Системні науки та кібернетика
System Sciences & Cybernetics

Електронний науковий журнал
Electronic Science Journal
№ 5 June, 2016 ISSN 2221-7851

Університетський науковий збірник «Системні науки та кібернетика» створено для публікації студентських наукових статтей при проведенні науково-дослідницької роботі у магістратурі.

Publisher: ESC «Institute for Applied System Analysis» NTUU «KPI»

www.mmsa.kpi.ua/ssc

СИСТЕМНІ НАУКИ ТА КІБЕРНЕТИКА

Науковий електронний збірник НТУУ «КПІ»
№ 5 • 2016
Виходить з вересня 2010 року

АДРЕСА РЕДАКЦІЇ:
03056, м. Київ пр. Перемоги, 37, корп. 35, каб. 307
Навчально-науковий комплекс «Інститут прикладного системного аналізу»
НТУУ «КПІ» МОН та НАН України
Тел.: (044) 406-8359
http://mmsa.kpi.ua/
Відповідальний секретар: Н.В. Кузнєцова
Верстка: Н.В. Кузнєцова

Шановні читачі!

Навчально-науковий комплекс «Інститут прикладного системного аналізу» НТУУ «КПІ» МОН та НАН України та факультет інформатики та обчислювальної техніки НТУУ «КПІ» видають університетський науковий електронний збірник:

«СИСТЕМНІ НАУКИ ТА КІБЕРНЕТИКА».

Збірник публікує роботи теоретичного і прикладного характеру з широкого спектру проблем, пов'язаних із системними науками та кібернетикою.

Провідні тематичні розділи журналу:

- методи обробки інформації;
- прикладні методи системного аналізу;
- проектування кібернетичних систем;
- системні дослідження;
- математичне і програмне забезпечення;
- системи підтримки прийняття рішень;
- комп'ютерні системи та мережі;
- розв'язування задач оптимізації математичної кібернетики;
- теорія управління та обчислювальна математика;
- математичне моделювання та прогнозування економічних, технічних, екологічних і соціальних процесів;
- аналіз процесів і систем в умовах ризиків.

Головний редактор наукового збірника — заступник директора Навчально-наукового комплексу «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут» МОН та НАН України, доктор технічних наук, професор Віктор Демидович РОМАНЕНКО.

Університетський науковий електронний збірник «Системні науки та кібернетика» створено для публікації студентських наукових статей при проведенні науково-дослідної роботи в магістратурі.

Статті у збірнику публікуються українською, російською та англійською мовами у тому вигляді, в якому були надані авторами. Автори повністю несуть відповідальність за усі матеріали, наведені у статтях.

СИСТЕМНІ НАУКИ ТА КІБЕРНЕТИКА

5 • 2016

ЗМІСТ

УДК 004.6

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ПРОЦЕСІВ НА БІРЖІ

Михайлюк О.О.

Науковий керівник: д.т.н., професор Бідюк П.І.

Вступ

Дослідження та передбачення процесів на біржі цінних паперів є надзвичайно актуальною темою, як і для простих трейдерів (посередників), так і для інвесторів компаній (адже динаміка зміни цін акцій свідчить про стан останньої), акції яких розміщені на біржі. При цьому важливо спрогнозувати не тільки майбутню ціну акцій, але й можливе відхилення від отриманого прогнозу.

Для того щоб отримати прибуток, більшість приватних інвесторів і керівників фондів повинні спостерігати за напрямком розвитку цін на активи. Традиційна, або фундаментальна стратегія вивчає всі аспекти ринку, всі чинники, які впливають або можуть вплинути на ціну, загальний стан економіки (якщо це необхідно), і визначає вартість даного механізму інвестування. Крім основних факторів, багато хто вивчає ще й так звані технічні фактори. Методи технічного аналізу, суть якого полягає в дослідженні динаміки цін за попередній період часу, використовують для вироблення рекомендацій щодо інвестування. Дослідження графіків для виявлення певних цінових моделей (виходячи з очікування того, що вони повторюються) або тенденцій є досить поширеним технічним інструментом. Сьогодні технічний аналіз застосовується настільки широко, що інвестиційні установи обов'язково наймають як мінімум одного аналітика, а мова і термінологія технічного аналізу використовуються у фінансовій пресі[1].

Проблемі прогнозу волатильності присвячена книга Кевіна Конноллі «Покупка і продаж волатильності» (перше посилання у списку літератури). Існує багато робіт присвячених самій торгівлі на біржі (зокрема на американській). Наприклад, книга Ніколаса Талеба «Обдурені випадковістю» і книга Марка Дагласа «Дисциплінований трейдер». Існує багато книг з економетрики, які описують моделі для прогнозування різних економічних процесів, в тому числі і процесів на біржі. В якості прикладів, які використовувалися в досліджені, можна привести наступні книги: «Економетричний аналіз часових рядів: конспект лекцій» Бідюка П.І. та «Эконометрика. Введение в регрессионный аналіз временних рядов» Носка В.П.

Дослідження присвячене аналізу процесів на біржі, побудові та застосуванню моделей ціноутворення акцій та моделей волатильності. На основі моделей АРКС і АРІКС реалізована СППР, із допомогою якої можливо визначати найкращі моделі ціноутворення акцій.

Постановка задачі

В рамках дослідження необхідно виконатинаступні операції:

а) Розробити власну СППР для моделювання і прогнозування процесів на біржі.

б) Зібрати статистичні дані про ціни акцій.

в) Побудувати модель, яка описує процес ціноутворення акцій.

г)Обчислити оцінки короткострокових прогнозів на основі побудованої моделі та статистичні параметри якості.

д) Виконати порівняльний аналіз результатів застосування власної системи із уже існуючими.

е) Змоделювати можливі відхилення отриманих прогнозів.

Види процесів на біржі та їх моделі

Опис типів процесів на біржі

Розрізняють стаціонарність слабку і сильну. Слабку стаціонарність називають ще

– стаціонарністю по коваріації;

– стаціонарністю другого порядку;

– стаціонарністю в широкому смислі.

Формально стохастичний процес із скінченним середнім та дисперсією називають стаціонарним по коваріації, якщо для всіх k і $k-s, (k=0,1,2,\ldots; s=0,1,2,\ldots)$ виконуються наступні умови:

– математичне сподівання не залежить від часу:

$$E\,[y(k)] = E\,[y(k-s)] = \mu = const, \qquad (1)$$

– дисперсія залишається постійною для всього часового інтервалу, на якому розглядається процес:

$$E\left\{[y(k)-\mu]^2\right\} = E\left\{[y(k-s)-\mu]^2\right\} = \sigma_y^2 = const, \qquad (2)$$

– автоковаріація залишається незмінною в часі для всього часового інтервалу:

$$E\left\{[y(k)-\mu]\,[y(k-s)-\mu]\right\} = E\left\{[y(k-j)-\mu]\,[y(k-j-s)-\mu]\right\} = \gamma(s) = const.$$

Сильна стаціонарність не вимагає, щоб математичне сподівання та дисперсія були константами [2].

Таким чином, слабка стаціонарність накладає більш строгі обмеження на поведінку випадкового процесу ніж сильна.

Серед нестаціонарних процесів виділяють інтегровані і гетероскедастичні.У випадку, коли математичне сподівання змінюється в часі(порушується умова (1)), то такий процес називають **процесом з трендом або інтегрованим процесом** (по аналогії із характером зміни сигналу на виході інтегратора) або процесом з одиничними коренями (відповідного характеристичного рівняння).

При цьому під трендом будемо розуміти поточне середнє значення процесу, яке може бути отримане за допомогою процедури цифрової фільтрації. В простому випадку це може бути формула для обчислення поточного середнього значення [2].

Гетероскедастичними називають процеси із змінною в часі дисперсією (порущується умова (2)), а **гомоскедастичними** – процеси із сталою дисперсією на відрізку часу, що розглядається при моделюванні та прогнозуванні.

Дисперсію та стандартне відхилення часто використовують як міру ризику при дослідженні процесів на біржі, а тому цій проблемі приділяється значна увага в спеціальній літературі [2]. При дослідженні фінансових процесів дисперсію та стандартне відхилення використовують як міру вольтинності (мінливості) процесу.

Моделі для опису процесів на біржі

Для опису **стаціонарних процесів** в досліджені використовувалася модель авторегресії з ковзним середнім, яка є узагальненням моделей авторегресії (можна отримати з моделі АРКС, поклавши q=0)і ковзного середнього (можна отримати з моделі АРКС, поклавши p=0). Описується наступним рівнянням:

$$y(k) = a_0 + \sum_{i=0}^{p} a_i y(k-i) + \sum_{j=0}^{q} b_j \varepsilon(k-j),$$

де $\varepsilon(k)$ - білий шум;

a_i і b_j - дійсні числа, авторегресійні коефіцієнти і коефіцієнти ковзного середнього, відповідно.

Для визначення параметрів рта qможна використати (фактично ми встановлюємо структуру моделі) автокореляційну і частково автокореляційну функції (приклад цього наведено в дослідженнях). Для визначення параметрів a_i і b_j можна застосувати метод найменших квадратів і його рекурентну модифікацію, яка не потребує обертання матриці, а також метод максимальної правдоподібності (але в дослідженнях його не застосовували і в розробленій СППР він не реалізований).

Для опису **інтегрованих процесів** застосовується модель авторегресії з інтегрованим ковзним середнім, яка базується на АРКС і описується наступним чином:

$$\Delta^d y(k) = a_0 + \sum_{i=0}^{p} a_i \Delta^d y(k-i) + \sum_{j=0}^{q} b_j \varepsilon(k-j),$$

де $\varepsilon(k)$ - білий шум;

a_i і b_j - дійсні числа, авторегресійні коефіцієнти і коефіцієнти ковзного середнього, відповідно;

Δ^d - оператор різниці часового ряду порядку d.

Для моделювання умовної дисперсії гетероскедастичних процесів застосовують модель авторегресії умовної гетероскедастичності :

$$E_k[\hat{\varepsilon}^2(k+1)] = \alpha_0 + \alpha_1\hat{\varepsilon}^2(k-1) + \alpha_2\hat{\varepsilon}^2(k-2) + ... + \alpha_q\hat{\varepsilon}^2(k-q),$$

і її модифікація – узагальнена авторегресійна умовно гетероскедастична модель :

$$h(k) = \alpha_0 + \sum_{i=1}^{q}\alpha_i\varepsilon^2(k-i) + \sum_{i=1}^{p}\beta_i h(k-i),$$

які теж базуються на моделі АРКС.

Опис розробленої СППР

Система підтримки прийняття рішень або СППР — це комп'ютерна система, яка шляхом збору і аналізу великої кількості інформації може впливати на процес ухвалення рішень організаційного плану в бізнесі чи підприємництві.

Реалізована СППР є найбільш простою з погляду архітектури, тому її впровадження буде доцільним в організаціях, що не ставлять перед собою глобальних завдань і що мають невисокий рівень розвитку інформаційних технологій.

Архітектура створеної СППР налічує наступні рівні:

- завантаження і обробка даних;
- аналіз даних;
- побудова та вибір кращої моделі;
- прогнозування.

Покажемо реалізовану СППР в термінології класів ООП, щоб відобразити взаємозв'язки між окремими компонентами, а також описати їх внутрішню структуру та типи відношень (рис.1):

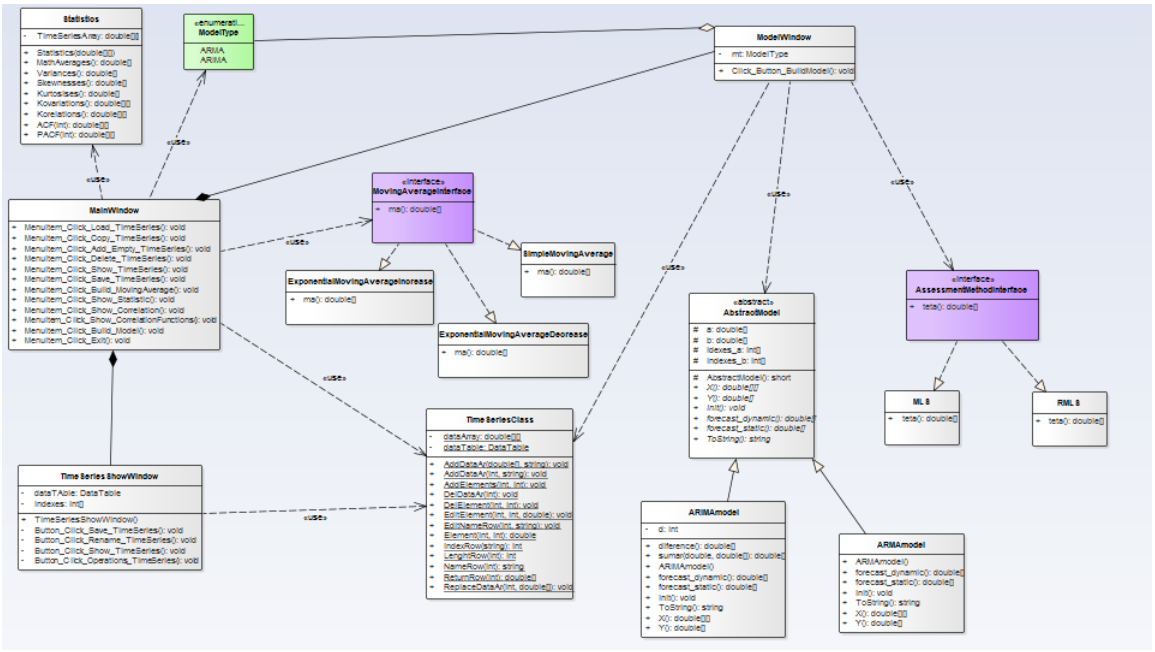


Рис.1 Діаграма основних класів

Приклад застосування розробленої СППР

Опис зібраних статистичних даних

Для дослідження були зібрані ціни акцій компанії «Укрнафта» у період з 14 березня 2005 року по 10 березня 2006 року. Графік цін акцій «Укрнафти» наведено на рис.2:

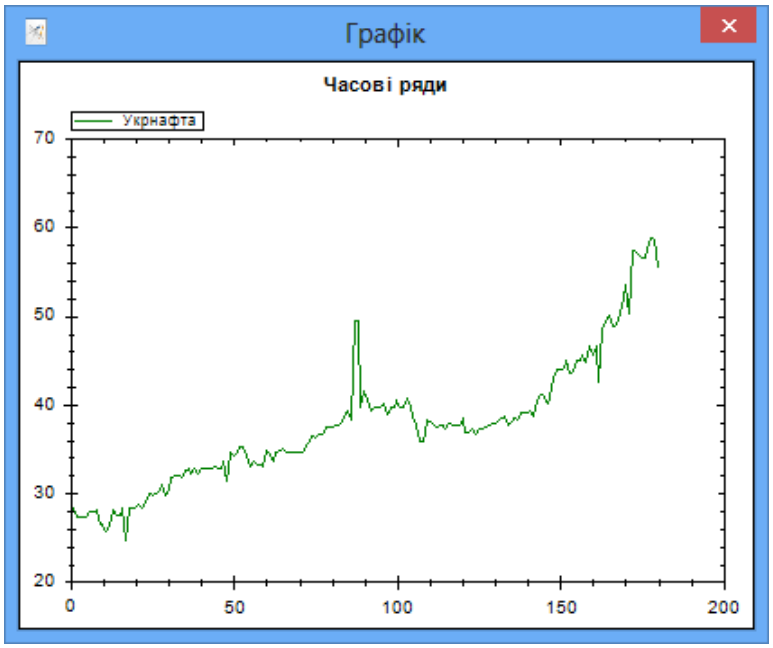
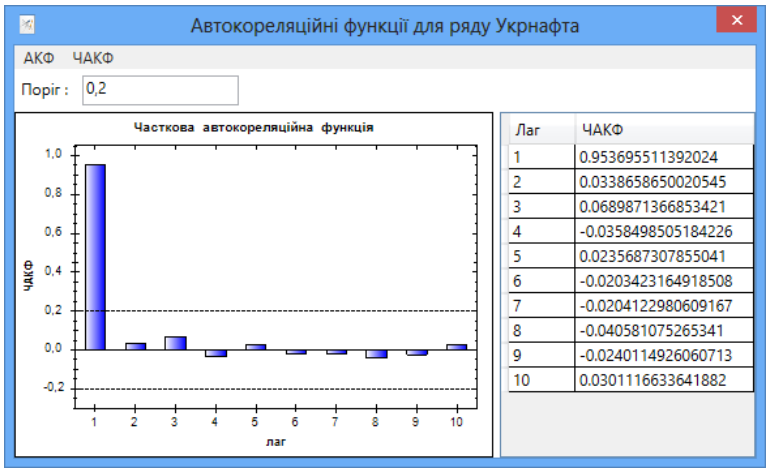


Рис.2 Графічне представлення цін акцій компанії «Укрнафта»

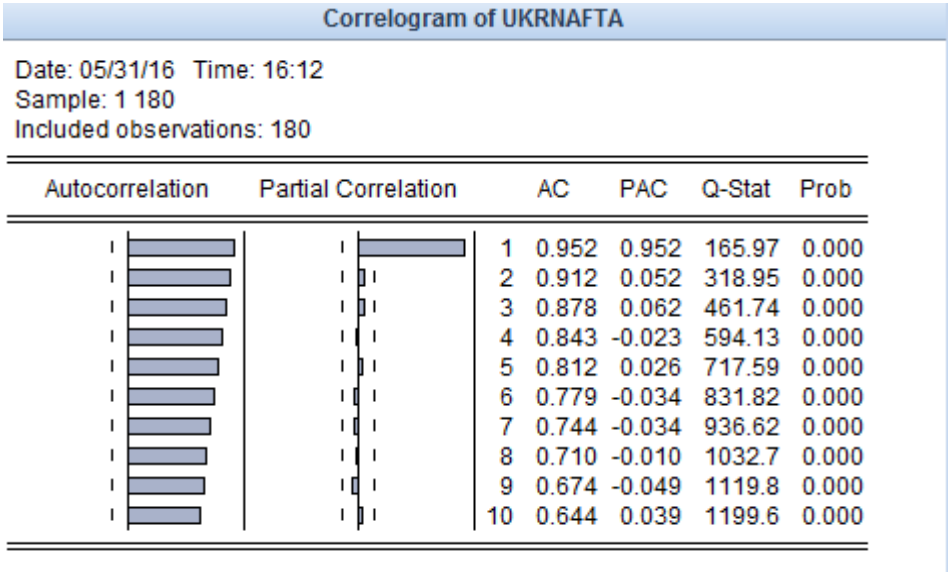
Побудова моделі

Будемо вважати, що цей процес є стаціонарним (візуальний аналіз).

