# МОРСЬКИЙ ТА РІЧКОВИЙ ТРАНСПОРТ

УДК 62-50

## РОЗРОБКА СИСТЕМИ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ (СППР) ДЛЯ ПРОГНОЗУВАННЯ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ З АВТОМАТИЗАЦІЄЮ ВИБОРУ КРАЩОЇ МОДЕЛІ

***Бідюк П.І.,***
*Національний технічний університет України «Київський політехнічний інститут»,*
***Бень А.П.,***
*Херсонський державний морський інститут*

**Вступ.** Для аналізу даних, побудови їх математичних моделей, оцінювання прогнозів та прийняття рішень широко застосовують системи підтримки прийняття рішень (СППР). На сьогодні існує велика кількість інформаційних систем, які дозволяють зберігати та обробляти статистичні дані різних процесів, накопичені знання та досвід у створенні методів прогнозування різних характеристик часових рядів [1, 2]. Також існує велика кількість критеріїв якості моделі та ефективності прогнозу. Однак не існує універсальних інформаційно-аналітичних систем, що дозволяють автоматично здійснювати вибір найкращої моделі або найкращого методу для прогнозування окремого часового ряду. Це суттєво обмежує коло користувачів СППР, вимагає від них спеціальної підготовки і вносить суттєвий елемент суб'єктивізму в процес вибору кращого результату.

У даній роботі розглядається створення експертної системи для аналізу, моделювання та прогнозування часових рядів за допомогою різних методів на основі авторегресії та множинної регресії. Система здійснює імпорт досліджуваного часового ряду і будує безліч математичних моделей за допомогою обраних методів. Потім автоматично вибирає найкращу модель за кожним методом, а так само визначає найкращий метод і модель для прогнозування даного часового ряду на основі інтегрального критерію, який ґрунтується на відомих статистичних критеріях якості.

**Постановка задачі.** Необхідно розробити інформаційно-аналітичну систему (ІАС) для автоматичної побудови моделі для досліджуваного часового ряду за кожним реалізованим методом, а також найкращу модель серед усіх побудованих для прогнозування ряду на декілька кроків вперед. Для оцінювання структури моделі застосувати автоматизоване тестування

процесів за допомогою тестів на інтегрованість і гетероскедастичність. Для побудови моделей вибрано такі моделі [3]:

- авторегресія;
- просте ковзне середнє;
- експоненційне ковзне середнє;
- авторегресія із ковзним середнім (ковзне середнє будується за залишками моделі АР);
- авторегресія із ковзним середнім (ковзне середнє будується по вихідному сигналу);
- множинна регресія;
- авторегресія з умовною гетероскедастичністю;
- узагальнена авторегресія з умовною гетероскедастичністю.

Також потрібно побудувати узагальнюючий (консолідований) критерій оцінки якості моделі, визначити можливість його застосування при прогнозуванні для автоматизації процесу вибору кращої моделі з множини кандидатів, побудованих у СППР. При створенні консолідованого критерію пропонується використати такі статистичні параметри якості моделі та прогнозу:

- коефіцієнт детермінації;
- сума квадратів похибок моделі;
- інформаційний критерій Акайке;
- критерій Байєса-Шварца;
- статистика Дарбіна-Уотсона;
- середньоквадратична похибка;
- середня абсолютна похибка в процентах (САПП);
- коефіцієнт Тейла.

**Розробка архітектури СППР.** Спочатку вибираємо архітектуру СППР, виконуємо розробку її функціональної структури, узгоджуємо функціональні вимоги з вимогами замовника. Далі розроблюємо інтерфейс системи на основі узгоджених функціональних вимог (рис. 1).

Система представлення результатів дає можливість представити результати моделювання та прогнозування фінансово-економічних процесів у таких формах:

- графічна форма (лінійні графіки, діаграми різних типів);
- таблична форма;
- текстова форма.

