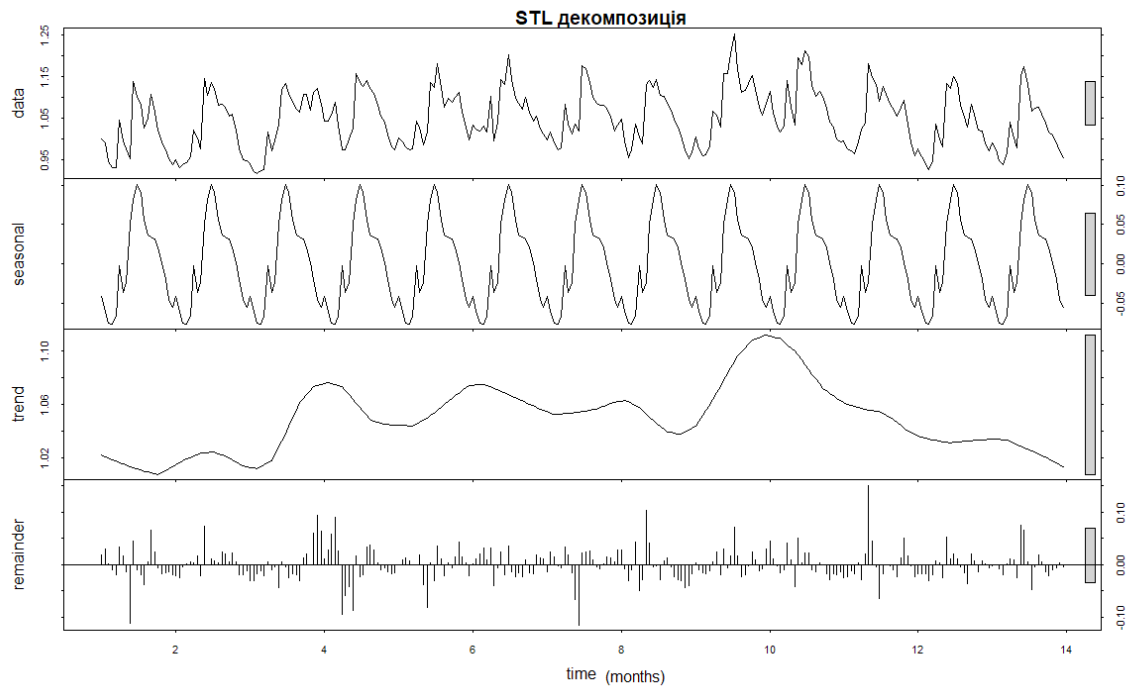


Рис.4.3. Сезонна декомпозиція

Сезонність також можна підтвердити візуально шляхом побудови корелограми, аналізом графіків за сезонами та сезонних серій.