Побудуємо ЧАКФ для даного процесу (рис.3):



а) розроблена СППР



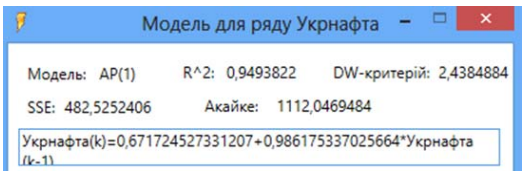
б) Eviews

Рис.3 ЧАКФ для цін акцій компанії «Укрнафта»

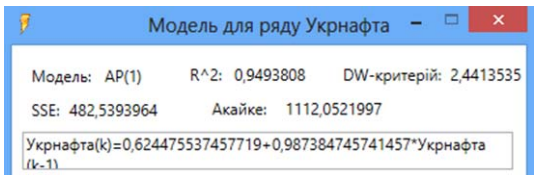
Як можна побачити розроблена СППР (рис.3, а) і Eviews (рис.3, б) дають майже однакові результати. Судячи із ЧАКФ доцільно буде побудувати модель АР(1), яку описує формула:

$$y(k) = a_0 + a_1 y(k-1).$$

Оцінимо коефіцієнти цього рівняння за допомогою МНК і РМНК у розробленій СППР та у Eviews (рис.4):



а) МНК розроблена СППР



б) РМНК розроблена СППР

Dependent Variable: UKRNAFTA
Method: Least Squares
Date: 05/31/16 Time: 16:37
Sample (adjusted): 2 180
Included observations: 179 after adjustments
UKRNAFTA=C(1)+C(2)*UKRNAFTA(-1)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.707196	0.665313	1.062952	0.2893
C(2)	0.985011	0.017596	55.97789	0.0000

R-squared	0.946534	Mean dependent var	37.26816
Adjusted R-squared	0.946232	S.D. dependent var	7.312114
S.E. of regression	1.695526	Akaike info criterion	3.904974
Sum squared resid	508.8414	Schwarz criterion	3.940587
Log likelihood	-347.4952	Hannan-Quinn criter.	3.919415
F-statistic	3133.524	Durbin-Watson stat	2.482144
Prob(F-statistic)	0.000000		

в)Eviews

Рис.4 Характеристика моделі АР(1) оцінена для цін акцій компанії «Укрнафта»

Як ми бачимо кращий результат оцінювання дала розроблена СППР (як за допомогою МНК (рис.4, а), так і за допомогою РМНК(рис.4, б)) ніж Eviews (рис.4, в). Найкращий результат дав МНК у розробленій СППР. Для зручності результати всіх оцінювань зведені у табл.1:

Табл.1Оцінки параметрів моделі АР(1) для цін акцій компанії «Укрнафта»

Коефіцієнт	Оцінка в СППР, метод МНК	Оцінка в СППР, метод РМНК	Оцінка в Eviews
a_0	0,6717245	0,624475	0,707196
a_1	0,986175	0,987384	0,985010

Для всіх трьох випадків маємо доволі високий показник адекватності (коефіцієнт детермінації вище 0,94). Тому не має необхідності покращувати модель.

Найкраща модель виражається такою формулою:

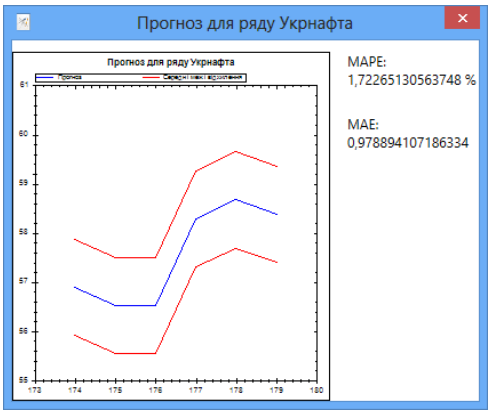
$$y(k) = 0{,}67172 + 0{,}98617 \cdot y(k-1)$$

.

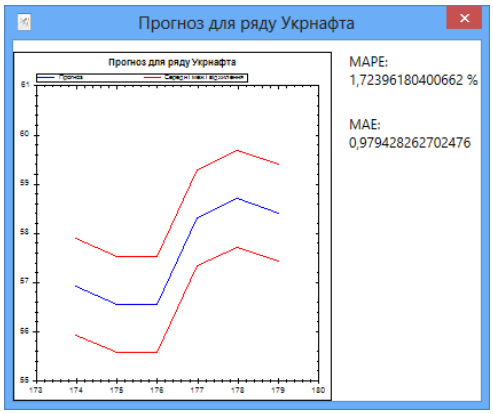
Враховуючи те, що модуль a_1 менше одиниці, то наше припущення щодо стаціонарності процесу є правильним (достатня умова стаціонарності моделі авторегресії).

Прогнозування цін акції

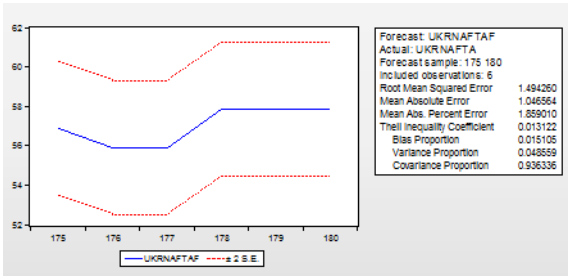
Тепер спробуємо спрогнозувати (статичним методом) п’ять останніх вимірів цін акцій (рис.5):



а) МНК розроблена СППР



б) РМНК розроблена СППР



в)Eviews

Рис.5 Прогноз для цін акцій компанії «Укрнафта» на основі моделі АР(1)

Як можна побачити, і прогнози на основі моделі АР(1), оціненої за допомогою МНК (рис.5, а) і РМНК (рис.5, б) у розробленій СППР та за допомогою Eviews (рис.5, в) дали схожі результати. Конкретні значення прогнозу наведені в табл.2:

Табл.2 Результати прогнозування цін акцій компанії «Укрнафта» на основі моделі АР(1)

Вимір	Реальне значення	Розроблена СППР			
		МНК		РМНК	
		Прогноз	Похибка	Прогноз	Похибка
175	56,63	56,8837	0,2537	56,9054	0,2754
176	56,63	56,5188	0,1112	56,5400	0,0899
177	58,42	56,5188	1,9012	56,5400	1,8799
178	58,81	58,2840	0,5259	58,3074	0,5025
179	58,52	58,6686	0,1487	58,6925	0,1726
180	55,45	58,3827	2,9327	58,4062	2,9562
Сер. Похибка			0,9789		0,9794

Як показує табл.2, обидва варіанти оцінювання моделі АР(1) дають прийнятні результати прогнозування, проте для прогнозу волатильності ми будемо працювати з квадратами залишків моделі одержаними при оцінці МНК моделі АР(1) у власній СППР (адже модель має найкращі показники і дає найкращий прогноз).

Моделювання волатильності

Графік квадратів похибок має такий вигляд (рис.6) :

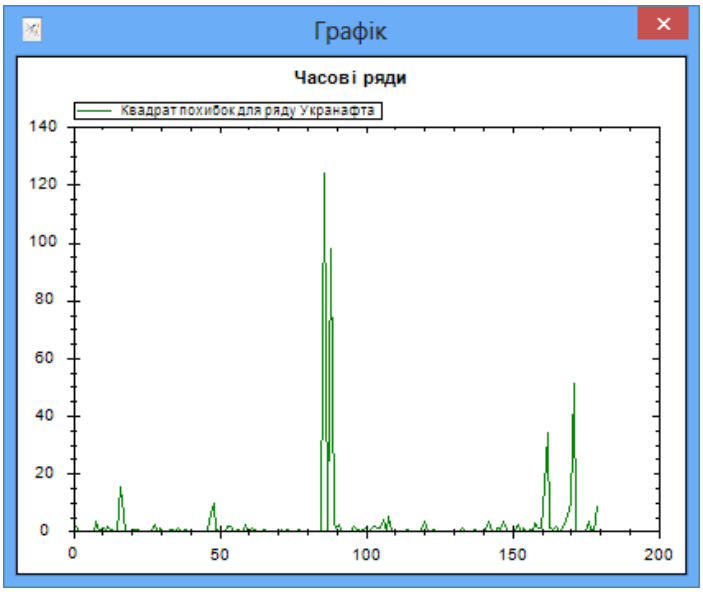


Рис.6Графік для квадратів похибок моделі АР(1) цін акцій «Укрнафта»

Для прогнозування волатильності (відхилення) оцінимо ЧАКФ залишків моделі, щоб визначитися з структурою моделі для волатильності (рис.7):

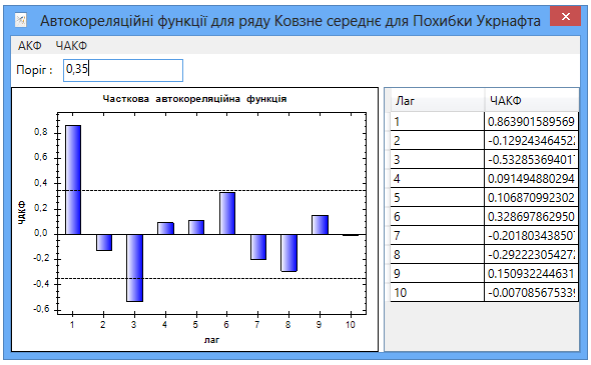


Рис.7 ЧАКФ для квадратів похибок моделі АР(1) цін акцій «Укрнафта»

Судячи з ЧАКФ, от похибки моделі будуть описуватися моделлю АРУГ(3), яка після оцінювання описується формулою:

$$\hat{\varepsilon}^2 = 0,46357 - 2,88878 \cdot \hat{\varepsilon}^2(k-1) + 0,54878 \cdot \hat{\varepsilon}^2(k-3).$$

Ця модель має наступні властивості (рис.8):

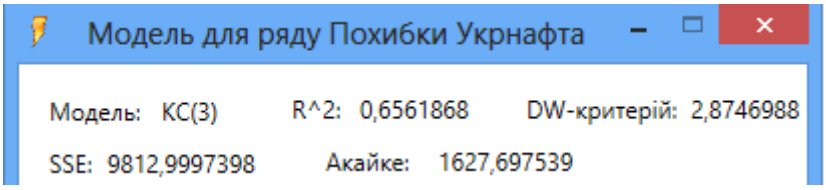


Рис.8Властивості моделі АРУГ(3) для акцій «Укрнафти»

Побудуємо по ній прогноз на 10 наступних цін акцій компанії «Укрнафта». На рис.9 наведений графік прогнозу цін акцій з урахуванням волатильності:

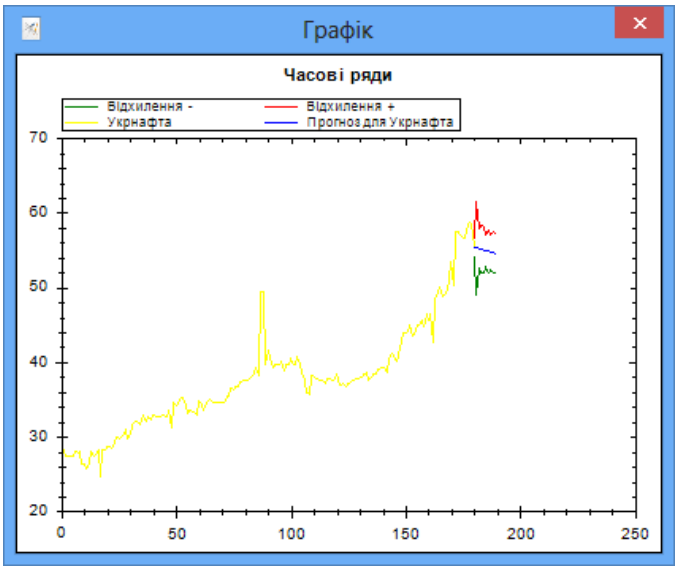


Рис.9 Прогноз цін акцій «Укрнафти» з можливим відхиленням

Аналіз отриманих результатів

В ході дослідження побудовано модель АР(1) ціноутворення акцій компанії «Укрнафта» за допомогою власної СППР і пакету Eviews. На основі цієї моделі можна робити коротко- та середньострокові прогнози. Також на основі залишків моделі була побудована модель АРУГ(3), яка дозволяє прогнозувати волатильність прогнозів отриманих за основною моделлю.

Такий результат можна застосовувати для прийняття стратегії дій на біржі. Також можливе подальше вдосконалення СППР, при якому вона сама прийматиме рішення щодо купівлі/продажу цінних паперів.

Висновки

В статті виконано огляд існуючих методів моделювання і прогнозування процесів на біржі.

Спроектована власна система підтримки прийняття рішень, яка дає майже ідентичні результати в порівнянні з вже існуючими, проте не потребує вивчення додаткових команд і має інтуїтивно зрозумілий інтерфейс.

В результаті дослідження для цін акцій компанії «Укранафта» була побудована модель АР(1). На основі цієї моделі був побудований прогноз на п’ять останніх вимірів, і його середня абсолютна похибка для найкращої моделі склала 0,9789.Для моделювання відхилення була побудована модель АРУГ(3). На основі обох моделей був побудований прогноз на 10 наступних вимірів з урахуванням можливого відхилення.

Література

1. Кевин Коннолли. Покупка и продажа волатильности / Коннолли Кевин: [пер. з англ. М. Чекулаєв]. –Москва: ИК «Аналитика», 2007. –243 с.
2. Бідюк П.І. Економетричний аналіз часових рядів: конспект лекцій / Петро Іванович Бідюк. –Київ НТУУ «КПИ», 2007. – 250 с.
3. Ставицький А.В. Навчально-методичний комплекс з курсів „Прогнозування” та „Фінансове прогнозування”/ А.В.Ставицький. – К.:Центр учб. літ., 2006. – 107 с.
4. Носко В.П. Эконометрика. Введение в регрессионный аналіз временних рядов/ В.П.Носко. –Москва: «Литкон», 2002. – 273 с.
5. Грешилов А. А.Математические методы построения прогнозов / А.А.Грешилов, В.А. Стакун, А.А. Стакун. – Москва: «Радио и связь», 1997. – 112 с.

УДК 004.02

Моделювання і прогнозування ВВП України за допомогою регресійних моделей

К.В. Богун

Науковий керівник: д.т.н., проф. П.І. Бідюк

Вступ

Актуальним питанням на сьогоднішній день залишається прогнозування макроекономічних показників, а саме ВВП. Оскільки цей показник показує сукупність всіх вироблених товарів і послуг за рік всередині країни, а також статистику зростання чи зниження економічного розвитоку держави загалом. Значення ВВП важливе як для наших кредиторів так і для інвесторів. За допомогою регресійного аналізу та прогнозу можливо визначити подальший економічний розвиток та чинники, що на нього впливають.

Аналіз останніх досліджень і публікацій. Аналіз динаміки внутрішнього валового продукту, а також залежність його від інших макроекономічних показників не є новим, оскільки досить багато авторів присвятили свої праці даній темі: Р. Мадяр, В. Геєця, А.Храмов, І. Багратяна, І.Крючкова, Н. Фатюга, О. Васильєв. [1-5]. Проте дані роботи мають описовий характер і демонструють залежність ВВП від інших показників, що не дає можливості зробити прогноз величини валового внутрішнього продукту України.

Постановка задачі

Розробити методологію моделювання ВВП України, побудувати математичні моделі для опису динаміки ВВП і використати

побудовані моделі для прогнозування величини ВВП на найближчі три роки.

Регресійні рівняння для моделювання ВВП

Одним із суттєвих показників ефективності економіки є валовий внутрішній продукт, динаміка якого залежить від багатьох чинників. Актуальним виступає аналіз динаміки ВВП України й визначення суттєвих чинників зовнішнього впливу.

Проте постає необхідність обґрунтування динаміки ВВП залежно від чинників зовнішнього впливу із застосуванням математичних методів аналізу.

Необхідність використання моделювання визначається тим, що багато об'єктів досліджувати або неможливо, або це вимагає багато часу і коштів. Але величезні масштаби економічної системи, розгалуженість зв'язків між елементами, відома інерційність зумовлюють майбутній її стан попереднім, що дає можливість передбачити розвиток системи з великою точністю.

Кожне економічне явище функціонує в часі, тому моделювання проводилося на основі часових рядів з використанням регресійних рівнянь для формалізації та побудови прогнозу, динаміки ВВП (показника економічного розвитку країни, дослідивши який можна дати оцінку економічного стану, тобто економічної системи).

Структура різницевого рівняння така, що воно дозволяє виконувати прогнозування на один крок (один період дискретизації вимірів) без додаткових перетворень. Тобто в праву частину необхідно підставити минулі значення змінних і обчислити оцінку прогнозу головної змінної в лівій частині. Але для того щоб знайти оцінку прогнозу на більше число кроків, необхідно застосувати деякі

попередні перетворення різницевих рівнянь (РР). Розглянемо деякі можливі підходи до обчислення прогнозованих значень.

Як приклад, розглянемо рівняння АР(1):

$$y(k) = a_0 + a_1 y(k-1) + \varepsilon(k), \qquad E[\varepsilon(k)] = 0. \tag{1}$$

Збільшимо незалежну змінну час на одну одиницю і запишемо рівняння знову:

$$y(k+1) = a_0 + a_1 y(k) + \varepsilon(k+1). \tag{2}$$

Якщо коефіцієнти a_0, a_1 відомі, то можна знайти умовне математичне сподівання на основі відомої інформації до моменту k включно:

$$E_k[y(k+1)] = E_k[y(k+1) \mid y(k), y(k-1), \ldots, \varepsilon(k), \varepsilon(k-1), \ldots] =$$

$$= a_0 + a_1 E_k[y(k)] = a_0 + a_1 y(k), \tag{3}$$

оскільки $y(k)$ в момент k є відомою константою.

По аналогії запишемо рівняння для моменту $k+2$:

$$y(k+2) = a_0 + a_1 y(k+1) + \varepsilon(k+2), \tag{4}$$

і знайдемо умовне математичне сподівання:

$$E_k[y(k+2)] = a_0 + a_1 E_k[y(k+1)] = a_0 + a_1 E_k[a_0 + a_1 y(k)] =$$

$$= a_0 + a_0 a_1 + a_1^2 y(k). \tag{5}$$

Для наступного моменту часу маємо:

$$E_k[y(k+3)] = a_0 + a_0 a_1 + a_0 a_1^2 + a_1^3 y(k). \tag{6}$$

Таким чином, для загального випадку прогнозування на s кроків можна записати:

$$E_S[y(k+s)] = a_0 \left(\sum_{i=0}^{S-1} a_1^i \right) + a_1^S y(k) = a_0 \sum_{i=0}^{S-1} a_1^i + a_1^S y(k). \tag{7}$$

Отримане рівняння називають функцією прогнозування на довільне число кроків. Прогноз представляє собою збіжний процес, якщо $|a_1|<1$, тобто

$$\lim_{s\to\infty} E_k[y(k+s)]=\frac{a_0}{1-a_1}, \tag{8}$$

де a_1 – знаменник геометричної прогресії. Вираз свідчить про те, що для будь-якого стаціонарного процесу АР чи АРКС оцінка умовного прогнозу асимптотично $(s\to\infty)$ збігається до безумовного середнього.