Основними запитами мовної системи, які необхідні для підтримки діалогу користувач-система, є:

- запити на модифікацію та доповнення бази даних і знань;
- введення нових алгоритмів оцінювання параметрів математичних моделей;
- розширення функцій системи за рахунок нових алгоритмів прогнозування фінансових показників;

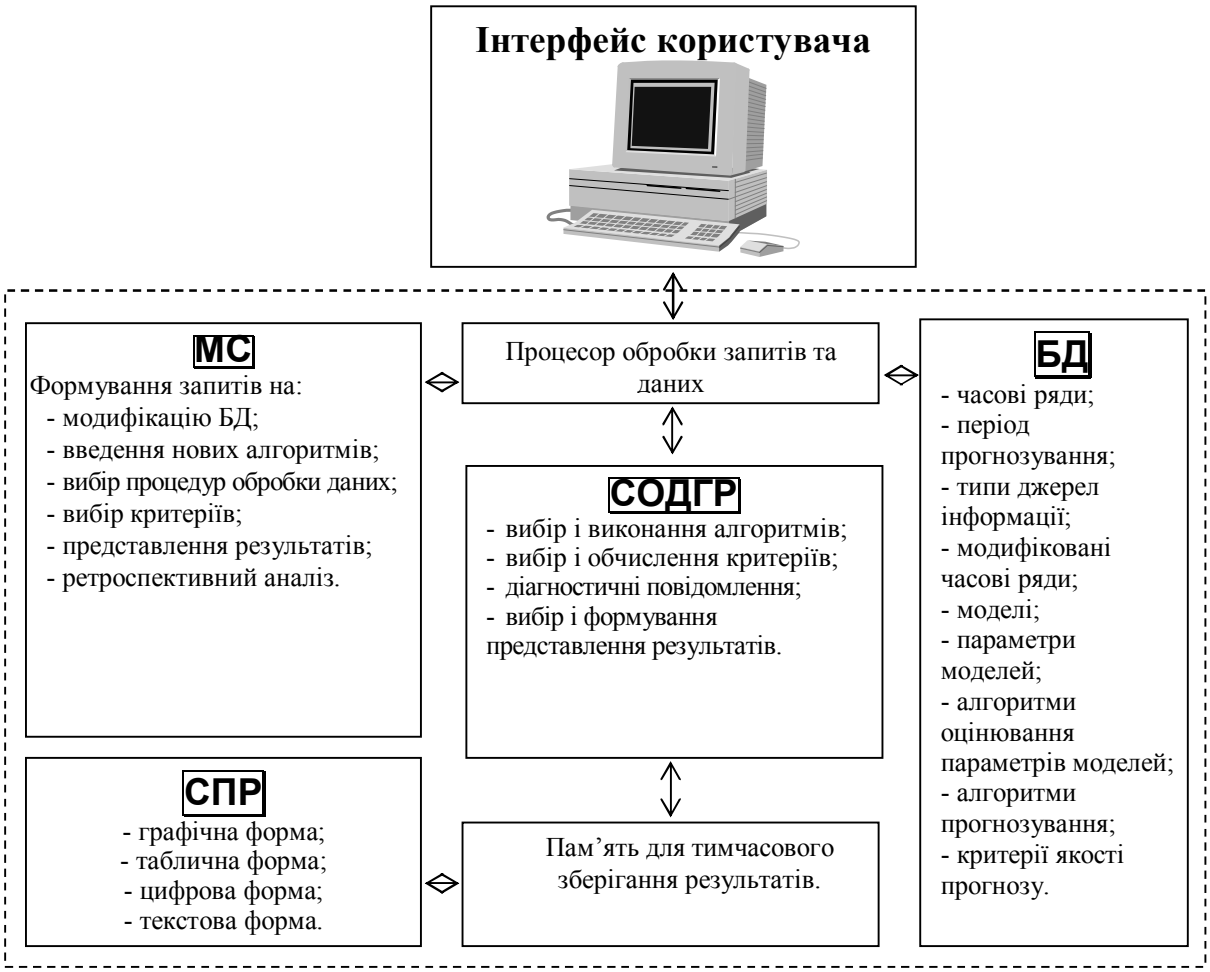


Рисунок 1 – Структурна схема СППР

- запити на формування конкретних процедур обробки даних та прогнозування;
- запити на вибір та формулювання критеріїв розв'язку задачі;
- запити на виконання задач моделювання і прогнозування;
- запити на форму представлення результатів;
- запит на ретроспективний аналіз розв'язування подібних задач у минулому з метою використання отриманих результатів;
- перевірка запитів на коректність та генерування підказок користувачу;
- запит стосовно поточного стану системи.