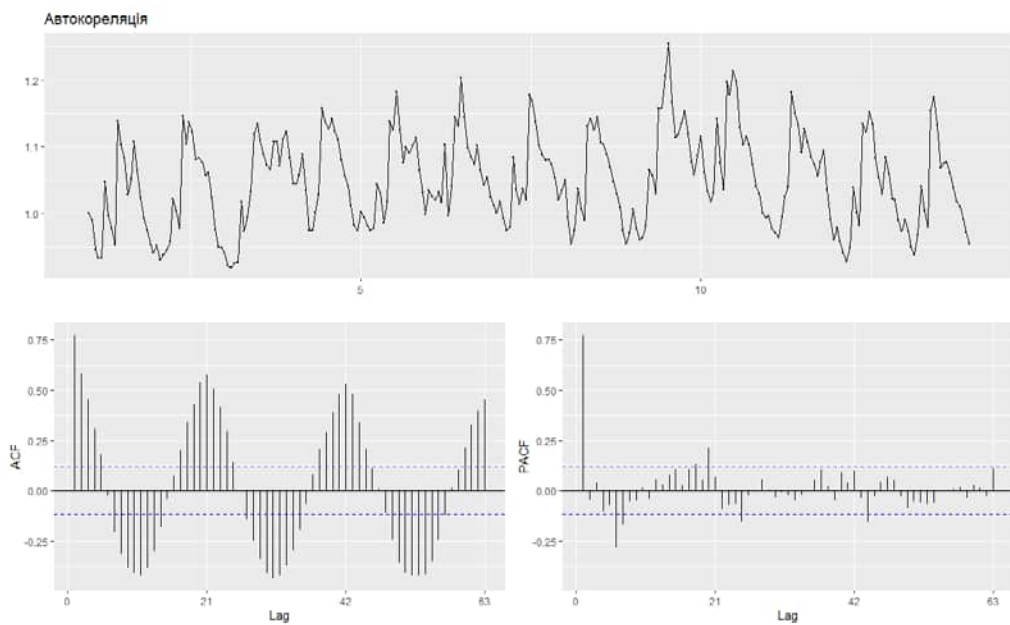


Рис.4.4. Автокореляційна та частково автокореляційна функції

На рис.4.4. показана корелограма вихідного часового ряду. Функція АСФ має максимуми з лагом 21, що говорить про місячну сезонність (за робочими днями).

Нижче показані значення ряду по сезонах в одних координатах – днях місяця (рис.4.5.), а також графіки серій, що відповідають окремим сезонам (рис.4.6.).