Знайдемо похибку прогнозування:

$$f_k(s)=y(k+s)-E_k[y(k+s)]. \tag{9}$$

Похибка прогнозу на один крок:

$$f_k(1)=y(k+1)-E_k[y(k+1)]=a_0+a_1y(k)+\varepsilon(k+1)-a_0-a_1y(k)=\varepsilon(k+1) \tag{10}$$

Похибка прогнозу на два кроки:

$$f_k(2)=y(k+2)-E_k[y(k+2)]=$$
$$=a_0+a_1[a_0+a_1y(k)+\varepsilon(k+1)]+\varepsilon(k+2)-E_k[y(k+2)]=$$
$$=a_0+a_0a_1+a_1^2y(k)+a_1\varepsilon(k+1)+\varepsilon(k+2)-a_0-a_0a_1-a_1^2y(k)=$$
$$=\varepsilon(k+2)+a_1\varepsilon(k+1). \tag{11}$$

Таким чином, можемо записати вираз для похибки для довільного числа кроків прогнозування як :

$$f_k(s)=\varepsilon(k+s)+a_1\varepsilon(k+s-1)+a_1^2\varepsilon(k+s-2)+\ldots+a_1^{S-1}\varepsilon(k+1). \tag{12}$$

Враховуючи те, що $E[f_k(s)]=0,$ оцінка прогнозу, яка обчислюється за виразом (12), є незміщеною. Дисперсія похибки прогнозування

$$Var[f_k(s)]=\sigma^2[1+a_1^2+a_1^4+a_1^6+\ldots+a_1^{2(S-1)}], \tag{13}$$

тобто дисперсія є функцією s. Асимптотичне значення дисперсії похибки прогнозу для стаціонарного процесу:

$$\lim_{S\to\infty} Var[f_k(s)] = \frac{\sigma^2}{1-a_1^{\,2}}, \qquad (14)$$

де $a_1^{\,2}-$ знаменник геометричної прогресії.

При побудові математичної моделі у вигляді різницевих рівнянь необхідно вибрати кращу модель з моделей-кандидатів. Оскільки процес оцінювання параметрів рівняння – це випадковий процес, то для підвищення надійності правильного вибору моделі необхідно використовувати, по можливості, більше статистичних параметрів, що характеризують адекватність моделі, які наведено нижче.

- Коефіцієнт множинної детермінації R^2

Це величина, яка визначається відношенням дисперсії змінної, оціненої за отриманою моделлю, до дисперсії виміру цієї змінної:

$$R^2 = \frac{Var(\hat{y})}{Var(y)}\ . \qquad (15)$$

Для адекватної моделі значення чисельника відношення (15) буде наближатися до значення знаменника, а тому для кращої моделі коефіцієнт детермінації буде мати значення найближче до одиниці (знизу),тобто $R^2 \to 1.$

- Сума квадратів похибок

$$\sum e^2(k) = SSE = \sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2\ , \qquad (16)$$

де $y(k)-$ вимір даних для моменту часу k ;

$\hat{y}(k)-$ оцінка виміру, знайдена за допомогою отриманої математичної моделі;

$N-$ довжина вибірки.

Очевидно, що необхідно вибирати ту модель, для якої значення цієї суми приймає найменше значення.

- Статистика Дарбіна-Уотсона DW визначається за виразом:

$$DW = \frac{\sum_{k=2}^{N}[e(k) - e(k-1)]^2}{\sum_{k=1}^{N} e^2(k)} \quad . \tag{17}$$

Тут використана сума перших різниць, тому що вона краще відображає наявність або відсутність кореляції; статистика змінюється в інтервалі: $0 \le DW \le 4$.

При повній відсутності кореляції між похибками e(k) статистика $DW = 2$, тобто це найбільш прийнятне значення цього параметра.

А також використано параметри для оцінювання якості прогнозу, наведені нижче.

- Середня абсолютна похибка в процентах (САПП) – це середнє абсолютних значень похибок оцінок прогнозу в процентах відносно фактичного значення показника:

$$\text{САПП} = \frac{1}{N}\sum_{k=1}^{N}\frac{|y(k)-\hat{y}(k)|}{|y(k)|} \times 100\%. \tag{18}$$

Оскільки ця міра характеризується відносну якість прогнозу, то її використовують, в основному, для порівняння точності прогнозів різновид них об’єктів(процесів) прогнозування. Однак, вона є завжди корисною при виконанні порівняльного аналізу якості прогнозування одного і того ж процесу різними методами, оскільки відносна міра є чіткою і зрозумілою.

Таблиця 1. Інтерпретація типових значень критерію САПП

САПП,%	Інтерпретація
<10	Висока точність

10-20	Хороша точність
20-30	Задовільна точність
>30	Незадовільна(неприйнятна) точність

Для побудови моделі ми використовували дані ВВП, індексу споживчих цін та індекс зайнятості, які були взяті з електронного джерела Державної служби статистики України та Світового банку.

В роботі розглянуто опис моделі ВВП за допомогою авто-регресійних рівняння різних ступенів, авто регресійні рівняння з ковзним середнім, авто регресійних рівняння з ковзним середнім з екзогенним фактором.

Продемонструємо моделі з найкращими результатами.

Розглянемо характеристики моделі вищого порядку. Нижче наведені результати оцінювання моделі AP(7). Вибрані статистичні параметри, які характеризують якість моделі зображені на рисунку 1.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	68.03756	25.87343	2.629630	0.0208
Y(-1)	0.697926	0.200134	3.487300	0.0040
Y(-6)	1.459056	0.900440	1.620381	0.1291
Y(-7)	-2.154600	1.098141	-1.962043	0.0715

R-squared	0.863419	Mean dependent var	101.1512
Adjusted R-squared	0.831901	S.D. dependent var	56.25286
S.E. of regression	23.06362	Akaike info criterion	9.316714
Sum squared resid	6915.097	Schwarz criterion	9.512764
Log likelihood	-75.19207	Hannan-Quinn criter.	9.336202
F-statistic	27.39396	Durbin-Watson stat	2.179179
Prob(F-statistic)	0.000007		

Рисунок 1 – Результати оцінювання моделі AP (7)

$GDP=a_0+a_1*gdp(k-1)+a_6*gdp(k-6)+a_7*gdp(k-7)+e(k) =$

$=68,03756+0,695926*gdp(k-1)+1,459056*gdp(k-6)-$

$- 2,154600*gdp(k-7)+e(k).$

Розглянемо можливість описання ВВП а за допомогою моделі ARIMA. Характеристика моделі ARIMA(1,3) наведені нижче. Вибрані статистичні параметри, які характеризують якість моделі зображені на рисунку 2.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	362.6741	1827.111	0.198496	0.8448
AR(1)	0.983778	0.105009	9.368486	0.0000
MA(3)	0.091002	0.246168	0.369675	0.7157

R-squared	0.845141	Mean dependent var	87.40136
Adjusted R-squared	0.828840	S.D. dependent var	52.78024
S.E. of regression	21.83600	Akaike info criterion	9.131121
Sum squared resid	9059.405	Schwarz criterion	9.279899
Log likelihood	-97.44233	Hannan-Quinn criter.	9.166168
F-statistic	51.84597	Durbin-Watson stat	1.943023
Prob(F-statistic)	0.000000		

Рисунок 2 – Результати оцінювання моделі ARIMA(1,3)

$GDP=a_0+a_1*gdp(k-1)+ma_3*gdp(e-3)+e(k) =$

$=362,6741+0,983778*gdp(k-1)+0,091002*\ e(k-3)+e(k).$

Характеристика моделі ARIMA(1,6) наведені нижче. Вибрані статистичні параметри, які характеризують якість моделі зображені на рисунку 3

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	55.52562	19.74949	2.811496	0.0132
AR(1)	0.784417	0.149436	5.249187	0.0001
MA(1)	0.779425	0.503608	1.547682	0.1425
MA(2)	-1.266162	0.483446	-2.619034	0.0194
MA(3)	0.537453	0.430888	1.247314	0.2314
MA(4)	0.366817	0.524353	0.699561	0.4949
MA(5)	1.233002	0.476862	2.585656	0.0207
MA(6)	1.052136	0.664358	1.583688	0.1341

R-squared	0.966695	Mean dependent var	89.33174
Adjusted R-squared	0.951152	S.D. dependent var	52.39117
S.E. of regression	11.57927	Akaike info criterion	8.004518
Sum squared resid	2011.191	Schwarz criterion	8.399472
Log likelihood	-84.05195	Hannan-Quinn criter.	8.103848
F-statistic	62.19680	Durbin-Watson stat	1.660721
Prob(F-statistic)	0.000000		

Рисунок 3 – Результати оцінювання моделі ARIMA(1,6)

$$GDP= a_0+a_1*gdp(k-1)+ ma_1* e(k-1)+ ma_2* e(k-2)+ ma_3*e(k-3)+$$
$$ma_4* e(-4)+ ma_5* e(-5)+ ma_6* e(-6)+ e(k) =$$
$$=55,52562+0,784417*gdp(k-1)+0,779425* e(k-1)– 1,266162*e(k-2)+$$
$$+0,537453*e(k-3)+0,366817* e(k-4)+1,233002* e(k-5)+$$
$$+1,052136*e(k-6)+ e(k)$$

Характеристики моделі ARIMA(6,1) наведені нижче. Вибрані статистичні параметри, які характеризують якість моделі зображені на рисунку 4

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	76.67986	13.57614	5.648134	0.0002
AR(1)	1.436646	0.359344	3.997974	0.0025
AR(2)	-0.582598	0.560839	-1.038798	0.3234
AR(3)	0.198762	0.586295	0.339014	0.7416
AR(4)	-0.104146	0.589848	-0.176564	0.8634
AR(5)	0.396203	0.568941	0.696386	0.5021
AR(6)	-0.582160	0.332404	-1.751365	0.1104
MA(1)	-0.999618	0.440577	-2.268882	0.0467

R-squared	0.910896	Mean dependent var	98.31778
Adjusted R-squared	0.848523	S.D. dependent var	55.88157
S.E. of regression	21.74910	Akaike info criterion	9.298124
Sum squared resid	4730.235	Schwarz criterion	9.693845
Log likelihood	-75.68312	Hannan-Quinn criter.	9.352689
F-statistic	14.60408	Durbin-Watson stat	2.074529
Prob(F-statistic)	0.000163		

Рисунок 4 – Результати оцінювання моделі ARIMA(6,1)

$$GDP= a_0+a_1*gdp(k-1)+ a_2*gdp(k-2)+ a_3*gdp(k-3)+ a_4*gdp(k-4)+$$
$$a_5*gdp(k-5)+ a_6*gdp(k-6)+ ma_1* e(k-1)+e(k) =$$
$$=76,67986+1,436646*gdp(k-1)-0,582598*gdp(k-2)+0,198762*gdp(k-3)-$$
$$-0,104146*gdp(k-4)+0,3396203*gdp(k-5)-0,582160*gdp(k-6)-$$
$$-0,999618* e(k-1)+ e(k).$$

Результати моделювання на три кроки, наведені нище в таблиці 1, свідчать про те, що всі моделі є придатними для моделювання,

оскільки коефіцієнт Тейла вимірюється сотими долями. Але у деяких моделях САПП більше 5, що свідчить про те, що у таких випадках потрібно покращувати структуру моделі.

Порівнюючі побудовані моделі, найкращі характеристи в моделях APIMA(1,6) та APIMA(3,3), так як у першої моделі кращі кофефіцієнти множинної детермінації = 0,966695,коєфіцієнтк Акайке = 8.004518 і сума квадратів похибок = 2011.191, що є у два рази краще від моделі APIMA(3,3), але статистика Дарбіна-Уотсона рівна 1,660721.

Тип моделі	Характеристики моделі				Характеристики прогнозу		
	R^2	Akaike	$\sum e^e(k)$	DW	СКП	САПП	Коеф-т Тейла
ARMAX+ICЦ	0.861030	9.631469	7377.469	2.2419	9,803565	6,423132	0,031532
AR (6)	0.867350	9.584922	7041.939	2.114848	6,258251	4,147031	0,020394
AR (7)	0.863419	9.316714	6915.097	2.179179	9,201781	7,944902	0,038579
ARIMA(1, 3)	0.867178	8.953026	8020.643	1.956997	12,00931	1,198774	0,008543
ARIMA(1, 6)	0.966695	8.004518	2011.191	1.660721	4,1711349	2,667609	0,013199
ARIMA(3,3)	0.919992	8.928393	4761.756	2.052934	7,309842	4,885511	0,023379
ARIMA(6,1)	0.910896	9.298124	4730.235	2.074529	1.308944	0.756722	0.004171

Рис.1. Зведені характеристики математичних моделей, побудовані для прогнозу ВВП України.

Результати моделювання на три кроки, наведені вище, свідчать про те, що всі моделі є придатними для моделювання, оскільки коефіцієнт Тейла вимірюється сотими долями. Але в деяких моделях

САПП більше 5, що свідчить про те, що у таких випадках потрібно покращувати структуру моделі.

Порівнюючи побудовані моделі, найкращі характеристики в моделях ARIMA(1,6) та ARIMA(3,3), оскільки у першої моделі кращі кофефіцієнти множинної детермінації = 0,966695, коєфіцієнт Акайке = 8.004518 і сума квадратів похибок = 2011.191, що є у два рази краще від моделі ARIMA(3,3), але статистика Дарбіна-Уотсона рівна 1.660721 .

Найкраща модель для оцінювання прогнозу ARIMA(6,1), оскільки середня абсолютна похибка в процентах дорівнює 1,308944, а коєфіцієнт Тейла 0,004171.

Висновки

В статті розроблена методологія економетричного моделювання динаміки ВВП України. Знайдено оптимальну модель для прогнозування. Розглянувши ВВП як головний показник соціально-економічного розвитку нашої країни та можливості реструктуризації економіки, можна зробити висновок про те, що він відіграє надзвичайно важливу роль у визначенні економічного здоров'я суспільства. Прогнозуючи його ми можемо визначити потенціал країни в цілому.

У подальших дослідженнях планується комбіноване використання регресійних моделей і методів інтелектуального аналізу даних. Зокрема, динамічних байєсівських мереж, нейро-нечітких моделей та імунного підходу до моделювання і прогнозування.

Література

1. Мадяр Р.О. Пріоритетні заходи з стабілізації обсягів ВВП та прогнозування його динаміки [Електронний ресурс]. – Режим

доступу: https://irbis-nbuv.gov.ua/cgibin/irbis_nbuv/cgiirbis_64.exe?C21CO

2. Іванюк М. М., Підгородецька В. М. Модель динаміки внутрішнього валового продукту зі сталими екзогенними величинами // Вісник Хмельницького національного університету. – 2006. – Т. 1. – № 1. – С. 158 – 163.

3. Геєць В.М., Клебанова Т.С., Черняк О.І., Іванов В.В., Дубровіна Н.А., Ставицький А.В. Моделі і методи соціально-економічного прогнозування. - Харків: ВД «ІНЖЕК», 2005.

4. Крючкова І. В. Диспропорції валового наявного доходу України та їхній вплив на динаміку ВВП // Економіка і прогнозування. – 2013. – № 3. – С. 21-37.

5. Васильєв О. Синергетична модель прогнозування динаміки ВВП України // Банківська справа. – 2012. – № 1. – С. 59-67.

6. Храмов А. В., Миннуллин А. А., Нуруллин Н. Н., Кадочникова Е. И. Эконометрический анализ валового внутреннего продукта России // Молодой ученый. — 2014. — №21. — С. 452-454.

7. Фатюха Н.Г., Аналіз динаміки ВВП України/ Н.Г.Фатюха, Т.П.Макушина//Ефективна економіка [Електронний ресурс]. – Режим доступу:http://www.economy.nayka.com.ua/pdf/12_2015/30.pdf

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

9. Горидько Н.П. Современный экономический рост: теория и регрессионный анализ / Горидько Н.П., Нижгородцев Р.М. – Новочеркасск: «НОК», 2011. – 343с.

10. Бідюк П.І. Аналіз часових рядів / Бідюк П.І.,Романенко В.Д., Тимощук О.Л. – К: Політехніка, 2010. – 317 с.

УДК 519.87

АНАЛІЗ КРЕДИТНИХ РИЗИКІВ РЕГРЕСІЙНИМИ МЕТОДАМИ

Бойко О.М., Бідюк П.І.

Вступ

Для того, щоб ефективно управляти кредитними ризиками, необхідно вміти його оцінюватиз максимально можливою точністю. Існує досить широка множина методів і моделей оцінювання кредитного ризику. При створенні системи управління кредитним ризиком банки спираються на власний досвід та напрацювання.

На даний момент існує безліч моделей оцінки кредитоспроможності на основі ринкових показників, а саме: модель Блека-Шоулза, модель Мертона[1, 2]. Перевагами таких моделей є висока прогнозуюча спроможність, абсорбуюча інформація про позичальника, доступна всім інвесторам, чинним на ринку. Недоліками є те, що інформація про позичальника буде показовою лише за умови ефективності ринку, необхідний великий масив даних. Методи реалізовані на практиці у вигляді програмного продукту CreditMonitor.

Підходи до розробки моделей оцінки кредитоспроможності на основі фундаментальних показників базуються на макроекономічних показників, фінансових показників та на даних рейтингових агентств. Особливості таких підходів є те, що вони враховують циклічність економіки, дають довгострокову оцінку, підходять для крос-аналізу. Основні переваги: доступність інформації, простота підрахунків і точність прогнозу. Проте присутні недоліки: важко визначити періодичність циклів економіки та оцінити ймовірність дефолту конкретного позичальника; не завжди надані дані достовірні;бухгалтерська звітність показує результати постфактум, що володіє недостатньою передбачувальною силою щодо майбутніх перспектив; переоцінка рейтингу відбувається з часовим лагом [1,3].

Для мінімізації кредитного ризику потрібно управляти ними, що представляє собою процес виявлення і оцінки ризиків, а також вибір методів та інструментів для цього. Традиційно кредитний ризик розглядається в розрізі кожного конкретного позичальника. Існує велика кількість моделей, що використовують складний математичний апарат для оцінки кредитного ризику. Ключовою задачею побудови математичної моделі кредитного ризику є

оцінювання розподілу збитків всього агрегованого кредитного портфеля. Це так звані системи скорингу [4,5].