Головною з точки зору обробки даних СППР є система обробки даних та генерування результатів. Вона сприймає коректні запити користувача і виконує наступні задачі:

- модифікація та доповнення БЗД;
- читання необхідних даних у вигляді часових рядів для побудови моделей;
- вибір з БЗД алгоритмів оцінювання та прогнозування;
- запуск на виконання модулів обробки даних та прогнозування;
- формування результатів обробки даних та їх зберігання в короткостроковій та довгостроковій пам'яті;

- ретроспективний аналіз розв'язків задач, отриманих раніше;
- порівняльний аналіз методів прогнозування;
- генерування діагностичних повідомлень;
- генерування вказівок користувачу щодо можливостей системи.

**Модульна структура системи**. Модульна структура розробленої СППР представлена на рисунку 2. Розглянемо коротко деякі функції, представлені на рисунку. Це функції тестування даних на наявність нестаціонарності, формування структури моделі, оцінювання параметрів моделей та оцінювання якості моделей і прогнозів.

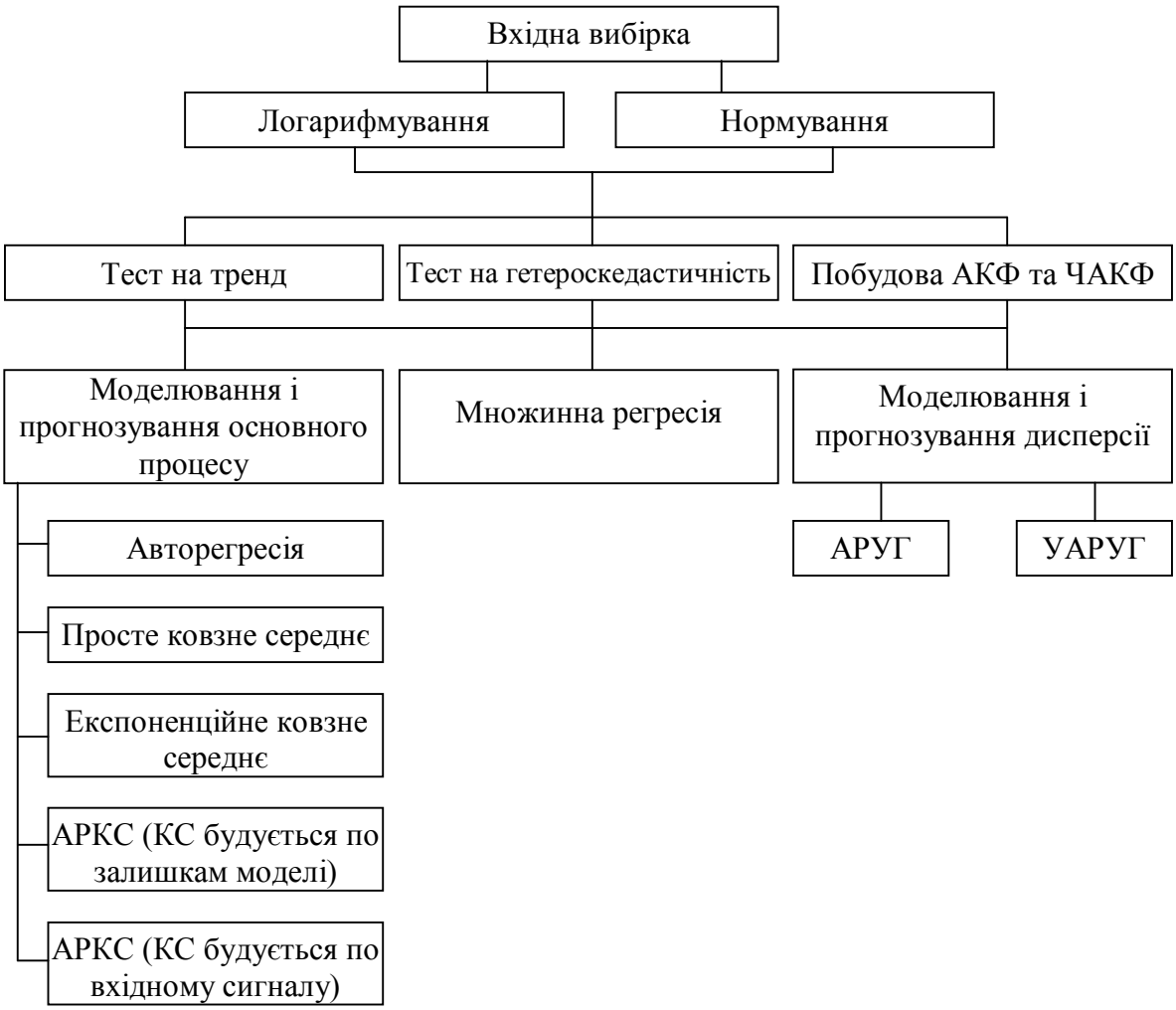


Рисунок 2 – Модульна структура системи

**Тести на виявлення нестаціонарності.** При визначенні наявності нестаціонарності (тобто наявності одиничного кореня в характеристичному рівнянні) пропонується скористатися тестом Дікі-Фуллера. За допомогою цього критерію визначають – яку величину має коефіцієнт $a_1$ у рівнянні:

$$y(k) = a_1 y(k-1) + \varepsilon(k),$$

тобто $a=1$ чи $a<1$. Якщо $a=1$, то дані містять одиничний корінь і ступінь інтегрованості дорівнює $I(1)$. Якщо ж $0<a_1<1$, то ряд стаціонарний, тобто має степінь інтегрованості $I(0)$. Для фінансово-економічних процесів значення $a_1>1$ не характерно, тому що такі значення означають наявність у процесах різко зростаючих (спадаючих) ефектів. Виникнення таких процесів є малоймовірним, оскільки фінансово-економічне середовище є достатньо інерційним і не дозволяє змінним приймати нескінченно великі значення за короткі проміжки часу.