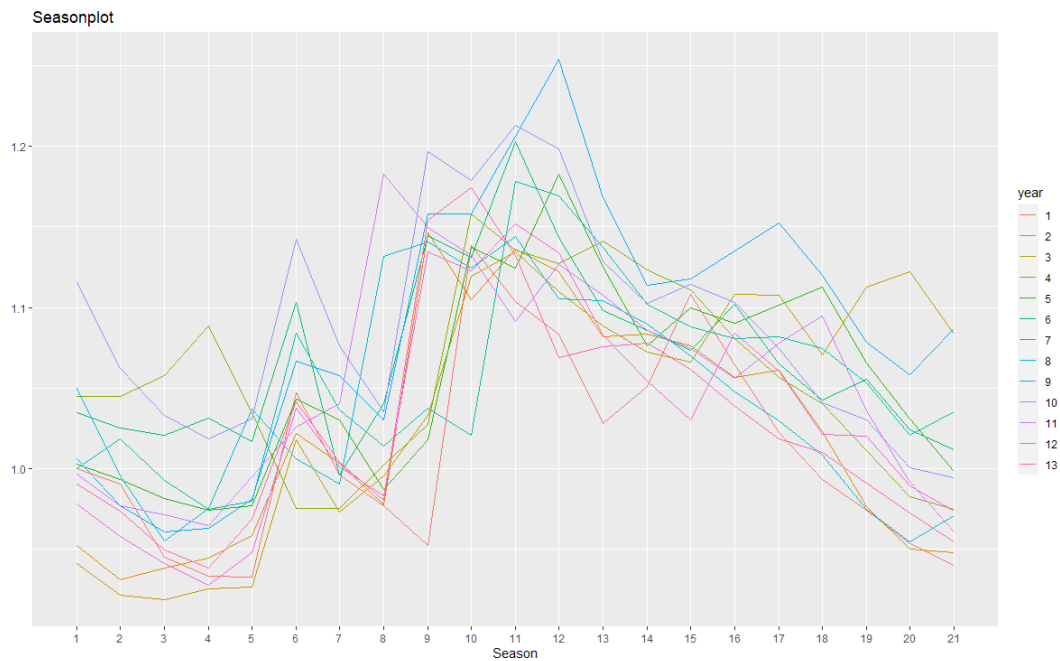


Рис.4.5. По-сезонний графік

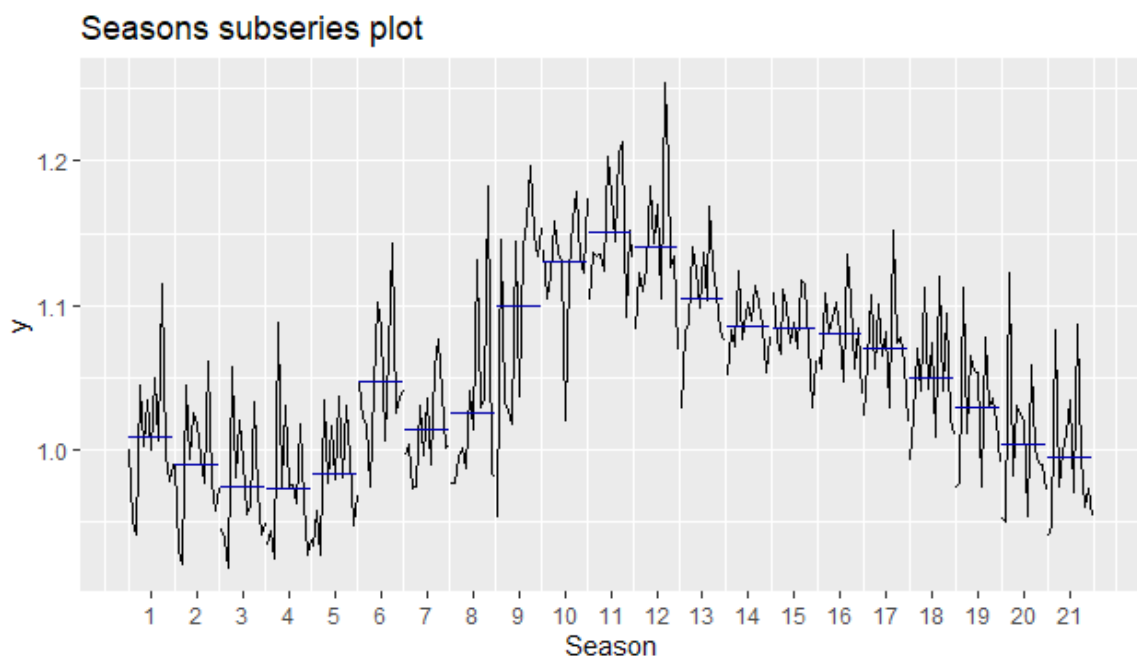


Рис.4.6. Графік сезонних серій

Проведений графічний аналіз беззаперечно показує місячну сезонність.

Відповідно до методики Бокса-Дженкінса [4] проведемо диференціювання з лагом $s = 21$. Отримані результати показано на рис. 4.7.



Рис.4.7. Результат сезонного диференціювання вихідного ряду

$$\Delta_s y_t = y_t - y_{t-s}$$

Отриманий ряд по вигляду є близьким до білого шуму. Проведемо тести на його стаціонарність.

Використаємо тест KPSS (Кв'ятковського – Філіпса – Шмідта – Шіна), нульовою гіпотезою якого є тренд-стаціонарність:

```
#####
# KPSS Unit Root Test #
#####
Test is of type: mu with 5 lags.
Value of test-statistic is: 0.2609
Critical value for a significance level of:
          10pct  5pct  2.5pct  1pct
critical values 0.347 0.463  0.574 0.739
```

Легко бачити, що немає підстав відкидати нульову гіпотезу, оскільки тестове значення статистики менше крайнього лівого рівня значущості.

Проведемо також тест ADF (Augmented Dickey-Fuller Test), нульовою гіпотезою якого є наявність одиничного кореня. Отримали такий результат:

```
Augmented Dickey-Fuller Test
data: y.sdifff
Dickey-Fuller = -4.1699, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
Warning message:
In adf.test(y.sdifff) : p-value smaller than printed p-value
```

Отримане значення p-value є менше 0.01, тому є підстави відхилити нульову гіпотезу про наявність одиничного кореня.

Тести підтверджують візуальний висновок про стаціонарність перших сезонних різниць. Тому дальше диференціювання – не потрібне.

Отже, ми встановили параметри диференціювання моделі SARIMA: $d = 0, D = 1$.

Далі потрібно визначити параметри p, q, P, Q . Це найпростіше зробити використовуючи процедуру `auto.arima`. Ця процедура перебирає різні значення параметрів p, q, P, Q і розраховує

Отримали результат:

```
Series: dep_ts
ARIMA(1, 0, 2)(2, 1, 1)[21]
```

Coefficients:

	ar1	ma1	ma2	sar1	sar2	sma1
	0.9465	-0.4134	-0.2464	-0.3347	-0.2125	-0.6026
s.e.	0.0325	0.0727	0.0695	0.1132	0.0978	0.1158

```
sigma^2 estimated as 0.001218: log likelihood=479.21
AIC=-944.43 AICc=-943.97 BIC=-919.72
```


Отже шуканою моделлю а модель ARIMA(1,0,2)(2,1,1)[21].