Дослідження присвячене аналізу можливості застосування лінійної і нелінійної регресії, а також дерев рішень та Байєсівських мереж; виконанню та аналізу результатів обчислювальних експериментів з оцінювання кредитоспроможності клієнтів за заданими статистичними даними; порівняння результатів застосованих методів стосовно розв'язання задачі оцінювання кредитів.

Постановка задачі

У роботі ставиться мета оцінювання кредитоспроможності позичальників кредитів фінансових установ з використанням математичних моделей. Для досягнення цієї мети необхідно розв'язати такі задачі: виконати огляд математичних моделей для оцінювання кредитного ризику і вибрати типи моделей для подальшого використання; зібрати статистичні дані, необхідні для використання обчислювальних експериментів; побудувати математичні моделі для оцінювання кредитоспроможності клієнтів і виконати порівняльний аналіз їх характеристик на основі наявних вибірок даних.

Огляд моделей для оцінювання кредитного ризику

Традиційними і найбільш поширеними є регресійні методи, насамперед лінійна багатофакторна регресія:

$$P = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n,$$

де p – ймовірність дефолту; w – вагові коефіцієнти; x – характеристики клієнта.

Недолік даної моделі полягає у тому, що в лівій частині рівняння знаходиться імовірність, яка приймає значення від 0 до 1, а змінні в правій частині можуть приймати будь-які значення від $-\infty$ до $+\infty$.

Нелінійна модель у формі логістичної регресії дозволяє подолати цей недолік:

$$\ln\left(\frac{p}{1-p}\right) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n.$$

Нас цікавить ймовірність появи події залежно від значень параметрів $x = \{x_1, x_2, \ldots, x_n\}$. Виходом є значення логіт-функції,яка приймає значення від 0 до 1, тобто:

$$g(z) = \frac{1}{1 + e^{-z}}.$$

Тоді модель матиме вигляд:

$$Pr = g\left(p_0 + p_1 x_1 + p_2 x_2 + \ldots + p_m x_m\right).$$

Припустимо тепер, що ми маємо вибірку:

$$\left(x_1^i, x_2^i, \ldots, x_m^i, y^i\right) \left(i = 1, \ldots, n\right),$$

де y^i – значення залежної змінної; n – кількість спостережень.

Потрібно оцінити коефіцієнти моделі. Для цього скористаємося принципом максимальної правдоподібності. Відповідного до цього принципу, за оцінки коефіцієнтів приймаються ті значення, які дають максимум функції правдоподібності, що має такий вигляд у нашому випадку:

$$L(\overline{p}) = \prod_{i=1}^{n} g\left(\overline{px^1}\right)^{y^i} \left[1 - g\left(\overline{px^1}\right)\right]^{1-y^i},$$

Для зручності позначимо:

$$\overline{p} = \left(p_0, p_1, p_2, \ldots, p_m\right),$$
$$\overline{x^1} = \left(x_1^i, x_2^i, \ldots, x_m^i\right),$$
$$\overline{px^1} = p_0 + p_1 x_1^i + p_2 x_2^i + \ldots + p_m x_m^i.$$

Зазвичай замість функції (наведена вище) використовується її логарифм, що не змінює суті задачі, але дозволяє позбавитись від добутку.

$$l(\overline{p}) = \sum_{i=1}^{n} y^i \, ln\left(g\left(\overline{px^1}\right)\right) + \left(1 - y^i\right) ln\left(1 - g\left(\overline{px^1}\right)\right).$$

Очевидно, функція має максимум. Для знаходження значень коефіцієнтів, при яких функція досягає максимуму, розраховуються часткові похідні та прирівнюються нулю, в результаті чого отримуємо систему рівнянь:

$$f(\overline{p}) = \begin{cases} \sum_{i=1}^{n} g\left(\overline{px^1}\right) - \sum_{\{i:y^i=1\}} 1 = 0, \\ \sum_{i=1}^{n} g\left(\overline{px^1}\right) - \sum_{\{i:y^i=1\}} x_j^i = 0 \qquad (j = 1,\ldots,m). \end{cases}$$

Для розв'язання даної системи скористаємося методом Ньютона-Расфона. Метод передбачає початкове наближення та наступне його покращення. Для вибору найкращого початкового наближення оцінимо ймовірність появи «1» за наступною формулою:

$$Pr = \frac{\sum_{\{i:y^i=1\}} 1}{n}.$$

Тобто кількість разів появи одиниці поділили на загальну кількість спостережень.

Початкове наближення для параметрів моделі отримаємо, якщо не будемо враховувати вплив факторів взагалі, тобто:

$$\overline{p^0} = \left(ln\left(\frac{Pr_0}{1 - Pr_0}\right), 0, \ldots, 0 \right).$$

Використано той факт, що для логіт-функції виконується тотожність:

$$z \equiv ln \frac{g(z)}{1 - g(z)}.$$

Послідовне наближення будемо обраховувати згідно формули, яка зв'язує наступне (q+1) з попереднім q-наближеним:

$$\overline{p}^{q+1} = \overline{p}^{q} - F\left(\overline{p}^{q}\right)J^{-1}\left(\overline{p}^{q}\right).$$

У даному співвідношенні використовується $F(\overline{p}^{\,q}) = (f_0(\overline{p}^{\,q}), f_1(\overline{p}^{\,q}), \ldots, f_m(\overline{p}^{\,q}))$, а також матриця Якобі.

Зауважимо, що замість того, щоб обраховувати обернену матрицю, яка фігурує і необхідна для обрахунку величини $\Delta = J^{-1}\left(\overline{p}^{q}\right)F\left(\overline{p}^{q}\right)$, ефективніше методом виключень Гауса розв'язувати лінійну систему рівнянь:

$$\Delta J\left(\overline{p}^{q}\right) = F\left(\overline{p}^{q}\right).$$

Для розв'язання системи використовується метод Гауса з вибором головного елемента по всій матриці коефіцієнтів. Розрахунки закінчуються, коли на поточному кроці норма вектора Δ буде менше деякої заданої величини, що характеризує точність обрахунків. Норма вектора Δ характеризує різницю між сусідніми наближеннями [8].

Для застосування логістичної регресії необхідні набагато більш складні розрахунки для отримання вагових коефіцієнтів. Але в даний час логістична регресія є лідером скорингових систем.Перевага логістичної регресії ще й у тому, що вона може поділяти клієнтів як на дві групи (0 – поганий, 1 – хороший), так і на кілька груп (1, 2, 3, 4 групи ризику).

Дерева рішень – це модель, що будується на логічному ланцюжку правил, які намагаються описати окремі взаємозв'язки між даними щодо очікуваного результату. Структура дерев рішень відкрито показує аргументацію правил і тому дозволяє легко зрозуміти процес прийняття рішення [6].

Існує множина критеріїв, які визначають якість побудованої моделі і якість прогнозу. Нижче наведено деякі з них, які використано в даній статті.

1) Інформаційний критерій Акайке

Критерій використовується для порівняння моделей з різним числом параметрів, коли потрібно вибрати найкращий набір пояснюючих змінних.

Для лінійної моделі множинної регресії значення критерію розраховується за такою формулою:

$$AIC = ln\left(\frac{\sum_{i=1}^{k}\varepsilon_i^2}{n}\right) + \frac{2k}{n},$$

де n – кількість спостережень; k – кількість параметрів моделі, $\sum_{i=1}^{k}\varepsilon_i^2$ – сума квадратів залишків моделі, отримана при оцінці коефіцієнтів моделі методом найменших квадратів.

При збільшенні кількості пояснюючих змінних перший доданок у правій частині зменшується, а другий збільшується. Таким чином, критерій не тільки винагороджує за якість наближення, але і штрафує за використання зайвої кількості параметрів моделі. Серед кількох альтернативних моделей перевага віддається тій, у якої значення AIC менше.

2) Інформаційний критерій Шварца

Цей критерій, аналогічно критерію Акайке, дозволяє порівняти моделі з різноюкількістю параметрів, коли потрібно вибрати кращумножину пояснюючих змінних.Для лінійної моделі множинної регресії значення критерію розраховується за такою формулою:

$$SC = ln\left(\frac{\sum_{i=1}^{k}\varepsilon_i^2}{n}\right) + \frac{k\,ln(n)}{n},$$

де n – кількість спостережень; k – кількість параметрів моделі; $\sum_{i=1}^{k}\varepsilon_i^2$ – сума квадратів залишків моделі, отриманих при оцінюванні коефіцієнтів моделі за методом найменших квадратів.

При збільшенні кількості пояснюючих змінних перший доданок у правій частині зменшується, а другий збільшується. Серед кількох альтернативних моделей перевага віддається тій, у якої значення SC менше.

3) Коефіцієнт детермінації

Коефіцієнт детермінації розглядають, як правило, в якості основного показника, що відображає міру якості регресійної моделі, яка описує зв'язок між залежною і незалежними змінними моделі. Коефіцієнт детермінації показує, яка частка варіації пояснюючої змінної y врахована в моделі і обумовлена впливом на неї факторів, включених в модель:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})}{\sum_{i=1}^{n}(y_i - \bar{y})},$$

де y_i – значення спостережуваної змінної; $\bar{y}$ – середнє значення по спостережуваних даних; $\hat{y}$ – модельні значення, побудовані за оціненими параметрами.Чим ближче R^2 до 1 , тим вища якість моделі.

4) Стандартна похибка

Стандартна похибка регресії (стандартна похибка оцінки) розглядається в якості міри розкиду даних спостережень від змодельованих значень. Стандартна похибка регресії розраховується як квадратний корінь з незміщеної оцінки дисперсії регресії :

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{1}{(n-2)}\left[\sum_{i=1}^{n}(y_i - \bar{y}) - \frac{\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]},$$

де n – загальне число спостережень; y_i – значення змінної, що спостерігається; x_i – значення пояснюючої змінної; $\bar{y}$ – середнє значення спостережуваної змінної за вибіркою; $\bar{x}$ – середнє значення пояснюючої змінної за вибіркою; $\hat{\sigma}^2$ – незміщена оцінка дисперсії регресії.Чим менше значення стандартної похибки регресії, тим якість моделі вище.

5) Сума квадратів залишків

Сума квадратів величин розбіжності між змодельованими і фактичними значеннями, які пояснюються змінними на періоді ідентифікації розраховується за такою формулою:

$$\sum_{i=1}^{n}\left(y_i-\hat{y}\right)^2,$$

де y_i – значення змінної, що спостерігається; $\hat{y}$ – модельні значення, побудовані за оціненими параметрами.

6) Статистика Дарбіна-Уотсона

Критерій Дарбіна-Уотсона (чи DW-критерій) — статистичний критерій, що використовується для знаходження автокореляції залишків першого порядку регресійної моделі. Обчислюється за формулою:

$$DW=2-2\rho,$$

де ρ – коефіцієнт автокореляції першого порядку і

$$\rho=\frac{\frac{1}{N-1}\sum_{k=2}^{N}\left[e(k)-\bar{e}\right]\left[e(k-1)-\bar{e}\right]}{\sigma_e^2},$$

$$\sigma_e^2=\frac{1}{N-1}\sum_{k=1}^{N}\left[e(k)-\bar{e}\right]^2,$$

$$e(k)=y(k)-\overline{y(k)}.$$

7) Критерій Ханана-Куїна

Критерій Хана-Куїна (HQ) використовується для порівняння моделей по співвідношенню між якістю підбору і кількістю оцінюваних параметрів. Обчислюється тільки для моделей бінарного, множинного вибору і моделей з урізаними даними.

$$HQ=-2\frac{\ln L}{n}+2*k*\frac{\ln(\ln n)}{n},$$

де L – функція правдоподібності; n – число спостережень; k – кількість змінних в моделі.Вибирається модель з найменшим значенням критерію.

8) Критерій Макфадена

Аналог коефіцієнта детермінації для звичайної регресії:

$$McFaddenRs\ quared\ =1-\frac{\ln L}{Re\ str\ ln L},$$

де $\ln L$ – логарифм функції правдоподібності; $Re\,str\,\ln L$ – залишок логарифма функції правдоподібності. Значення критерію знаходиться в діапазоні (0, 1).

Обчислюється, тільки якщо модель містить константу. Найкращим вважається значення найбільш близьке до одиниці.

9) Середньоквадратична похибка

Середньоквадратична похибка використовується для оцінки адекватності моделі і обчислюється за формулою:

$$СКП = \sqrt{\frac{1}{n}\sum_{k=1}^{N}\left[y(k) - \hat{y}(k)\right]^2},$$

де $y(k)$ – значення змінної, що спостерігається; $\hat{y}(k)$ – модельні значення, побудовані за оціненими параметрами.

10) Середня абсолютна похибка у процентах

Середня абсолютна похибка в процентах (САПП) – це середнє абсолютних значень похибок оцінок прогнозу в процентах відносно фактичного значення показника:

$$САПП = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|y(k) - \hat{y}(k)\right|}{\left|y(k)\right|}100\%,$$

де $y(k)$ – значення змінної, що спостерігається; $\hat{y}(k)$ – модельні значення, побудовані за оціненими параметрами.

Оскільки ця міра характеризує відносну якість прогнозу, то її використовують, в основному, для порівняння точності прогнозів різнорідних об’єктів (процесів) прогнозування. Однак, вона є завжди корисною при виконанні порівняльного аналізу якості прогнозування одного й того ж процесу різнимиметодами, оскільки відносна міра є чіткою і зрозумілою для дослідника і практичного користувача[7,8].

Приклад застосування методики

Для побудови моделі лінійної регресії, логістичної регресії та дерева рішень було використано статистичні дані першої вибірки –All_1 (довжина вибірки даних складала 15000 значень).

– $x1$ – змінна, яка характеризує стать позичальника (gender);

– $x2$ – змінна, яка характеризує вік позичальника (Age);

– $x3$ – змінна, яка характеризує суму кредиту (Credit_sum);

– $x4$ – змінна, яка характеризує термін кредитування в днях (Term_of_crediting_in day);

– $x5$ – змінна, яка характеризує сімейний стан позичальника (Maritial_status);

– $x6$ – змінна, яка характеризує кількість дітей позичальника (Children);

– $x7$ – змінна, яка характеризує кількість найманих робітників в компанії позичальника (Number_of_employees_in_kompany);

– $x8$ – змінна, яка характеризує дохід позичальника (Income_customer);

– $x9$ – змінна, яка характеризує витрати позичальника (Costs_customer);

– y – змінна, яка характеризує результат повернення кредиту (Result).

Особливістю цієї вибірки є її висока асиметрія стосовно типів позичальників, тобто вона містить характеристики 750 клієнтів, які не повертають кредити , і 14250 клієнтів, які повертають кредити.

Нижче наведено результат побудови лінійної регресії (рис.1).

Equation: UNTITLED Workfile: ALL_1

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 04/30/14 Time: 22:45
Sample: 1 15000
Included observations: 15000

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.965872	0.009212	104.8455	0.0000
X1	0.016636	0.003582	4.644929	0.0000
X2	-0.000539	0.000188	-2.864474	0.0042
X3	7.47E-08	3.17E-08	2.352971	0.0186
X4	-2.23E-09	1.54E-09	-1.442450	0.1492
X5	0.002748	0.001968	1.396726	0.1625
X6	0.003767	0.002594	1.452080	0.1465
X7	-0.005106	0.001223	-4.176210	0.0000
X8	-6.93E-08	9.26E-08	-0.748356	0.4543
X9	3.50E-07	7.02E-07	0.497868	0.6186

R-squared	0.004825	Mean dependent var	0.950000
Adjusted R-squared	0.004227	S.D. dependent var	0.217952
S.E. of regression	0.217491	Akaike info criterion	-0.212651
Sum squared resid	709.0625	Schwarz criterion	-0.207574
Log likelihood	1604.885	F-statistic	8.074477
Durbin-Watson stat	1.854974	Prob(F-statistic)	0.000000

Рисунок 1 – Результати лінійної регресії та її характеристики

Для вибору кращої моделі із множини різних специфікацій дано значення інформаційних критеріїв Акайке (AIC) та Шварца (SC), логарифмічну функцію правдоподібності. Для визначення міри якості лінійної регресійної моделі показано значення R-квадрат (коефіцієнт детермінації)та скорегований коефіцієнт детермінації, стандартну похибку регресії, статистика Дарбіна-Уотсона,

статистика Фішера (F-статистика) та її ймовірність.Значення стандартної похибки регресії показує, що лінійна модель дає гірші результати ($S.E = 0.217491$)ніж модель логіт ($S.E = 0.217336$). Але критерії Акайке (для лінійної регресії $AIC = -0.212651$, для логіт $AIC = 0.393290$)та Шварца (для лінійної регресії $SC = -0.207574$,для логіт $SC = 0.398368$) дають протилежні результати.

Також було взято іншу вибірку (BASE_ALL) для порівняння з статистичними даними, які описують такі змінні (довжина вибірки даних складала 7568 значень):

– змінна яка характеризує вік позичальника (Age years);
– змінна яка характеризує стать позичальника (Sex);
– змінна яка характеризує освіту позичальника (Education);
– змінна яка характеризує сімейний стан позичальника (Marital status);
– змінна яка характеризує суму кредиту (Credit);
– змінна яка характеризує регіон проживання позичальника(Region);
– змінна яка характеризує сферу, де працює позичальник(PrevEmployerSector);
– змінна яка характеризує статус в суспільстві позичальника(OccupationStatus);
– змінна яка характеризує досвід роботи позичальника(WorkExperienceMonth);
– змінна яка характеризує результат повернення кредиту(Good/Bad).

Для побудови моделі відцифровано дані:

Стать позичальника(Sex):

– Чоловіча(Male) – 1;
– Жіноча(Female) – 0;

Освіта (Education):

– Початкова (Elementary) – 0;
– Середня (Maddle) – 1;
– Вища (High) – 2;
– Дві вищі чи ступінь (Two high and/or degree) – 3

Сімейний стан (Marital status):

– Неодружений (Not married) – 0;

– Одружений (Registering marriage) – 1;
– Вдова(ець) (Widow(er)) – 2;
– Розведений (Divorced) – 3;
– Цивільний шлюб (Civil marriage) – 4

Сфера праці (PrevEmployerSector):

– Невизначена(none) – 0;
– Інші (other) – 1;
– Медицина (Medicine) – 2;
– Сільське господарство (Agriculture) – 3;
– Послуги (Services) – 4;
– Будівництво (Building) – 5;
– Освіта (Education) – 6 ;
– Торгівля (Trade) – 7;
– Виробництво (Manufacturing) – 8;
– Фінанси (Finance) – 9;
– Видобуток (Mining) – 10.