Застосування МНК до розв'язання задачі оцінювання коефіцієнтів моделі часового ряду передбачає скінченність дисперсії залишків (похибок) $e(k)$ моделі. Наявність нестаціонарності призводить до порушення цього припущення. Розглянемо рівняння:

$$y(k)=y(k-1)+e(k)=[y(k-2)-e(k-1)]+e(k)=...$$

$$=y(0)+e(k)+e(k-1)+e(k-2)+\ ...\ +e(1).$$

Оскільки залишки $e(k)$ незалежні і мають постійну дисперсію, то дисперсія $y(k)$ зростає до нескінченності при $k\to\infty$. У такому випадку для описання динаміки ряду можна скористатись рівнянням:

$$y(k)-y(k-1)=a_1\ y(k-1)-y(k-1)+e(k)$$

або

$$\Delta y(k)=b\,y(k-1)+e(k),$$

де $b=a_1-1$. Якщо $b=0$, то ряд містить одиничний корінь і має степінь інтегрованості $I(1)$, а ряд $\{\Delta y(k)\}$ може бути вже стаціонарним. Якщо ж $b<0$, то $a<1$ і стаціонарним буде сам ряд $\{y(k)\}$.

У рівнянні $y(k)=a_1y(k-1)+\varepsilon(k)$ відсутнє середнє значення (перетин) і опис тренду. Якщо включити середнє, то воно приймає вигляд:

$$y(k)=a_0+a_1y(k-1)+\varepsilon(k),$$

або

$$\Delta y(k)=a_0+a_1y(k-1)-y(k-1)+\varepsilon(k)=a_0+by(k-1)+\varepsilon(k).$$

З урахуванням тренду останнє рівняння приймає вигляд:

$$y(k)=a_0+a_1k+a_2y(k-1)+\varepsilon(k),$$

де $k-$ дискретний час. Це рівняння можна записати для першої різниці

$$y(k)-y(k-1)=a_0+a_1k+a_1y(k-1)-y(k-1)+\varepsilon(k),$$

або

$$\Delta y(k-1) = a_0 + a_1 k + by(k-1) + \varepsilon(k).$$

Для такої моделі було б некоректно використовувати $t-$статистику з метою визначення значущості коефіцієнта $b$, оскільки застосування регресії для оцінювання цього коефіцієнта передбачає, що $b<0$ $(a_1<1)$. Тобто при $b\approx 0$ великий процент оцінок за $t-$статистикою не буде прийматися як значущий, тобто нульова гіпотеза стосовно існування одиничного кореня буде часто відкидатись.