З використанням цієї моделі побудували прогноз на місяць (h=21):

Poi nt	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
14. 00000	0. 9787196	0. 9339964	1. 0234428	0. 9103214	1. 047118
14. 04762	0. 9582183	0. 9075372	1. 0088995	0. 8807082	1. 035728
14. 09524	0. 9429924	0. 8910129	0. 9949718	0. 8634967	1. 022488
14. 14286	0. 9384792	0. 8853635	0. 9915948	0. 8572458	1. 019713
14. 19048	0. 9625364	0. 9084230	1. 0166497	0. 8797771	1. 045296
14. 23810	1. 0289551	0. 9739632	1. 0839470	0. 9448523	1. 113058
14. 28571	1. 0070133	0. 9512461	1. 0627804	0. 9217247	1. 092302
14. 33333	1. 0466748	0. 9902220	1. 1031275	0. 9603377	1. 133012
14. 38095	1. 1231654	1. 0661054	1. 1802254	1. 0358997	1. 210431
14. 42857	1. 1167970	1. 0591984	1. 1743956	1. 0287076	1. 204886
14. 47619	1. 1285870	1. 0705101	1. 1866639	1. 0397661	1. 217408
14. 52381	1. 1324894	1. 0739873	1. 1909914	1. 0430182	1. 221961
14. 57143	1. 0935719	1. 0346915	1. 1524523	1. 0035221	1. 183622
14. 61905	1. 0687496	1. 0095324	1. 1279669	0. 9781846	1. 159315
14. 66667	1. 0622323	1. 0027148	1. 1217498	0. 9712081	1. 153256
14. 71429	1. 0630278	1. 0032426	1. 1228130	0. 9715942	1. 154461
14. 76190	1. 0601628	1. 0001387	1. 1201868	0. 9683640	1. 151962
14. 80952	1. 0467091	0. 9864719	1. 1069462	0. 9545843	1. 138834
14. 85714	1. 0174193	0. 9569918	1. 0778468	0. 9250034	1. 109835
14. 90476	0. 9863919	0. 9257944	1. 0469894	0. 8937160	1. 079068
14. 95238	0. 9764607	0. 9157113	1. 0372102	0. 8835525	1. 069369