Статус в суспільстві (OccupationStatus):

– Найманий робітник(Employee) – 0;
– Урядовий офіцер (Government officer) – 1;
– Пенсіонер (Pensioner) – 2;
– Власник/співвласник (Owner/coowner) – 3;
– Військовий (Military) – 4.

Результат (Good/Bad):

– Повернено кредит (Good) – 1;
– Не повернено кредит (Bad) – 0.

Узагальнена порівняльна характеристика лінійної моделі та моделі логіт показана в таблиці 1:

Таблиця 1 – Результати лінійної і нелінійної регресії

Варіант розрахунків	Якість моделі			Якість прогнозу		
	Коефіцієнт детермінації	Сума квадратів залишків	Статистика Дарбіна-Уотсона	Середня квадратична похибка	Середня абсолютна похибка	Коефіцієнт Тейла
Лінійна регресія						
All_1	0,004825	709,0625	1,854974	0,217419	9,4637	0,112956
BASE_743	0,384027	114,4168	0,271820	0,392419	34,0567	0,302704
Нелінійна регресія						
All_1	-	708,0545	-	0,217264	9,4463	0,112876
BASE_743	-	106,1273	-	0,377937	28,4419	0,289203

Отже, нелінійна регресія за всіма показниками якості прогнозу дає кращі результати ніж лінійна регресія. Побудована модель нелінійної регресії є кращою за показниками якості.

Побудова моделей дерев рішень за статистичними даними

В результаті використання системи SPSS побудовано дерево рішень з відповідними характеристиками для вибірки All_1.

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	y
	Independent Variables	x9, x8, x6, x7, x5, x1, x4, x2, x3
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	x8, x9, x3, x4, x1, x7
	Number of Nodes	31
	Number of Terminal Nodes	21
	Depth	3

Рисунок 2 – Зведена таблиця про загальні специфікації дерева рішень

За допомогою методу CHAID(CHi-squared Automatic Interaction Detection) отримано результат, який свідчить, що рівень доходу ($x8$) є кращим предикатом кредитного рейтингу.

Для категорії з низьким рівнем доходів, наступним найкращим предикатом є рівень витрат ($x9$). Для малих витрат найкращим предикатом виступає сума

кредиту ($x3$). Оскільки немає ніяких розгалужень, то це термінальний вузол. Клієнти цих вузлів мають досить непоганий кредитний рейтинг (0,69 або 0,78), незважаючи на суму кредиту. Для великих значень витрат немає розгалужень;це термінальний вузол.

Для категорій середнього та високого доходу, наступний найкращим предикатом є термін кредитування ($x4$). При середньому терміні найкращим предикатом виступає сума кредиту ($x3$), для більш великого терміну – стать позичальника ($x1$).

Для клієнтів з високим рівнем доходів наступний найкращим предикатом є термін кредитування ($x4$). Дивлячись від терміну наступним найкращим предикатом є дохід позичальника ($x8$), кількість найманих робітників в компанії ($x7$) або суму кредиту($x3$).

Узагальнені результати по дереву наведено нижче (рис. 3). Оцінка ризику та його стандартна похибка, тобто міра точності прогнозу, отриманого за деревом:

Gain Summary for Nodes

Node	N	Percent	Mean
30	1091	7,3%	1,00
27	1001	6,7%	,99
13	233	1,6%	,99
11	947	6,3%	,98
7	2042	13,6%	,98
21	1912	12,7%	,97
29	68	,5%	,97
15	2737	18,2%	,97
23	507	3,4%	,97
28	176	1,2%	,97
25	417	2,8%	,96
6	108	,7%	,94
20	318	2,1%	,94
24	689	4,6%	,94
22	298	2,0%	,94
9	678	4,5%	,93
26	171	1,1%	,91
12	106	,7%	,90
5	366	2,4%	,86
19	612	4,1%	,79
18	523	3,5%	,69

Growing Method: CHAID
Dependent Variable: y

Risk

Estimate	Std. Error
,043	,001

Growing Method: CHAID

Рисунок 3 – Узагальнені результаті та ризики

Отже, за допомогою дерев рішень отримано ризик неправильної оцінки клієнта 4,3% та стандартне відхилення 0,001.

Узагальнені результати по дереву другої вибірки (BASE_743) наведено нижче. Діаграма дерева (рис.4) являє собою графічне представлення моделі дерева. Ця діаграма показує, що: за допомогою методу CHAID, регіон проживання позичальника ($x6$) є кращим предикатом кредитного рейтингу. Оскільки глибина дерева один, то не відбувається більш ніякого розгалуження.

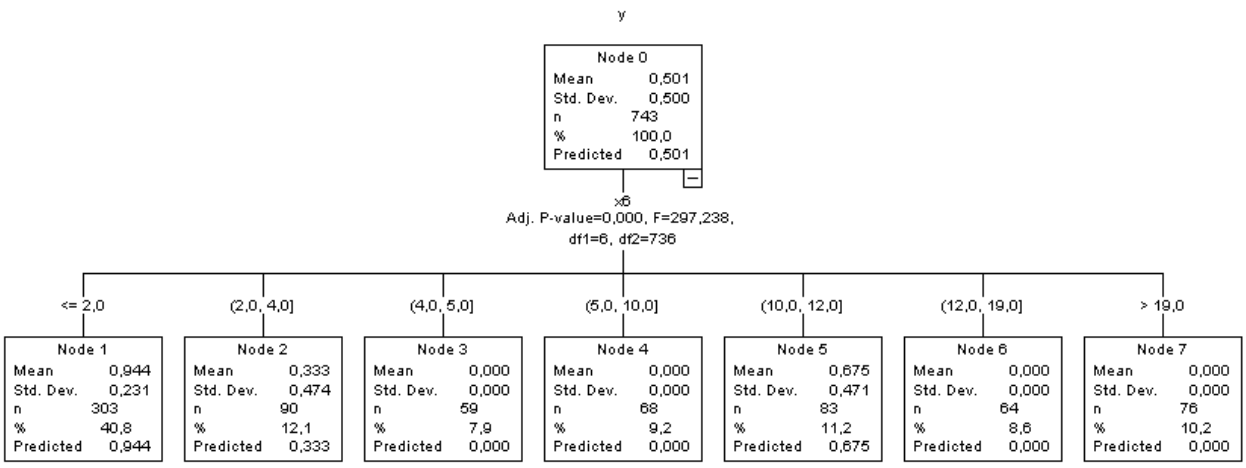


Рисунок 4 – Діаграма дерево рішень

І оцінка ризику та його стандартна помилка, тобто міра точності прогнозу дерева:

Gain Summary for Nodes

Node	N	Percent	Mean
1	303	40,8%	,94
5	83	11,2%	,67
2	90	12,1%	,33
7	76	10,2%	,00
4	68	9,2%	,00
6	64	8,6%	,00
3	59	7,9%	,00

Growing Method: CHAID
Dependent Variable: y

Risk

Estimate	Std. Error
,073	,006

Growing Method: CHAID

Рисунок 5 – Узагальнені результаті та ризики

Отже, за допомогою дерев рішень отримано ризик неправильної оцінки клієнта 7,3%. Та стандартне відхилення 0,06.

Нижче наведено результати ризику неправильної оцінки та стандартне відхилення для двох вибірок (таблиця2).

Таблиця 2 – Оцінка ризику та стандартна похибка дерева рішень трьох статистичних даних

Вибірки	Ризик неправильної оцінки	Стандартне відхилення
1) ALL_1	0,043	0,001
2) BASE_743	0,073	0,006

Порівняльна характеристика логістичної регресії, лінійної регресії та дерев рішень

Таким чином, використано три методи інтелектуального аналізу даних для прогнозування кредитоспроможності фізичних осіб за статистичними даними: лінійна регресія, логістична регресія та дерева рішень. Для виконання обчислювальних експериментів використано вибірку даних для 15000 клієнтів банку та вибірку для 7568 клієнтів. Вибрано множину статистичних критеріїв для аналізу якості класифікації клієнтів.

Кращі результати класифікації клієнтів на дві групи отримано за допомогою дерев рішень(таблиця 3), що можна пояснити можливістю досягнення високої якості класифікації на дві групи за допомогою цього методу.

Таблиця 3 – Стандартне відхилення трьох моделей

Вибірки	Лінійна регресія	Логістична регресія	Дерева рішень
1) ALL_1	0,217419	0,217264	0,001
2) BASE_743	0,392419	0,377937	0,006

Побудова мережі Байєса

Для побудови першої моделі(ALL – 15000 значень) використовувалися статистичні дані по 15000 виданих кредитах, строк яких закінчився. Вибірка була поділена на навчальну (13000 випадків) та перевірочну (200 випадків). Навчальна вибірка завантажується в блок побудови моделі. Для побудови моделі необхідно формалізувати дані у зручному для обробки вигляді, тобто перевести їх у заданий формат, а в разі неперервних змінних – дискретизувати їх. Для дискретизації використано ієрархічна дискретизація.

На наступному кроці виконується співставлення завантаженим характеристикам взаємовиключних змінних і вибір тих змінних, які будуть використовуватись на етапі побудови мережі. Далі вибирається відповідний алгоритм навчання мережі у разі необхідності використовується експертне знання і виконується навчання мережі.

При побудові структури мережі Байєса в програмі GeNIe слід пам'ятати, що обраний алгоритм впливає на швидкість і на саму побудовану структуру. Найшвидшим реально є алгоритм Greedy Thick Thinning, його і будемо використовувати для аналізу прикладів. У результаті роботи алгоритму отримуємо тільки одну структуру, яка є логічною і оптимальною(рис 6).

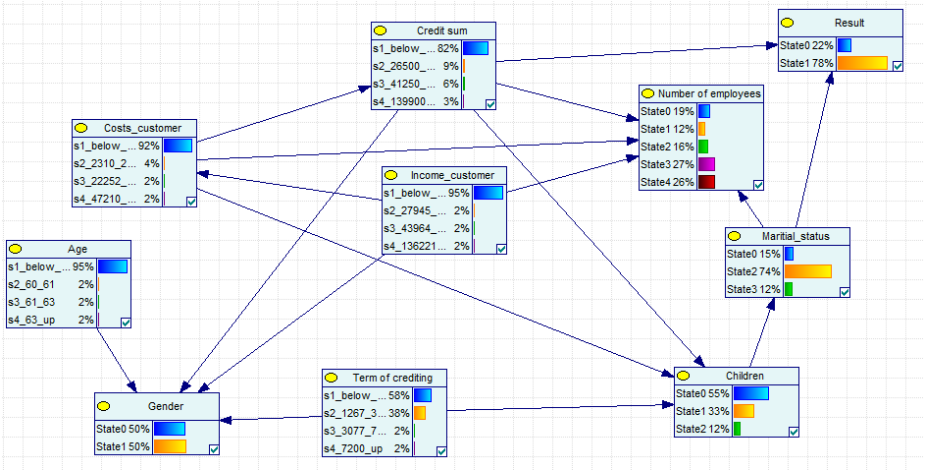


Рисунок 6 – Структура мережа Байєса у вигляді гістограм вибірки ALL

На основі перевірочної вибірки здійснюється перевірка якості моделі з використанням відомих оцінок: загальної точності, помилок першого та другого роду. Для різних порогів відсікання обчислено помилки I-го та II-го роду (табл.4).

Таблиця 4– Загальна точність моделі та помилки I-го та II-го роду для різних рівнів порогу відсікання, отримані для мереж Байєса(ALL)

	Прогноз: Повернення кредиту (0)	**Прогноз: Дефолт (1)**	**Процнт точності**
Cut-off=0,5			
Факт: Повернення кредиту (0)	82	50	0.620
Факт: Дефолт (1)	12	56	0.82
Загальна точність моделі			**0.69**

Cut-off=0,4			
Факт: Повернення кредиту (0)	92	49	0.65
Факт: Дефолт (1)	3	56	0.94
Загальна точність моделі			**0.74**
Cut-off=0,3			
Факт: Повернення кредиту (0)	101	49	0.67
Факт: Дефолт (1)	4	46	0.92
Загальна точність моделі			**0.735**

Найбільша точність моделі досягається на рівні 0,75 при встановленні порогу 0,3; при цьому буде пропущено 10 дефолтів. Очевидно, що мережа Байєса має схильність до «перестраховки», тобто вона частіше помиляється, відкидаючи тих клієнтів, які б повернули кредит. Зрозуміло, що точність моделі та кількість помилок I-го та II-го роду буде залежати від порогу відсікання, який буде встановлений банком. Слід пам'ятати, що встановлюючи поріг відсікання, ми визначаємо не лише процент відсіяних клієнтів, а нижню границю ймовірності повернення кредиту, тобто поріг, нижче якого клієнт вважається таким, що не поверне кредит (або нижню границю ймовірності дефолту, нижче якої вважається, що клієнту слід видати кредит). При цьому значення ймовірності дефолту 0,1 або 0,2 для клієнта є доволі незначними та статистично малими, а тому поріг відсікання доцільно було б встановити на рівні 0,25 – 0,3. Зрозуміло, що встановлений поріг відсікання впливає на кількість помилок I-го та II-го роду. Для МБ побудована ROC-крива, представлена на рис.7.

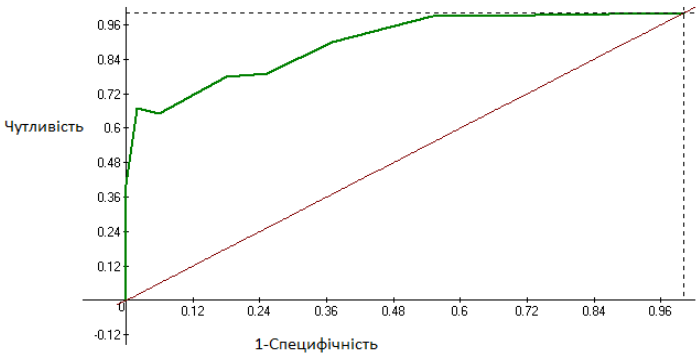


Рисунок 7 – ROC-крива для мережі Байєса

Для того щоб порівняти моделі, використаємо більш адекватний класифікатор – індекс GINI. Для цього спочатку обчислимо значення площі під кривою: $AUC = 0,86$. Відповідно, індекс GINI становить: $GINI = 2 * AUC - 1 = 0,72$.

Аналіз отриманих результатів

Таким чином, використано три методи інтелектуального аналізу даних для прогнозування кредитоспроможності фізичних осіб за статистичними даними: лінійна регресія, логістична регресія та дерева рішень. Для виконання обчислювальних експериментів використано вибірку даних для 15000 клієнтів банку та вибірку для 7568 клієнтів. Вибрано множину статистичних критеріїв для аналізу якості класифікації клієнтів.

Кращі результати класифікації клієнтів на дві групи отримано за допомогою дерев рішень (таблиця 5), що можна пояснити можливістю досягнення високої якості класифікації на дві групи за допомогою цього методу.

Таблиця 5 – Стандартне відхилення трьох моделей

Вибірки	Лінійна регресія	Логістична регресія	Дерева рішень
1) ALL_1	0,217419	0,217264	0,001
2) BASE_743	0,392419	0,377937	0,006

Результати порівняння якостей моделей на основі бінарної логістичної регресії та мереж Байєса показано у таблиці 5. Так, на основі моделей у формі мереж Байєса і бінарної логістичної регресії отримані дуже високі значення індексу GINI. Наведені результаті показують, що у подальшому банкам доцільно використовувати поряд з відомими скоринговими моделями ще й мережі Байєса, оскільки отримані результати та прогнозуючі якості мереж Байєса дозволяють відсіяти недобросовісних позичальників і таким чином зменшити обсяги втрат банків від неповернення кредитів.

Найкращі значення точності моделі також дає логістична регресія. Ці результати ще раз підтверджують доцільність використання логістичної регресії при оцінюванні кредитоспроможності позичальника.

Очевидно, що мережа Байєса має схильність до «перестраховки», тобто вона частіше помиляється, відкидаючи тих клієнтів, які б повернули кредит. Зрозуміло, що точність моделі та кількість помилок І-го та ІІ-го роду буде залежати від порогу відсікання, який буде встановлений банком.

Таблиця 6–Порівняльна таблиця характеристик для моделей Байєса та бінарної логістичної регресії

All				
Назва методу	**Індекс GINI**	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	**0,74**	0,87	**0,77**	Дуже висока
Мережа Байєса	0,72	0,86	0,74	Дуже висока
BASE				
Назва методу	**Індекс GINI**	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	0,74	0,87	**0,79**	Дуже висока
Мережа Байєса	**0,76**	0,88	0,74	Дуже висока

Висновки

Виконано короткий огляд моделей оцінки кредитоспроможності.Проаналізовано їх переваги та недоліки. З наведеного огляду випливає, що існує потреба у створенні нових сучасних систем для оцінювання ризиків з метою їх мінімізації та розв’язання задач ризик-менеджменту.

Докладно описано три методи, які можна використати для прогнозування кредитоспроможності фізичних осіб: логістична регресія, лінійна регресія та дерева рішень. Незважаючи на те, що лінійна регресія використовується нечасто, її також можна використати для порівняльного аналізу результатів.

Також зроблено загальний огляд статистичних критеріїв аналізу якості класифікації клієнтів. Необхідно зазначити, що розрахунок критеріїв якості дещо відрізняється для лінійних і нелінійних моделей. При цьому існує можливість вибрати саме ті критерії, які необхідні для аналізу якості класифікації клієнтів.

Оцінено якість прогнозу та якість моделі за такими критеріями: коефіцієнт детермінації, сума квадратів залишків, статистика Дарбіна-Уотсона, середня квадратична похибка та середня абсолютна похибка.

Нелінійна регресія має значно кращі якісні показники, ніж лінійна регресія за рядом критеріїв. Встановлено, що кращі результати класифікації клієнтів у даному випадку отримано за допомогою дерев рішень (стандартне відхилення – від 0,1 до 0,6 %). Показано, що дерева рішень дають можливість отримати прийнятний за якістю результат класифікації.

За допомогою програми GeNIe 2.0 і програми EViews 7.0 побудовано мережі Байєса та нелінійні логістичні регресії. Сформульовано рекомендації для їх практичного застосування. Встановлено, що кращі результати класифікації клієнтів у порівнянні з логістичною регресією отримано за допомогою мережі Байєса(Індекс GINI=0,76, Значення AUC=0,88).

У подальших дослідженнях для оцінювання кредитоспроможності клієнтів фінансової установи доцільно побудувати спеціалізовану систему підтримки прийняття рішень на основі методів регресійного та інтелектуального аналізу даних. Також, для підвищення достовірності результатів, доцільно використати дані, отримані з альтернативних джерел.