Крім того, одиничні корені *робастні* (зберігаються і можуть бути виявлені) при різних ступенях гетероскедастичності, але можуть виникати проблеми з автокореляцією залишків моделі. В умовах наявності автокореляції залишків задача тестування на стаціонарність розв'язується за допомогою розширеного тесту Дікі-Фуллера. При використанні цього методу значення залежної змінної вводяться в рівняння регресії з великими значеннями лагу, достатніми для того, щоб уникнути автокореляції залишків. Це рівняння може мати такий вигляд:

$$\Delta y(k) = a_0 + by(k-1) + c_1\,\Delta y(k-1) + c_2\,\Delta y(k-2) + \ldots + c_n\,\Delta y(k-n) + \varepsilon(k).$$

Форма критерію значущості залежить від вигляду моделі, що тестується, тобто чи введено у модель середнє значення і член, який описує тренд.

**Критерії якості моделі та прогнозу.** Для оцінювання якості моделей та оцінок прогнозів, які будуються за допомогою СППР, вибрані статистичні критерії якості, наведені нижче.

**Коефіцієнт детермінації ($R^2$).** За міру інформативності часового ряду часто використовують його дисперсію. Коефіцієнт $R^2$ – це відношення дисперсії тієї частини часового ряду основної змінної, що описується отриманим рівнянням, до вибіркової дисперсії цієї змінної. Він обчислюється за формулою:

$$R^2 = \frac{\mathrm{var}(\hat{y})}{\mathrm{var}(y)}. \qquad (1)$$

Для адекватної моделі коефіцієнт детермінації повинен прямувати до одиниці, тобто: $R^2 \to 1$.

**Сума квадратів похибок моделі (*SSE*).** $\sum e^2(k)$, тобто

$$SSE = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2. \qquad (2)$$

де $\hat{y}(k) = \hat{a}_0 + \hat{a}_1\hat{y}(k-1) + \hat{a}_2\hat{y}(k-2) + \hat{b}_1 x(k) + \hat{b}_2 z(k)$; $y(k)-$вимірювання; $N-$ довжина вибірки. З можливих кандидатів необхідно вибирати ту модель, для

якої $\sum e^2(k)$ приймає мінімальне значення на множині оцінок векторів параметрів моделей-кандидатів.

**Інформаційний критерій Акайке (*AIC*).** Цей критерій враховує суму квадратів похибок, кількість вимірів $N$ і кількість оцінюваних параметрів $p$:

$$AIC = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + 2p. \tag{3}$$

Для кращої моделі критерій приймає менше значення, оскільки він залежить від СКП.

**Критерій Байєса-Шварца (*BSC*).** Даний критерій схожий на попередній, проте він додатково враховує довжину вибірки за допомогою члена $\ln(N)$:

$$BSC = N \ln\left[\sum_{k=1}^{N} e^2(k)\right] + p \ln(N). \tag{4}$$

Його рекомендують використовувати при довгих вибірках вимірювальних даних**.**

**Статистика Дарбіна-Уотсона (*DW*).** Статистика Дарбіна-Уотсона обчислюється за формулою:

$$DW = 2 - 2p, \tag{5}$$

де $\rho -$ коефіцієнт кореляції між значеннями випадкової змінної $\varepsilon(k) \approx e(k)$, тобто $\rho = \mathrm{cov}[e(k)] = E[e(k)e(k-1)]$. Цей параметр дозволяє визначити ступінь корельованості похибок моделі. При повній відсутності кореляції між похибками $DW = 2$, – це ідеальне значення даного параметра.