Графік прогнозу показано на

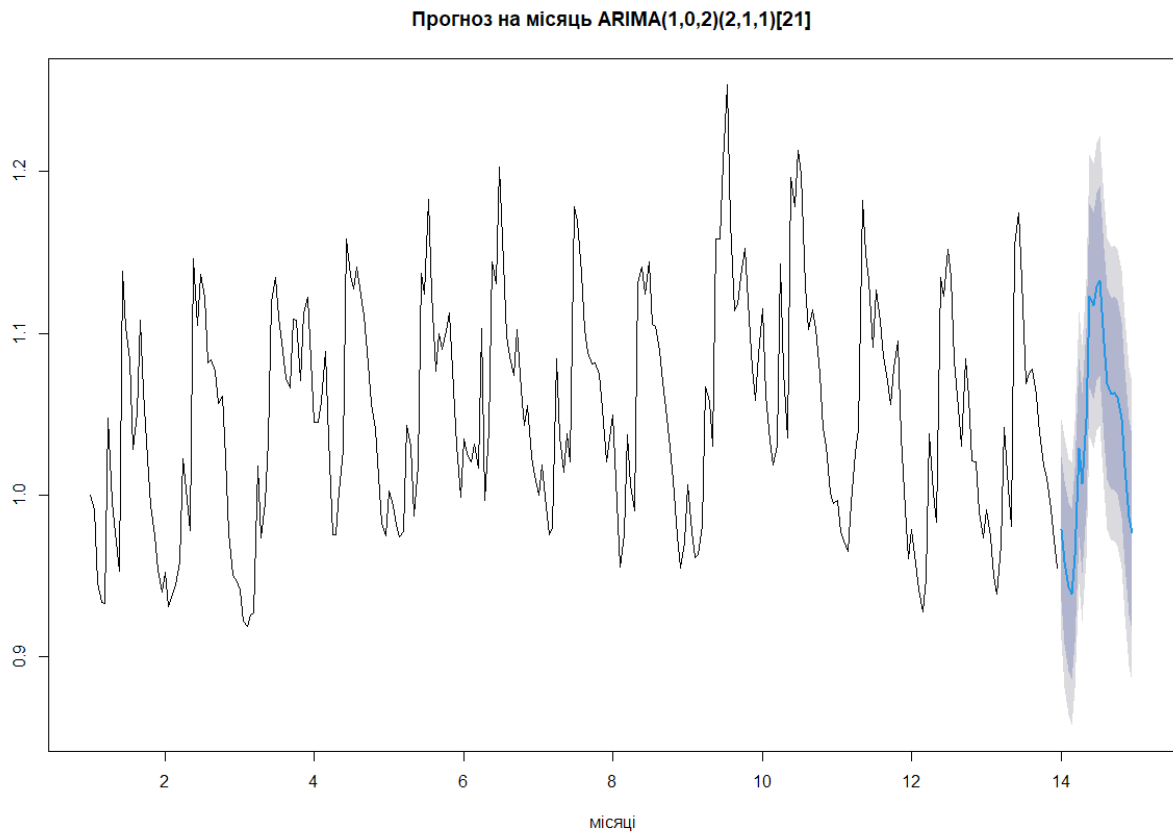


Рис.4.8. Прогноз за моделлю ARIMA(1,0,2)(2,1,1)[21] на один місяць.

На наступному рисунку деталізовано показано порівняння прогнозу та реальних даних

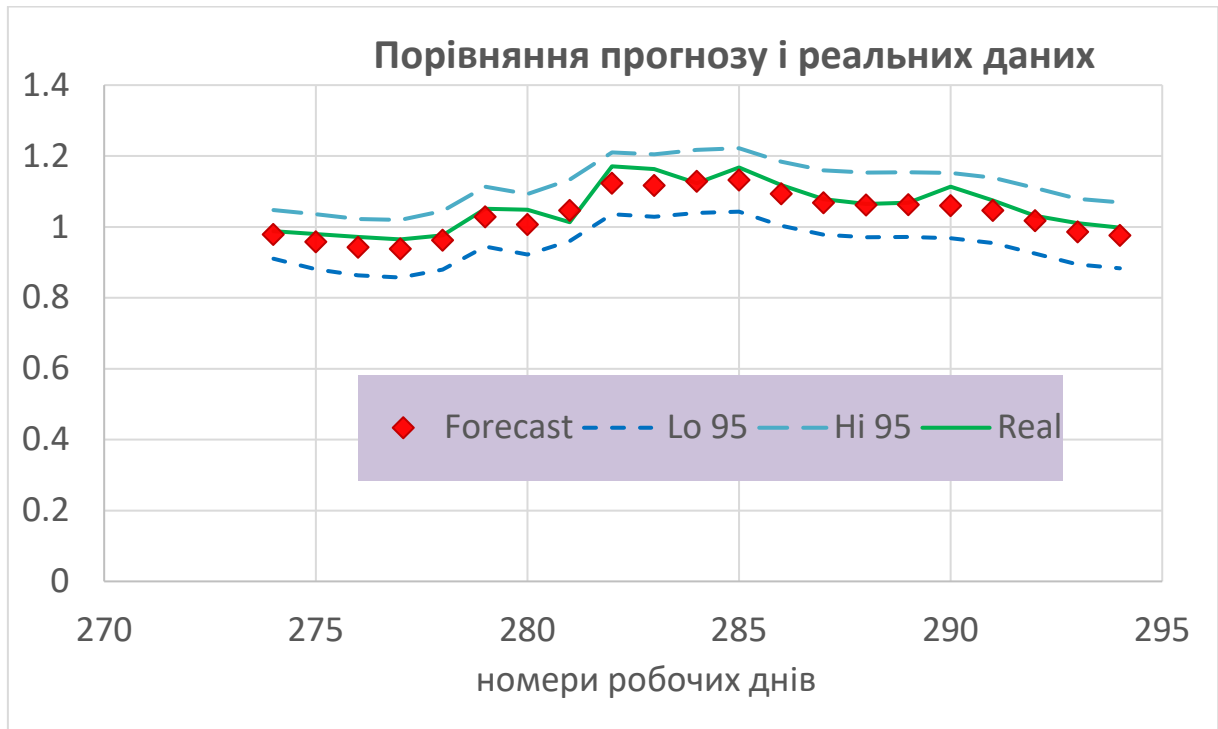


Рис.4.9. Похибка прогнозу

Для оцінки точності прогнозу використали показник MAPE (2.21). Його значення дорівнює $MAPE = 0.023$, що вказує на високу якість прогнозу.