Література

1. *Матигорова И. Ю.* Характеристика основных подходов к оценке кредитного риска // Экономическая наука и практика: материалы междунар. науч. конф. (г. Чита, февраль 2012 г.). — Чита: Издательство Молодой ученый, 2012. — с. 68 – 69.

2. *Джонс Ч.Т.*Анализ и расчет кредитных рисков –http://www.franklin-grant.ru/ru/services/06.shtml

3. *Liu Y.* The evaluation of classification models for credit scoring. Arbeitsbericht 02/2002, Institut fuËr Wirtschaftsinformatik, 2002.

4. *Агеев В.И.* Основные модели оценки кредитного риска в коммерческом банке –http://webcache.googleusercontent.com/search?q=cache:87exLBuErlMJ:zhurnal.ape.relarn.ru/articles/2011/066.pdf+&cd=2&hl=en&ct=clnk&gl=ua

5. *Андреева Г.*Скоринг как метод оценки кредитного риска – http://www.cfin.ru/finanalysis/banks/scoring.shtml

6. *Hosmer D. W., Lemeshow S.* Applied Logistic Regression. New York: John Wiley & Sons, Inc. 1989.

7. *Бідюк П.І., Романенко В.Д., Тимощук О.Л.* Аналіз часових рядів – К.: Політехніка, 2013. –600 с.

8. *Бідюк П.І.* Системний підхід до прогнозування на основі моделей часових рядів// *Системні дослідження та інформаційні технології,* 2003, №3, с. 88 – 110.

УДК 004.942

СКОРИНГОВІ МОДЕЛІ ПОВЕДІНКИ КЛІЄНТІВ-ВЛАСНИКІВ КРЕДИТНИХ КАРТОК ДЛЯ ОЦІНКИ ЇХ ПЛАТОСПРОМОЖНОСТІ

О. В. Фомін, Н. В. Кузнєцова

Вступ

Знаходження ймовірності дефолту споживчих кредитів, позик чи кредитних карток (КК) є однією з головних проблем, що постають перед банками, кредитними кооперативами та іншими установами, що надають фінансові послуги. Це перший крок, який необхідний для розрахунку капіталовкладень у резервування, спричинене ризиком втрати платоспроможності, тобто коли клієнт не виплачує кредит, і переходить у стан дефолту. Ризики, що виникають в такій ситуації, називаються кредитними [1]. Такі ризики стали об’єктом дослідження відносно не так давно – в середині минулого століття [2,3].

Одним із основних елементів управління кредитними ризиками є оцінювання клієнтів за допомогою скорингових карт. При цьому на сьогодні банки приділяють увагу переважно аплікаційному скорингу, коли проводиться оцінювання нового клієнта [3-5]. Однак у процесі функціонування будь-якої фінансової компанії формується певний кредитний портфель і постає задача оцінювання платоспроможності вже існуючих клієнтів [6-10]. Особливо відчутна така потреба при моніторингу поведінки клієнтів-власників кредитних карток, оскільки сама природа такої позики диктує динамічність ризиків, що пов’язані із цим видом кредиту.

Для вирішення задачі оцінки платоспроможності клієнтів банку застосовуються поведінкові скорингові моделі. У вітчизняних банках такі моделі в основному реалізують у вигляді звичайної логістичної регресії. Однак цей підхід має ряд недоліків, зокрема: статичність, неефективність та непредикативність. Наявність таких недоліків спричинила появу та розвиток альтернативних моделей. Одним із варіантів альтернативних моделей є моделі теорії аналізу виживання, що дозволяють оцінювати не лише сам факт втрати

платоспроможності, а й час, через який це відбудеться. Дослідженню таких моделей та порівнянню їх ефективності із загальноприйнятими і присвячена дана стаття.

Постановка задачі

Для забезпечення компромісу між класифікацією клієнтів та прогнозуванням їх поведінки у даній роботі виконано огляд існуючих методів побудови скорингових моделей. Метою роботи є виявлення найбільш актуальних та перспективних підходів до розробки моделей, на практичних прикладах для реальних даних виявлення тих моделей, які краще підходять до скорингу поведінки для кредитних карток, розроблення власних моделей поведінки клієнтів-власників кредитних карт на базі підходу з теорії аналізу виживання; аналіз результатів та напрацювання рекомендацій щодо подальшого використання скорингових моделей поведінки.

Модель кредитної картки

Модель кредитної картки може бути представлена у вигляді сукупності певних параметрів, що характеризують саму картку, цикл її життя та кредитний продукт (табл. 1).

Таблиця 1.Модель КК у вигляді фрейму

Слоти	**Опис**
beginDate	Дата початку
finshDate	Дата номінального закінчення
closeDate	Дата фактичного закінчення
gracePeriod	Грейс-період
setLimit()	Встановлення ліміту
getLimit()	Перегляд поточного ліміту
balance	Баланс
postDuePayments	Кількість прострочених платежів
setLoan()	Надання кредиту
getOverdue()	Отримання величини прострочки
getOutstanding()	Отримання величини термінової заборгованості
getPayment()	Отримання щомісячного мінімального платежу

Кредитні картки, за якими було прострочено щонайменше 3 платежі на суму не менше 100 грн. вважаються дефолтними, тобто є «поганими» в контексті даного дослідження. Усі інші картки будуть цензуруватися, тобто вважатися такими, що не дійшли до свого логічного кінця.

Побудова скорингових моделей різних типів

На основі певних соціально-демографічних та кредитних параметрів позичальника розробляють скорингову математичну модель оцінки повернення кредитів [11]:

$$S = F(w^j, x_i^j),$$

де w^j – ваги параметрів x_i^j .

За скоринговою моделлю для кожного набору параметрів конкретного клієнту та кредитного продукту визначаються скорингові бали [4]. Для зручності роботи кредитних спеціалістів розробляють спеціальні скорингові карти, які є узагальненням скорингових оцінок, отриманих за існуючою статистикою у банку та експертних знань, що базуються на досвіді експертів в галузі кредитування та рекомендацій НБУ. В скоринговій таблиці визначена сукупність балів для кожної суттєвої змінної-характеристики клієнту та кредитного продукту в залежності від значення, яке вона набуває.

Наприклад, для змінної «Age» позичальник віком до 23 років отримує 63 бали, від 23 до 25 років – 76 балів, 25..28 – 79 балів, 28..34 – 85 бали, 34..46 – 94 бали, 46–51 – 103 бали, більше 51 року – 105 балів. Таке саме ранжування міститься в скорингової таблиці і по інших змінних. Якщо у вибірці існують пропущені значення, то передбачається їх оцінювання, зазвичай, у вигляді штрафних балів. Таким чином, скорингова таблиця є зручним, наочним та універсальним інструментам, що дозволяє автоматизувати процес розгляду кредитних заявок та скоротити час розгляду та прийняття рішення щодо видачі кредиту [12].

У роботі ставиться задача побудувати скорингові моделі на основі лінійної регресії, іншу – на основі моделі пропорційних ризиків Кокса (PH)[13], а результати моделювання порівнюються. Вимоги до розроблених моделей:

1. Модель на основі лінійної регресії повинна оцінювати ймовірність виходу в дефолт кредитної картки на основі її історії в банку за певний проміжок часу.
2. Модель РН повинна реалізовувати можливість динамічного оцінювання поведінки КК, порівнюючи ймовірність її виходу в дефолт з іншими КК.

Обидві моделі забезпечують поведінковий скоринг, однак представляють собою дві різні парадигми і підходи до вирішення проблеми динамічного оцінювання клієнта чи КК.

Логістична регресія

Логістична регресія [1, 12] використовує логіт-перетворення залежної змінної:

$$f(\mathbb{E}(y|x)) = f(p) = \ln\left(\frac{p}{1-p}\right) = \beta^T x \quad (1)$$

де $Y \in \{0,1\}$.

Функцію f у такому випадку називають логіт-перетворенням, а відношення $\frac{p}{1-p}$ – шансами (з англ. odds).

$$p = \mathbb{E}(y|x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} = \sigma\left(e^{\beta^T x}\right) \quad (2)$$

Функцію σ називають сигмоїдою, або логістичною функцією. Важливою особливістю цієї функції є її область значень: $E(\sigma) = [0; 1]$, що як ніяк краще підходить для оцінювання імовірності.

Таким чином з припущення:

$$\mathbf{P}(y = 1|x) = \sigma(\beta^T x) \quad (3)$$

В іншій формі запису має вигляд: $\mathbf{P}(y = 0|x) = 1 - \sigma(\beta^T x)$, тож при підстановці отримаємо:

$$\mathbf{P}(y|x) = \left(\sigma(\beta^T x)\right)^y \left(1 - \sigma(\beta^T x)\right)^{1-y} \quad (4)$$

Нехай маємо вектор спостережень виходу та матрицю значень незалежних змінних: $Y \in \mathbf{R}^n$, $X = (X_1 X_2 \dots X_n) \in Mat(m \times n)$, - відповідно. Тоді функція правдоподібності та її натуральний логарифм матимуть вигляд:

$$L(\beta) = \prod_{i=1}^{n} \mathbf{P}(y = Y_i | x = X_i) \quad (5)$$

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{n} (Y_i \ln(\sigma(\beta^T X_i)) + (1 - Y_i) \ln(1 - \sigma(\beta^T X_i))) \quad (6)$$

Природнім чином постає задача максимізації цих функцій. Знаючи градієнт (7) логарифмічної функції правдоподібності та її матрицю Гессе (7), можна знайти оцінку вектору β (9), наприклад, методом Ньютона.

$$l'(\beta) = \nabla l(\beta) = \sum_{i=1}^{n} (Y_i - \sigma(\beta^T X_i)) X_i \quad (7)$$

$$l''(\beta) = \boldsymbol{H}(\beta) = -\left(D(\beta)\right)^T X = -X^T D(\beta) X \quad (8)$$

де

$$D = diag(\sigma(\beta^T X_1)(1 - \sigma(\beta^T X_1)), \quad \sigma(\beta^T X_2)(1 - \sigma(\beta^T X_2)), \dots, \sigma(\beta^T X_n)(1 - \sigma(\beta^T X_n))$$

$$\beta^* = \arg\max_{\beta} L(\beta) = \arg\max_{\beta} l(\beta) \quad (9)$$

Аналіз виживання

Загальний підхід [14] до аналізу даних базується на використанні функції ризику, значення якої відповідає ймовірності смерті в певний час:

$$h(t) = \lim_{\delta \to 0+} \left(\frac{\mathbf{P}(t \le T < t + \delta | T \ge t)}{\delta} \right) \tag{10}$$

де T – випадкова величина, що відповідає часу виживання.

Ймовірність виживання в певний час t може бути записаною в термінах функції ризику:

$$S(t) := \mathbf{P}(T \ge t) \tag{11}$$

Формулу (11) можна трактувати, як ймовірність виживання в інтервалі часу з 0 до t . Звідси можна отримати ймовірність дефолту:

$$P_D(t) = 1 - S(t), \tag{12}$$

що являється нічим іншим, як функцією розподілу $F = \mathbb{P}(T < t)$ випадкової величини T . Відповідно до формул (11), (12):

$$S(t) = 1 - F(t), \qquad f(t) = -\frac{dS(t)}{dt} \tag{13}$$

Використовуючи введену функцію ризику (10) можна отримати ще один вираз для її обчислення:

$$h(t) = \frac{\lim_{\delta \to 0} \left(\frac{\mathbb{P}(t \le T < t + \delta)}{\delta} \right)}{\mathbb{P}(T \ge t)} = \frac{f(t)}{S(t)} \tag{14}$$

де f – відома нам щільність випадкової величини T .

Звідси неважко отримати вирази для $h = h(S)$:

$$h(t) = -\frac{\frac{dS(t)}{dt}}{S(t)} \tag{15}$$

Розв’язуючи диференціальне рівняння (16) отримуємо вираз для S , що залежить від h :

$$\int_0^t h(u)du = \int_0^t -\frac{S'(u)}{S(u)}du = \int_{S(0)}^{S(t)} -\frac{1}{s}ds = \log S(0) - \log S(t) = \log 1 - \log S(t) = - \quad (17)$$

З рівності (17):

$$S(t) = e^{-\int_0^t h(u)du} \quad (18)$$

Існує декілька підходів до оцінювання функції ризику. Найбільш розповсюджений – модель пропорційних ризиків Кокса [13]:

$$h(t, x(t), \beta) = h_0(t)e^{\beta^T x(t)} \quad (19)$$

Потрібно оцінити вектор коефіцієнтів β .

Якщо x не залежить від t , то

$$\frac{h(t, x_1, \beta)}{h(t, x_2, \beta)} = e^{\beta(x_1 - x_2)} = const \quad (20)$$

Це пояснює слово «пропорційні» в назві методу. Однак у випадку, коли коваріанти залежать від часу: $x = x(t)$, - це не так.

Виявляється, що для оцінки вектору параметрів β достатньо розглядати функцію часткової правдоподібності. Як імовірність того, що відбулася загибель певного конкретного індивіда i , за умови, що нам відомо про його загибель.

- t_i – час спостереження, тобто загибель або цензура;
- c_i – індикатор дефолту, $c_i = 1 \Leftrightarrow t_i$ – час загибелі;

Ймовірність того, що спостереження провалюється в певний час t серед інших спостережень обраховується наступним чином:

$$\frac{h(t, x_i(t), \beta)}{\sum_{j \in R(t_i)} h(t_i, x_j(t), \beta)} = \frac{e^{\beta^T x_i(t)}}{\sum_{j \in R(t)} e^{\beta^T x_j(t)}} \quad (21)$$

де $R(t) = \{j | t_j \geq t\}$.

Позначимо $\theta_j(t) \coloneqq e^{\beta^t x_j(t_i)}$, $\theta_j^i \coloneqq \theta_j(t_i)$. Тоді часткова функція правдоподібності та лог-перетворення від неї матимуть наступний вигляд:

$$L_p(\beta) = \prod_{i=1}^{n} \left(\frac{\theta_i^i}{\sum_{j \in R(t_i)} \theta_j^i} \right)^{c_i} \quad (22)$$

$$l_p(\beta) = \log L_p(\beta) = \sum_{i:c_i=1} \left(\beta^T x_i(t_i) - \log \sum_{j \in R(t_i)} \theta_j^i \right) \quad (23)$$

Звідси можна отримати вирази для градієнта та матриці Гессе, що застосовуються при знаходженні оптимального значення.

$$l'_p(\beta) = \sum_{i:c_i=1} \left(x_i(t_i) - \frac{\sum_{j \in R(t_i)} \theta_j^i x_j(t_i)}{\sum_{j \in R(t_i)} \theta_j^i} \right),$$

$$l''_p(\beta) = -\sum_{i:c_i=1} \left(\frac{\sum_{j \in R(t_i)} \theta_j^i x_j(t_i) x_j^T(t_i)}{\sum_{j \in R(t_i)} \theta_j^i} - \frac{\left(\sum_{j \in R(t_i)} \theta_j^i x_j(t_i)\right)\left(\sum_{j \in R(t_i)} \theta_j^i x_j^T(t_i)\right)}{\left(\sum_{j \in R(t_i)} \theta_j^i\right)^2} \right)$$

За такими характеристиками проводиться максимізація функції l_p для знаходження оцінки параметра β^* аналогічно формулі (9).

Проведення обчислюваних експериментів

Для експерименту було відібрано 55286 спостережень за 4037 кредитними картками зі щомісячною деталізацією. За весь період було зафіксовано 714 випадків дефолту. Для побудови моделей було сформовано список параметрів (табл. 2).

Таблиця 2. Список відібраних параметрів

Ідентифікатори	
dealb2id	Ідентифікатор угоди
monnum	Порядковий номер місяця життя
Дані про поведінку	
outbody	Залишок за тілом кредиту
outextra	Залишок за комісією та/або відсотком
ovdbody	Прострочена заборгованість за тілом
ovdextra	Прострочена заборгованість за комісією + відсотками
pddbody_new	Кількість днів прострочки за тілом
pddextra	Кількість днів прострочки за комісією + відсотками
ovd	Величина простроченої заборгованості
pdd_new	Кількість днів прострочки за КК
limit	Величина ліміту
montake	Кількість знятих коштів
monpay	Кількість повернених коштів
minpd	Номер місяця першого простроченого платежу понад 100 грн.
Аплікаційні дані	
clienttype	Тип клієнта
age	Вік
beginyear	Рік видачі
term_month	Тривалість угоди
dealamount	Ліміт на початок угоди
requestamount	Сума, на яку розраховував клієнт
is_satisfied	Признак того, що очікування клієнта справдилися
liv_is_reg	Признак того, що адреса прописки є такою ж як і адреса місця постійного проживання
childcnt	Кількість дітей
dependantcnt	Кількість утриманців
income	Дохід
u_score	Загальний скоринг-бал
has_u_score	Признак того, що загальний скоринг-бал розраховувався
with_bank_mon	Вік клієнта як контрагента
reg_mon	Тривалість останньої прописки
Агреговані дані	
max_outbody	Максимальна заборгованість за тілом
max_ovdbody	Максимальна прострочена заборгованість за тілом
max_ovd	Максимальна прострочена заборгованість
max_pdd_new	Максимальна кількість прострочених місяців
max_limit	Максимальна величина ліміту
avg_montake	Середнє значення щомісячного зняття коштів
avg_monpay	Середнє значення щомісячного платежу
max_ovd_over_limit	Максимальне значення відношення простроченої

	заборгованості до встановленого ліміту
Цільове поле	
defaulted	Цільове поле (дефолт/не дефолт)

Цільове поле, що описує дефолт/не дефолт, формувалося за критерієм:

- прострочка більше 100 грн.;

- більше трьох послідовних прострочених платежів.

Спочатку була побудована модель на основі логістичної регресії (4). Для оцінки коефіцієнтів використовувався метод Ньютона. Для цього застосовувалася функція glm стандартного набору пакетів мови R.