**Середньоквадратична похибка (*СКП*)**:

$$СКП = \sqrt{\frac{1}{S}\sum_{i=1}^{S} y(k+s) - \hat{y}(k+s,k)}. \tag{6}$$

**Середня абсолютна похибка в процентах (*САПП*)**:

$$САПП = \frac{1}{S}\sum_{i=1}^{S} \frac{\left|y(k+s) - \hat{y}(k+s,k)\right|}{\left|y(k+s)\right|} 100\%. \tag{7}$$

**Коефіцієнт Тейла (*U*).** Коефіцієнт Тейла – дуже важливий індикатор точності моделі і її сумісності з вихідним рядом даних:

$$U = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i)^2} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i)^2}}. \quad (8)$$

За побудовою, його величина знаходиться між 0 і 1. Якщо $U = 1$, модель не може бути використана для прогнозу. Якщо $U$ *прямує до нуля*, то прогнозовані ряди співпадають з реальними рядами і модель є придатною для оцінювання прогнозів.

**Формалізація консолідованого критерію.** При розв'язанні поставленої задачі ми виходили з того, що необхідно об'єднати (1)-(8) таким чином, щоб усі критерії, які входять до консолідованого, мали однаковий вплив на кінцевий результат. Для цього пропонується нормувати значення деяких з них.

Таким чином, у консолідованому критерії пропонується застосовувати наступні значення: замість $R^2$ використовується значення $e^{1-R^2}$; замість $SSE$ $\rightarrow \frac{SSE}{N}$, де $N$ – довжина вибірки. Оскільки інформаційний критерій Акайке та критерій Байєса-Шварца схожі за змістом, пропонується згрупувати їх в один елемент консолідованого критерію, а для врахування від'ємних значень пропонується використати наступне перетворення:

$$\begin{cases} \ln(AIC + BSC), & AIC + BSC \geq 0 \\ e^{AIC+BSC}, & AIC + BSC < 0 \end{cases}, \; DW \rightarrow e^{2-DW}, \; СКП \rightarrow \ln(СКП), \; САПП \rightarrow \ln(САПП), \; U \rightarrow e^{U}.$$

Таким чином, запропонований евристичний консолідований критерій має такий вигляд:

$$KK = e^{1-R^2} + \frac{SSE}{N} + \left\{ \begin{matrix} \ln(AIC + BSC), & AIC + BSC > 0 \\ e^{AIC+BSC}, & AIC + BSC \leq 0 \end{matrix} \right\} + e^{2-DW} + \\ + \ln(СКП) + \ln(САПП) + e^{U}$$

**Приклади застосування.** Перевірка запропонованої СППР та консолідованого критерію виконувалась на прикладі ІСЦ (рис. 3.) України (щомісячні показники).

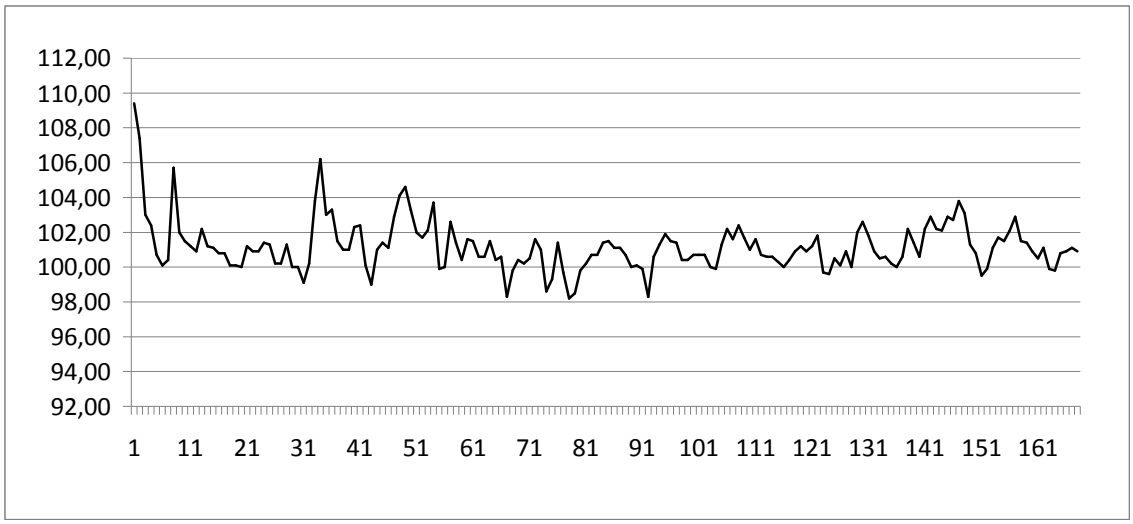


Рисунок 3 – ІСЦ України за період з січня 1996 по грудень 2009 (щомісячні показники)