Аналогічні дослідження проведено для ділянка 1 зростання з лінійним трендом (04.01.2016 – 30.09.2018). Отримано модель $ARIMA(0,1,3)(0,0,2)[21]$:

Point	Forecast	Lo 80	Hi 80	Lo 95	Hi 95
14. 90476	1. 809970	1. 744040	1. 875901	1. 709138	1. 910802
14. 95238	1. 793408	1. 715936	1. 870880	1. 674925	1. 911892
15. 00000	1. 837020	1. 755790	1. 918249	1. 712789	1. 961250
15. 04762	1. 798824	1. 717102	1. 880547	1. 673841	1. 923808
15. 09524	1. 777708	1. 695095	1. 860322	1. 651362	1. 904055
15. 14286	1. 788574	1. 702710	1. 874437	1. 657257	1. 919890
15. 19048	1. 821485	1. 732924	1. 910046	1. 686043	1. 956927
15. 23810	1. 850179	1. 759365	1. 940994	1. 711291	1. 989068
15. 28571	1. 815951	1. 723243	1. 908658	1. 674167	1. 957734
15. 33333	1. 846529	1. 752226	1. 940832	1. 702304	1. 990754
15. 38095	1. 874066	1. 778412	1. 969720	1. 727776	2. 020356
15. 42857	1. 985079	1. 888280	2. 081877	1. 837038	2. 133119
15. 47619	1. 999379	1. 901607	2. 097151	1. 849849	2. 148908
15. 52381	2. 000856	1. 902256	2. 099457	1. 850060	2. 151653
15. 57143	1. 968190	1. 868883	2. 067497	1. 816313	2. 120067
15. 61905	1. 944915	1. 845005	2. 044825	1. 792116	2. 097715
15. 66667	1. 933546	1. 833121	2. 033972	1. 779959	2. 087134
15. 71429	1. 910859	1. 809993	2. 011725	1. 756598	2. 065120
15. 76190	1. 875991	1. 774748	1. 977234	1. 721153	2. 030829
15. 80952	1. 840709	1. 739143	1. 942276	1. 685378	1. 996041
15. 85714	1. 821242	1. 719399	1. 923084	1. 665487	1. 976997

Графік прогнозу подано на Рис. 4.10

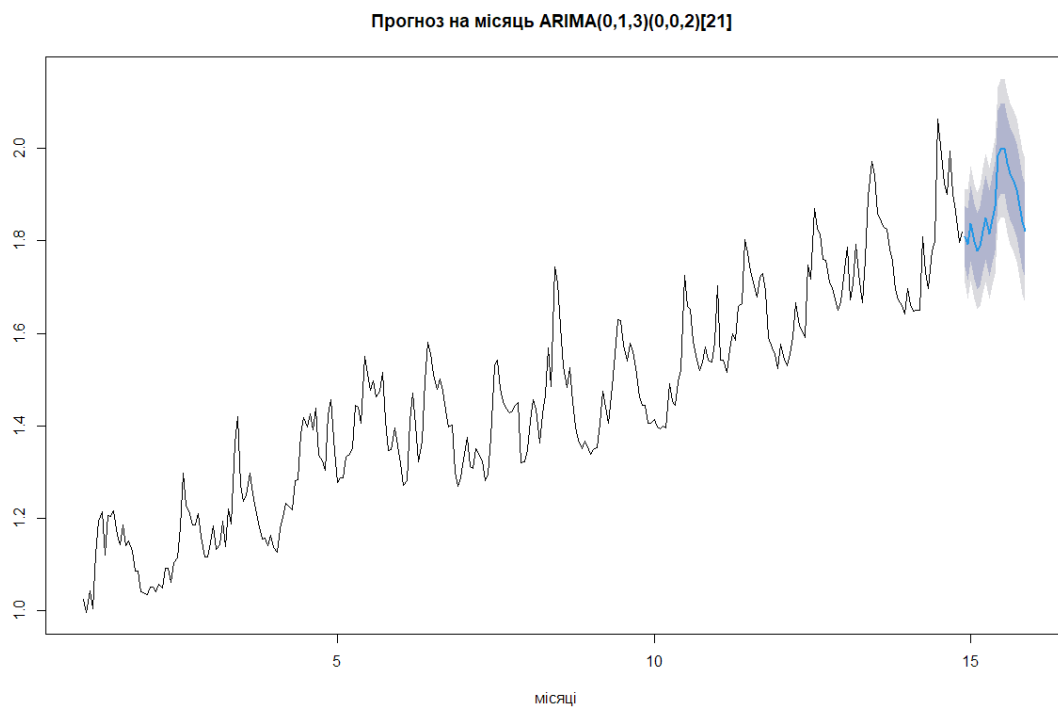


Рис.4.10. Прогноз для ділянки 1.