Табл. 3. Оцінені коефіцієнти та основні статистики

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.863284   0.613071  -1.408 0.159093
age                0.004807   0.005337   0.901 0.367745
is_satisfiedt     -0.411450   0.119419  -3.445 0.000570 ***
liv_is_regt       -0.018226   0.129879  -0.140 0.888402
childcnt           0.030690   0.078934   0.389 0.697421
dependantcnt       0.011642   0.124057   0.094 0.925233
log(1 + income)    0.029583   0.022285   1.327 0.184346
has_u_scoret       0.879906   0.580642   1.515 0.129671
with_bank_mon     -0.018143   0.009391  -1.932 0.053368 .
reg_mon           -0.004135   0.003742  -1.105 0.269086
clienttypeClient   0.479826   0.319386   1.502 0.133009
clienttypeEmployee 0.275121   0.411532   0.669 0.503796
clienttypeNormal   0.690322   0.242271   2.849 0.004380 **
clienttypeSalary   0.050843   0.246264   0.206 0.836434
log(1 + max_outbody)  0.142916   0.140520   1.017 0.309129
log(1 + max_ovdbody)  0.380554   0.186531   2.040 0.041334 *
log(1 + max_ovd)  -0.023716   0.183434  -0.129 0.897131
log(1 + max_limit)   -0.333863   0.087932  -3.797 0.000147 ***
log(1 + avg_montake)  0.367150   0.197322   1.861 0.062792 .
log(1 + avg_monpay)  -0.556334   0.065413  -8.505  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Перший стовпчик табл. 3 – це вектор незалежних змінних (вктор x) моделі, другий – це оцінені коефіцієнти (вектор β^*) при відповідних змінних. При цьому сама модель має вигляд: $p^* = \frac{e^{\beta^{*T} x}}{1 + e^{\beta^{*T} x}}$, де p – цільове поле, - дефолт власника кредитної карти.

Відповідно до описаної у попередньому розділі методології було побудовано декілька моделей PH, використовуючи функцію coxph.

Початкова модель M_0 обирала змінні динамічної поведінки клієнтів без лагів. Її основні показники наведено у табл. 4 (структура таблиці аналогічна табл. 3).

Таблиця 4. Основні показники моделі M_0

	coef	exp(coef)	se(coef)	z	p
log(1 + outbody)	0.14317	1.15392	0.03477	4.12	3.8e-05
log(1 + ovd)	0.74007	2.09607	0.04711	15.71	< 2e-16
pdd_new	1.49379	4.45395	0.04751	31.44	< 2e-16
log(1 + limit)	-0.21662	0.80524	0.01981	-10.93	< 2e-16
log(1 + montake)	-0.43572	0.64680	0.13098	-3.33	0.00088
log(1 + abs(monpay))	-0.03364	0.96692	0.02740	-1.23	0.21953
clienttypeClient	-0.05674	0.94484	0.23544	-0.24	0.80957
clienttypeEmployee	-0.19553	0.82240	0.40485	-0.48	0.62912
clienttypeInsider	0.65419	1.92359	0.64436	1.02	0.30998
clienttypeNormal	-0.24745	0.78079	0.18826	-1.31	0.18872
clienttypeOldSalary	0.04742	1.04856	0.24114	0.20	0.84410
clienttypeSalary	0.23902	1.27000	0.19967	1.20	0.23129
age	-0.01370	0.98639	0.00395	-3.47	0.00052
is_satisfiedt	-0.44390	0.64153	0.08898	-4.99	6.1e-07
liv_is_regt	-0.02701	0.97335	0.09588	-0.28	0.77814
childcnt	0.12223	1.13001	0.05829	2.10	0.03600
dependantcnt	-0.23561	0.79009	0.09375	-2.51	0.01196
log(1 + income)	-0.13613	0.87273	0.01892	-7.19	6.3e-13
has_u_scoret	0.38762	1.47347	0.49089	0.79	0.42975
with_bank_mon	-0.01518	0.98493	0.00874	-1.74	0.08254
reg_mon	0.00275	1.00275	0.00307	0.90	0.36960

Однак більш детальний аналіз параметрів даної моделі виявив необхідність її доповнення. Було вирішено доповнити модель з використанням зміщених у часі змінних. Таким чином були отримані три моделі M1, M2, M3, у яких замість поточних значень outbody, ovd, montake, monpay використовуються значення з певним лагом (табл. 5).

Така модифікація моделі спричиняє втрату потужності класифікації моделі, однак забезпечує її більшу предикативну здатність. Доцільність та компроміс між цими факторами буде предметом подальших досліджень.

Таблиця 5 Основні показники моделей $M1, M2, M3$

Скорочена назва змінної	Оцінені коефіцієнти моделі Mi		
	M1	M2	M3
log(1 + outbody(t-i))	-0.1984	-0.1679	-0.0500
log(1 + ovd(t-i))	1.0041	0.9217	0.6563
log(1 + limit)	-0.1422	-0.1458	-0.2052
log(1 + montake(t-i))	-0.8695	-0.5845	-0.0513
log(1 + abs(monpay(t-i)))	-0.2420	-0.1002	-0.0302
clienttypeClient	0.2738	0.0777	0.3777
clienttypeEmployee	0.3499	0.0680	-0.4239
clienttypeInsider	0.9472	0.8103	0.6047
clienttypeNormal	0.2748	0.1149	0.3139
clienttypeOldSalary	0.6650	0.4982	0.5961
clienttypeSalary	0.6318	0.1192	-0.0524

age	-0.0021	0.0018	0.0053
is_satisfiedt	-0.2529	-0.3165	-0.2657
liv_is_regt	0.1521	0.1887	0.1976
childcnt	0.0275	-0.0015	0.0083
dependantcnt	-0.0610	-0.0747	-0.0593
log(1 + income)	-0.0862	-0.0363	0.0012
has_u_scoret	0.1002	0.1747	0.4204
with_bank_mon	-0.0142	-0.0146	-0.0087
reg_mon	-0.0040	-0.0077	-0.0128

Аналіз результатів

Аналіз моделей проводився на основі співставлення вигляду ROC-кривих та площ під цими графіками (AUC)[1]. Безпосереднє порівняння кривих говорить про те, що моделі РН виявились більш ефективними (рис. 1). Також на це вказують значення AUC (табл. 6).

Однак значення критерію Акайке (AIC) говорять про те, що необхідно шукати компроміс між моделями. Зважаючи на це, найкращою моделлю являється М3, яка крім того, що краще класифікує «гарних» клієнтів та відзначається відносною «нескладністю», забезпечує прогноз поведінки клієнтів на три місяці вперед завдяки використанню лагових змінних.

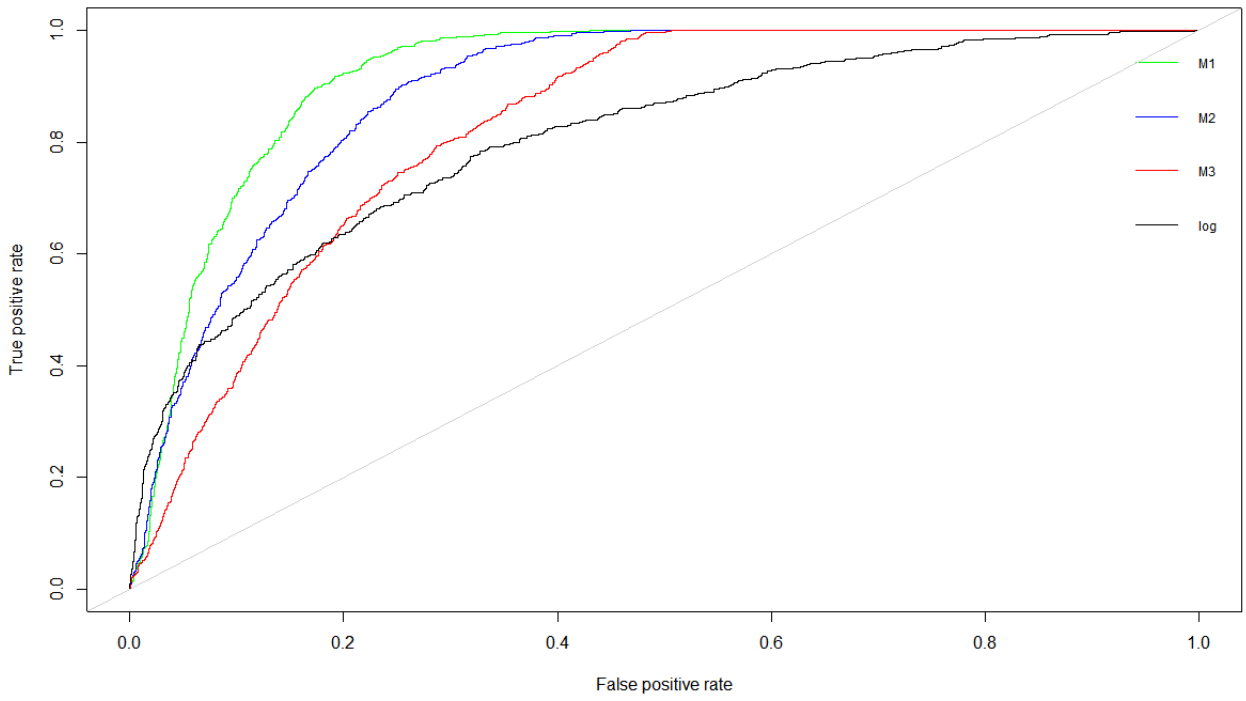


Рис. 1. Графіки ROC-кривих для моделей М1, М2, М3

Таблиця 6. Порівняння моделей

Моделі	M1	M2	M3	Logit
AUC	0.919	0.888	0.829	0,804
AIC	8946.005	7819.972	7495.905	2334.007

Висновки

У процесі дослідження виконано розробку та побудову альтернативних моделей у формі пропорційних ризиків Кокса з використанням лагових змінних на основі кредитної історії клієнтів банку. Результати обчислювальних експериментів відносно класифікації клієнтів для оцінки їх платоспроможності та прогнозування часу дефолту свідчать, що модель пропорційних ризиків Кокса, M3 з використанням поведінкових характеристик зміщених у часі на 3 місяці, надає найбільш близькі до реальних значення. Така модель забезпечує необхідний компроміс для управління ризиками, адже її дискримінативна сила у вигляді AUC = 82,9%, а також високі показники класифікації у верхніх сегментах є найбільш підходящими для застосування в реальному житті.

Наше дослідження підтвердило можливість використання моделей теорії аналізу виживання для оцінки платоспроможності клієнтів-власників кредитних карт, забезпечуючи більший рівень предикативності та надаючи ефективний інструмент для прогнозування поведінки клієнтів, що є основними вимогами до моделей поведінкового скорингу.

Література

1. Кузнєцова Н. В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / Н. В. Кузнєцова, П. І. Бідюк // Наукові вісті НТУУ “КПІ”. – 2010. – №1. – С. 42–53.

2. Narain B. Survival analysis and the credit granting decision / B. Narain // Credit Scoring and Credit Control. — 1992. — No. 1. — P. 1-2.

3. Carling K. Duration of consumer loans and bank lending policy: dormancy versus default risk / K. Carling, T. Jacobson, K. Roszbach // Working Paper Series in Economics and Finance. — 1998. — No. 280. — P. 109–122.

4. Stepanova M. Survival analysis methods for personal loan data / M. Stepanova, L. C. Thomas // Operations Research. — 2002. — Vol. 50, No. 2. — P. 277–289.

5. Roszbach K. Bank lending policy, credit scoring and the survival of loans / K. Roszbach // Sverriges Riksbank Working Paper Series. — 2003. — No. 154. — P. 1-2.

6. Glennon D. Measuring the default risk of small business loans: a survival analysis approach / D. Glennon, P. Nigro // Journal of Money, Credit, and Banking. — 2005. — Vol. 37. — P. 923–947.

7. Allen L. N. Financial survival analysis of defaulted debtors / L. N. Allen, L. C. Rose // Journal of Operational Research Society. — 2006. — Vol. 57. — P. 630–636.

8. Baba N. Survival analysis of hedge funds / N. Baba, H. Goko // Working Papers Series. — 2006. — Vol. 5, No. 6. — P. 1–2.

9. Malik M. Modelling credit risk of portfolio of consumer loans / M. Malik, T. L. // School of Management Working Paper Series. — 2006. — Vol. 7, No. 12. — P. 1–2.

10. Beran J. Credit risk modeling based on survival analysis with immunes / J. Beran, A. K. Dja // Statistical Methodology. — 2007. — Vol. 4. — P. 251–276.

11. Кузнєцова Н.В. Деякі аспекти мінімізації інформаційних ризиків у банківській діяльності / Н. В. Кузнєцова // Системні дослідження та інформаційні технології. – 2014. – № 1. – С. 7–19.

12. Siddiqi N. Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. – 2005. – 196p.

13. Cox D. R. Regression models and life-tables / D. R. Cox, S. Society, S. B. Methodological // 2007. — Vol. 34, No. 2. — P. 187–220.

14. Marimo M. Survival analysis of bank loans and credit risk prognosis master of science mathematical statistics / M. Marimo // [Електронний ресурс]. — Режим доступу : http://wiredspace.wits.ac.za/jspui/bitstream/10539/18597/1/Mercy%20Marimo%20Thesis_Survival%20Analysis_28.03.%202015_v1.pdf.

УДК 004.023

ВЕЙВЛЕТ АНАЛІЗ ХВИЛЬ ЕЛЛІОТТА

А. В. Слюсар, О. А. Гусєв, Д.С. Дульський

Вступ

В глобальній економіці періоди рецесій і зростань перемежовуються, тому аналіз передвісників є важливою проблемою [1]. Тому аналіз часових рядів валютного котирування є актуальним. Однією з найбільш ґрунтовних праць в області аналізу економічних тенденцій є так званий хвильовий аналіз Елліотта. Хоча сам аналіз був придуманий ще в 30-х роках XX століття, він досі залишається популярним напрямком досліджень. Математичною основою теорії Елліотта, за визнанням самого автора, стали так звані числа Фібоначчі — послідовність чисел, відкрита Фібоначчі [2]. Теорією хвиль Елліотта на даний момент займаються багато фінансових аналітиків[4],[5].

В даний час використання вейвлет-перетвореньдля аналізу фінансового ринку є обмеженим. Першою спробою використання вейвлет-перетворення для аналізу фінансових часових рядів було дослідження вчених університету міста Орхус, Данія [3]. У цій статті запропоновано використання дискретного вейвлет-перетворення лише для фільтрації високих частот та згладжування графіку.

Метою даної роботи є створення нового індикатора на основі вейвлет-перетворень. Окрім дискретного вейвлет-перетворення, пропонується також використання неперервного вейвлет-перетворення для виділення саме хвиль Елліотта.

Постановка задачі

Метою даної роботи є дослідження існуючих методів знаходження хвиль Елліотта і розробка власного алгоритму для виявлення хвиль Елліотта за допомогою попередньої обробки рядів котирувань та нового вейвлет-перетворень.

У даному дослідженні виділяємо такі підзадачі:

1. Аналіз структури хвиль Елліотта для подальшого дослідження
2. Дослідження фрактальної структури хвиль Елліотта
3. Побудова нового індикатора для виявлення хвиль Елліотта
4. Порівняльна характеристика вейвлет-аналізу та подвійного Фур'є-перетворення для якісного виявлення хвиль Елліотта

Аналіз структури хвиль Елліотта

Для теорії Елліотта було проведено дослідження стосовно виду та логічного опису структур хвиль [4]. Основою теорії хвиль Елліотта є три поняття: модель, співвідношення і час. В дещо спрощеному вигляді основний постулат теорії свідчить: фондовий ринок підпорядковується повторюваному ритму - п'ять хвиль зростання, три хвилі падіння [4].

П'ять хвиль(рис. 1), що складають фазу зростання, на малюнку пронумеровані. Висхідні хвилі (1, 3, 5) називають імпульсними хвилями.

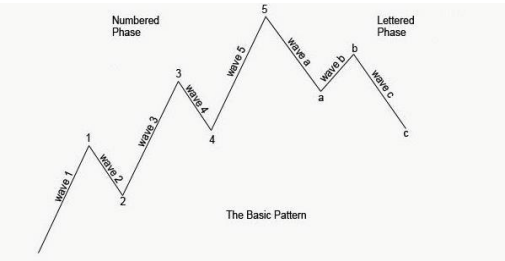


Рис. 1. Класичний приклад одного повного циклу хвилі Елліотта у фазі зростання

Спадні хвилі 2 і 4 розвиваються в протилежному тенденції напрямку. Їх називають коригуючими хвилями, так як вони вносять поправки в рух хвиль 1 і 3. Після того як зростання, що складається з п’яти хвиль, завершується, починається трьохвильове коригування. Три коригувальні хвилі позначені на малюнку буквами "a""b"і "c". ТакожЕлліоттом були задані наступні жорсткі правила щодо виявлення хвиль:

1. Хвиля 2 не може опуститися нижче початку хвилі 1
2. Хвиля 3 не може бути найменшою з імпульсних 1, 3 і 5
3. Хвиля 4 не може перетнутися з хвилею 1 по осі Y
4. Хвилі a, b, c не опускаються нижче початку хвилі 5

Крім того, хвилі Елліотта мають фрактальну структуру (рис. 2), а залежність наступної ланки від попередньої визначається за допомогою рівнів Фібоначчі [2].

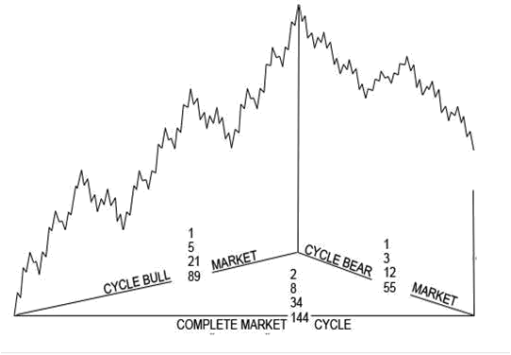


Рис. 2. Зразок повного цикл хвилі Елліотта до фондового ринку

Дослідження фрактальної структури хвиль Елліотта

Для виявлення наявності хвиль Елліотта в часовому ряді використовується індикатор зміни фрактальної розмірності. Використовується рухому вікно з інтервалом 20 значень, на якому обчислюється фрактальна розмірність Мінковського. Було встановлено, що для хвиль Елліотта характе-

рна розмірність в межах 1,15-1,45. Якщо фрактальна розмірність виходить за межі цього періоду, то можна стверджувати, що хвиль Елліотта на даному проміжку немає. На графіку(рис. 3)зображений аналіз валютних котирувань японська єна/долар за період с 1 січня 1971 року по 4 квітня 2016 року з місячними показниками.