**Приклад 1. Моделювання ІСЦ України**

Таблиця 1 – Результати моделювання індексу оптових цін

| Тип моделі | Характеристики моделі | | | | | Характеристики прогнозу | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ІК | $R^2$ | AIC | DW | BSC | СКП | САПП | SSE | Коеф. Тейла |
| АР(8) | 0,4133 | 0,6213 | -4,6583 | 2,6910 | -4,6583 | 0,2807 | 0,2572 | 0,3939 | 0,0014 |
| SMA(6) | -1,3974 | 0,1814 | 0,007 | 1,7130 | 0,007 | 0,4475 | 0,3253 | 1,0014 | 0,0022 |
| EMA(5) | 0,5925 | 0,5722 | -5,5354 | 2,1760 | -5,5354 | 0,2571 | 0,2256 | 0,3305 | 0,0013 |
| АРМА1(8,3) | 0,0174 | 0,9104 | -4,7419 | 2,8189 | -4,7419 | 0,2783 | 0,2679 | 0,3874 | 0,0014 |
| АРМА2(4,2) | 0,0540 | 0,7273 | -5,4040 | 2,8960 | -5,4040 | 0,2605 | 0,2486 | 0,3393 | 0,0013 |

**Приклад 2. Моделювання дисперсії ІСЦ України**

Таблиця 2 – Результати моделювання дисперсії індексу оптових цін

| Тип моделі | Характеристики моделі | | | | | Характеристики прогнозу | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ІК | $R^2$ | AIC | DW | BSC | СКП | САПП | SSE | Коеф. Тейла |
| АРУГ(7) | 23,872 | 0,0279 | 19,174 | 1,3583 | 19,174 | 3,0427 | 40,162 | 46,288 | 0,4818 |
| УАРУГ(6,5) | 20,810 | 0,0421 | 17,782 | 1,9083 | 17,782 | 2,6472 | 62,080 | 35,037 | 0,3557 |

**Приклад 3. Множинна регресія. Моделювання цін на акції підприємства «АзовСталь»**

Таблиця 3 – Результати моделювання індексу оптових цін

| Тип моделі | Характеристики моделі | | | | | Характеристики прогнозу | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ІК | $R^2$ | AIC | DW | BSC | СКП | САПП | SSE | Коеф. Тейла |
| Множинна регресія | 10,545 | 0,0226 | -14,406 | 1,0062 | -14,406 | 0,1059 | 5,7476 | 0,0561 | 0,0298 |

**Висновки.** Побудована комп'ютерна інформаційна система підтримки прийняття рішень відповідає сучасним вимогам до систем такого типу. Сформульовано вимоги користувача та функціональні вимоги до СППР. Система включає множину тестів для попереднього тестування даних з метою їх віднесення до визначеного класу – стаціонарні чи нестаціонарні. Також передбачена попередня обробка даних з метою приведення їх до форми, яка забезпечує належні умови для оцінювання параметрів моделей-кандидатів. Структура моделі оцінюється за допомогою автокореляційної, часткової автокореляційної функції, кореляційної матриці та функцій взаємної кореляції. Для оцінювання параметрів моделей передбачена множина методів, які дають можливість оцінювати лінійні та нелінійні за параметрами моделі. З метою автоматизації процесу побудови та вибору моделі запропоновано інтегрований критерій, який ґрунтується на множині статистичних параметрів якості. Виконано перевірку функціонування СППР на фінансово-економічних процесах. Передбачається подальше удосконалення функціональних можливостей системи за рахунок введення нових методів оцінювання параметрів і тестів для визначення наявності можливих складових випадкових процесів.

## ЛІТЕРАТУРА

1. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. – М.: Мир, 1974. – 408 с.

2. Бідюк П.І., Половцев О.В. Аналіз та математичне моделювання економічних процесів перехідного періоду. – К.: ПЛАБ-75, 1999. – 209 с.

3. Бідюк П.І., Савенков О.І., Баклан І.В. Часові ряди: моделювання та прогнозування. – К.: ЕКМО, 2003. – 144 с.