Значення показника $MARE = 0.081$, якість прогнозу – добра.

,

ВИСНОВКИ

У магістерській роботі проведено аналіз динаміки залишків на рахунках клієнтів банку та їх прогнозування з використанням авторегресійної інтегрованої моделі ковзного середнього з урахуванням сезонності – моделі SARIMA.

Перший розділ є вступним. У ньому розглянуто проблеми аналізу та прогнозування залишків на рахунках клієнтів банку, так званого осаду. Приведено короткий огляд публікацій по цьому питанню.

У другому розділі викладено основи теорії часових рядів, загальні проблеми їх моделювання та методи побудови прогнозу. Розглянуто методи оцінювання моделей, критерії, які використовують для порівняння моделей.

У третьому розділі вивчаються авторегресійні моделі, зокрема авторегресійна інтегрована модель ковзного середнього з врахуванням сезонності (Seasonal Autoregressive Integrated Moving Average – SARIMA).

Результати імплементації моделі SARIMA розглянуто в четвертому розділі. Проведено аналіз залишків на рахунках клієнтів банку та здійснено прогноз. Розрахунок показника похибки моделі MAPE показав високу якість прогнозу.

СПИСОК ЛІТЕРАТУРИ

1. Білава О. С., Татяніна С. М. Аналіз та оцінка ризиків у процесі залучення коштів вітчизняними банками // Економіка і суспільство. – 2018. – № 19. – С. 943-946. DOI: <https://doi.org/10.32782/2524-0072/2018-19-141>.
2. Білик О. І., Сидоренко А. Ю. Прогнозування динаміки залишку коштів на депозитах до запитання // Регіональна економіка. – 2005, № 4. – С. 104-111.
3. Грас Дж. Data Science. Наука о данных с нуля. – СПб.: БХВ-Петербург, 2017. – 336 с.
4. Бокс Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1974. – Кн. 1. – 406 с.
5. Карташов М. В. Імовірність, процеси, статистика. – Київ: Видавничо-поліграфічний центр "Київський університет", 2007. – 494 с.
6. Кендел М. Временные ряды. – М.: Финансы и статистика, 1981. – 199 с.
7. Костіна Н., Черняхівська П. Прогнозування надходження готівкових коштів до установи комерційного банку // Банківська справа. – 2000. – № 1. – С. 17-24.
8. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов: Учеб. пособие. – М.: Финансы и статистика, 2003. – 416 с.
9. Носко В.П. Эконометрика. Введение в регрессионный анализ временных рядов / В.П. Носко. – М., 2002. – 273 с.
10. Орлов Ю. К., Кир'ян О. М. Прогнозування залишків коштів на поточних і депозитних рахунках клієнтів // Штучний інтелект. – 2011. № 2. – С. 171-178.
11. Олійник Д. Управління ліквідністю комерційного банку на основі оптимального використання залишків тимчасово вільних коштів на поточних рахунках клієнтів // Банківська справа. – 2001, №2. – С.31-34.
12. Оліскевич М. О. Основи економетрії часових рядів. – Львів: Видавн. центр ЛНУ ім. Івана Франка, 2008. – 327 с.
13. Суслов В. И. Эконометрия / В. И. Суслов В. И., Н. М. Ибрагимов, Л. П. Талышева, А. А. Цыплаков. – Новосибирск: Изд-во СО РАН, 2005. – 744 с.
14. Hyndman R. J. Athanasopoulos G. Electronic Sources: Forecasting: principles and practice. – 2018. – Available from: <https://otexts.org/fpp2/>.
15. Pollock D. S. G. A Handbook of Time-Series Analysis, Signal Processing and Dynamics. – ACADEMIC PRESS, 1999. – 782 с.