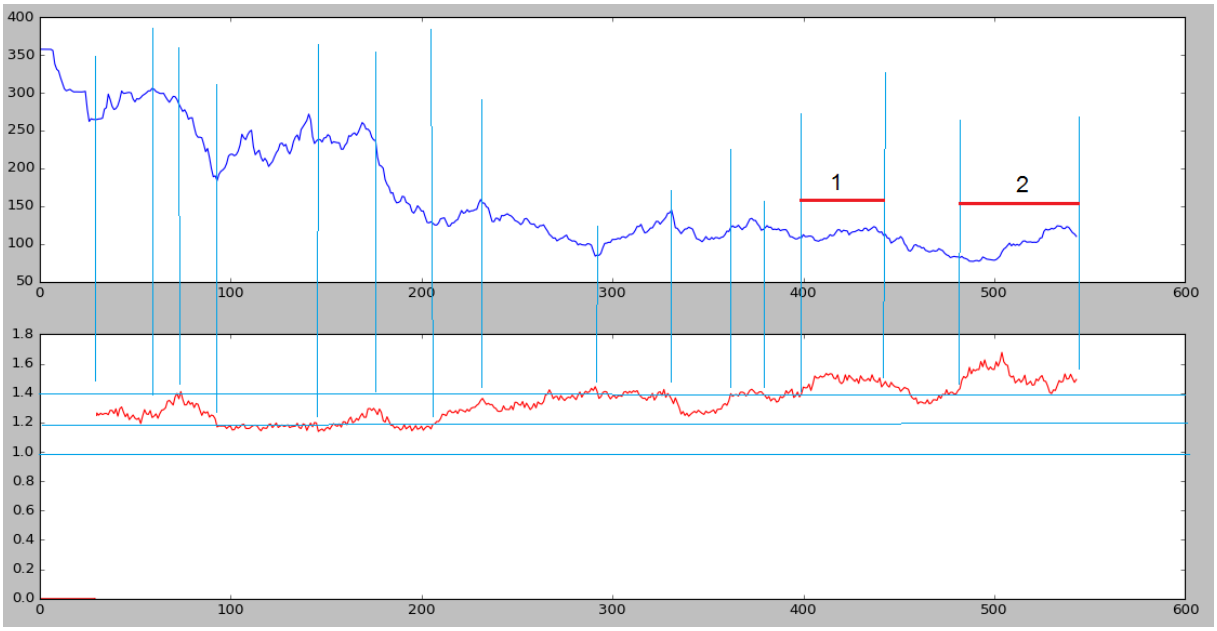


Рис 3. Вікно зміни фрактальної розмірності

Розбивши графіки на періоди окремих хвиль Елліотта, знаходимо, що 2 часові інтервали не містять хвиль Елліотта. Як бачимо, цим інтервалам відповідає фрактальна розмірність > 1,45.

Фур'є-аналіз хвиль Елліотта

Фур'є-аналіз - це потужний інструмент для обробки часових рядів. Він дозволяє виділяти частоти хвиль, які знаходяться в часовому ряді, та представити ряд у вигляді частотного спектру. Такий підхід дозволяє побачити циклічні хвилі, які впливають на формування ряду [10]. Результати

такого дослідження можуть бути неоднозначними, оскільки в основному ряд має велику зашумленність.

Для виконання такого дослідження використвується швидке фур'є-перетворення [8]. Якщо для прямого обчислення дискретного перетворення Фур'є з N точок даних потрібно O(x^N)арифметичних операцій, то FFT дозволяє обчислити такий же результат використовуючи O(N log(N)) операцій. Дискретне перетворення Фур'є перетворює набір чисел a_0,a_1, …, a_{n-1} в набір чисел b_0,b_1, …, b_{n-1} такий, що $b_i = \sum_{j=0}^{n-1} a_j \varepsilon^{ij}$, де $\varepsilon^n = 1$ і $\varepsilon^k \neq 1$ при 0<k<n.

$$y(k) = a_0 + a_1\, y(k-1) + \varepsilon(k)$$

Наступним кроком в дослідженні за допомогою Фур'є-перетворень є кепстри, є результатом прийняття зворотнього перетворення Фур'є (IFFT) логарифма оціненого спектру сигналу. Назва "cepstrum" була отримана шляхом обернення перших чотирьох букв слова «spectrum». Кепстр є зворотним перетворенням Фур'є, взятого від логарифму модуля Фур'є перетворення [9].

Результат роботи Фур'є-перетворення та кепстр продемонстровані на рисунках4 та 5 відповідно. На цих рисунках виведено графіки Фур'є перетворень та кепстри. Порівнявши результати можна побачити, що кепстра чіткіше виділяє піки.

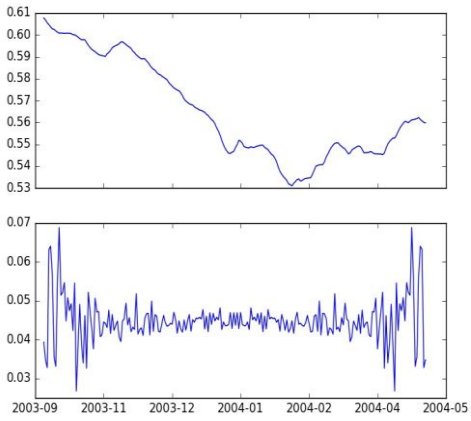


Рис. 4. Фур'є перетворення хвилі Елліотта

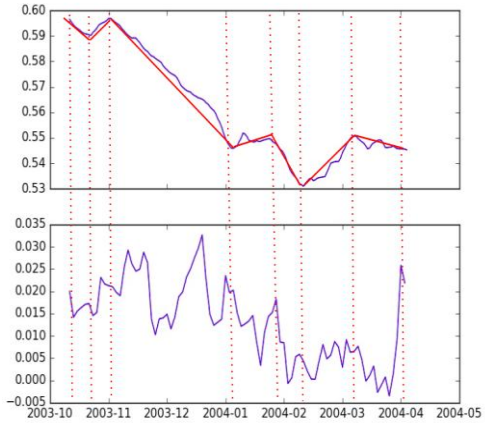


Рис. 5. Кепстр перетворення хвилі Елліотта

Вейвлет-аналіз хвиль Елліотта

Усі вейвлет-перетворення розглядають функцію (взяту як функцією від часу) у термінах коливань, локалізованих за часом (простором) і частотою.

Локальність у просторі означає, що енергія хвильок (вейвлетів) сконцентрована на скінченному інтервалі, так звана функція на компактному носії. Частотна локалізація означає, що перетворення Фур'є локалізоване. Частотна локалізація функції зводиться до понять гладкості та кількості зникаючих моментів. Вейвлет-перетворення звичайно поділяють на дискретне вейвлет-перетворення (DWT) та неперервне вейвлет-перетворення (CWT).

В нашому дослідженні використовувалися вейвлети Хаара, Добеши та Койфлет. Далі було застосованно вейвлети в двох випадків: після розбиття на інтервали зростання сталого тренду та без попередньої обробки. Оскільки вейвлети краще опрацьовують виділенні хвилі було виконане попереднє розбиття на основі показників Херста та MACD. Також було про-

ведено порівняння результатів роботи вейвлетів для всього часового ряду (рис. 6).

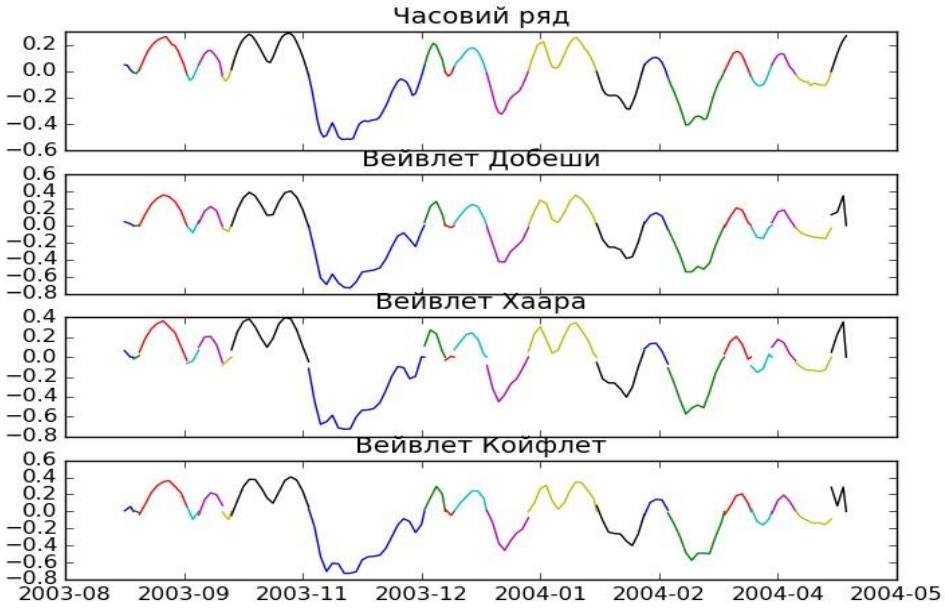


Рис. 6. Порівняння дискретного вейвлет перетворення на основі показника Херста

На рисунку 6 видно, що вейвлет Добеши краще фільтрує високі частоти та зберігає загальну поведінку хвилі при перетворенні менших хвиль, що взяті на основі показника Херста. Тому вейвлет Добеши краще використовувати при фільтрації хвиль Елліотта.

Приклад застосування методики для виділення хвиль Елліотта

Використовуючи класичний технічний аналіз [4], в часовому ряді валютного котирування виявлено класичну хвилю Елліотта. Як правило(рис. 7), хвиля Елліотта проявляється розмито(нечітко), тому для її виявлення запропоновано використання наступного алгоритму. До опису кроків алгоритму додано приклад його роботи на фінансовому ринку валютних котирувань американський долар/євро за 2003-2004 роки.

Крок 1. Для очищення часового ряду від “білого шуму”, використовується індикатор для первинного згладжування даних. У нашому дослідженні використано технічний індикатор MACD, що потрібен для знаходження сили і напрямку тренду. При цьому на часових рядах з поденною деталізацією даних рекомендується[5] користуватися різницею між 12-денним та 26-денним експоненційним ковзним середнім.

Крім того, досліджується можливість використання показника Херста, що є індикатором персистентності ряду[6].

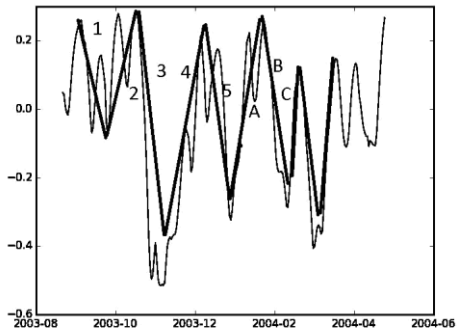


Рис. 7. Фактично виявлена хвиля Елліотта на графіку, де ланки 1-5 є імпульсом на спадання, а ланки a-c - їх корекцією

Крок 2. Після обробки часового ряду одним із індикаторів отримуємо згладжений ряд за показником MACD або Херста(рис. 8), на якому вже детальніше видно ланки хвиль Елліотта.

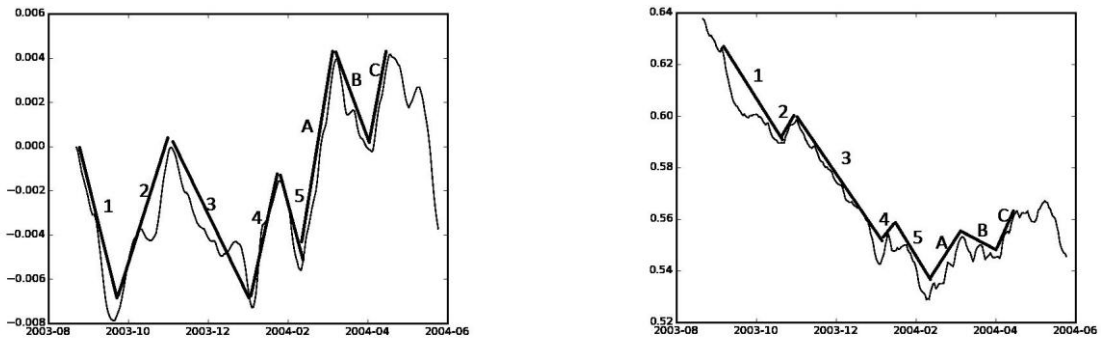


Рис. 8. Положення ланок хвилі Елліотта на графіку після застосування MACD та показника Херста

Оброблені за допомогою показників дані розбиваються на області зі зростаючою та спадаючою тенденцією. У контексті MACD зміна тенденції означала перетин індикатором нульового значення, у контексті показника Херста це означало перетин індикатором значення 0,5.

Крок 3.До кожної з виділених областей застосувується дискретне вейвлет-перетворення [7] із використанням одного із вейвлетів класичного набору(вейвлет Хаара(рис. 9), Добеші(рис. 10), Койфлет(рис. 11)).

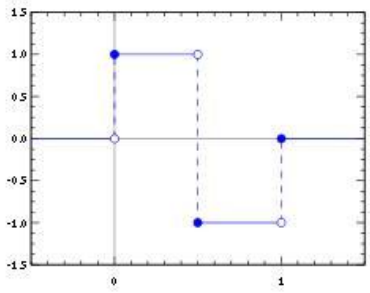


Рис. 9. Вейвлет Хаара

Рис. 10. Вейвлет Добеши

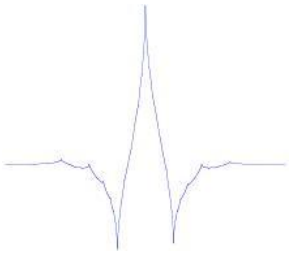


Рис. 11. Вейвлет Койфлет

Крок 4. Отримані за допомогою вейвлет-перетворення частоти пропускаються через фільтр (рис. 12) високих і низьких частот, вибираються лише низькі частоти. Побудований таким чином “низькочастотний” графік (рис. 13) і є новим індикатором, який дозволяє чітке виявлення хвиль Елліотта.

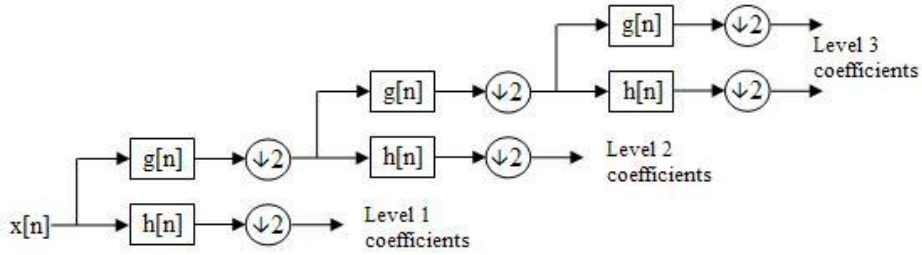


Рис. 12. Гребінка фільтрів високих і низьких частот

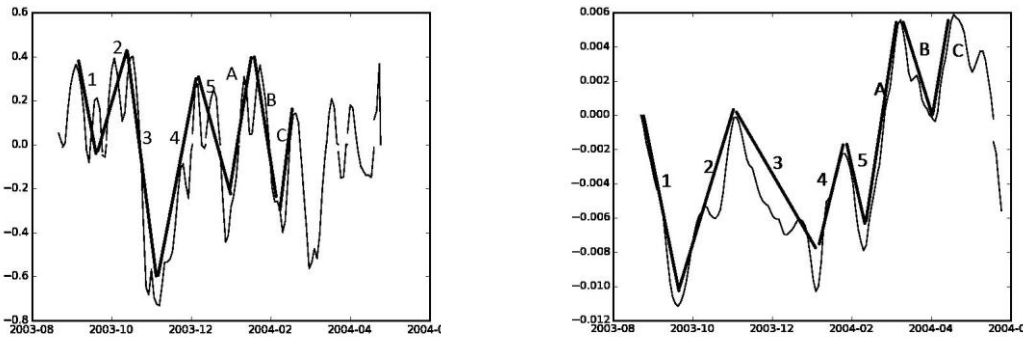


Рис. 13. Положення ланок хвилі Елліотта на графіку після застосування вейвлет-перетворення для показника MACD та Херста

Крок 5. До відфільтрованого графіка застосуємо неперервне вейвлет-перетворення. На отриманому спектрі (рис. 14) виділяються всі функціональні складові з циклічним характером, що дозволяє виявити можливу хвилю Елліотта.

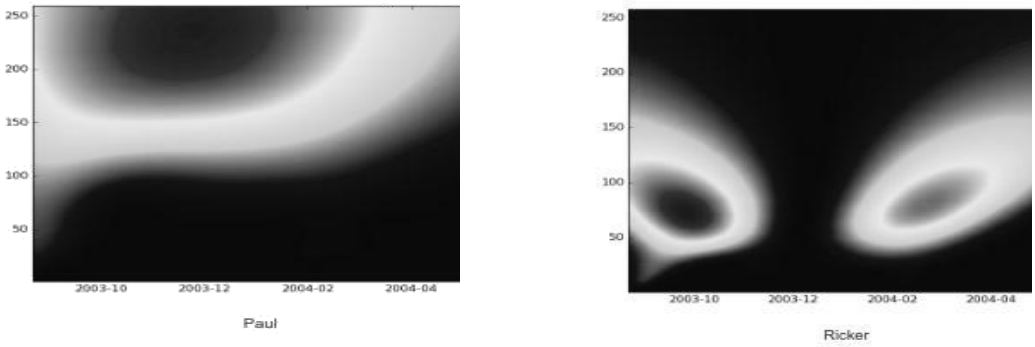


Рис. 14. Спектр вейвлет-перетворення вейвлета Paul та вейвлета Ricker

Також для виділення хвиль Елліотта використовується Wave-вейвлет, який виходить як перша похідна від функції Гаусса. За допомогою цього вейвлета виділяється точний початок циклу хвиль Елліотта(рис. 15).

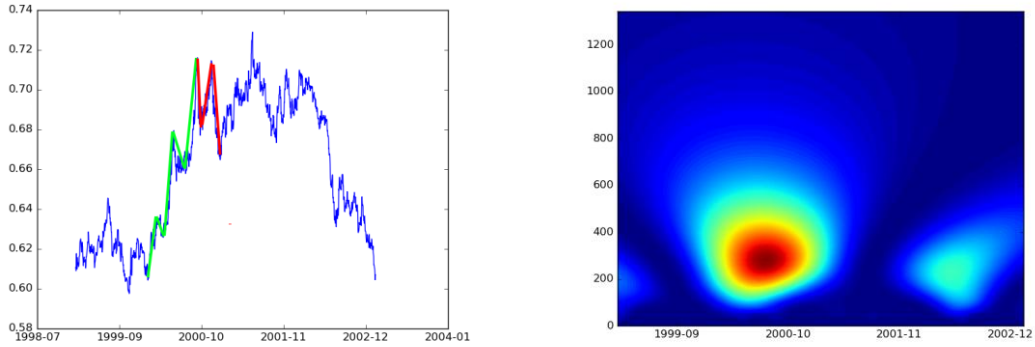


Рис. 15. Виділення хвилі Елліотта за допомогою Wave-вейвлета

Висновки

У роботі було розглянуто класичний підхід до виявлення хвиль Елліотта та створено новий індикатор(кроки 1-5) чіткого виявлення хвиль Елліотта(у порівнянні з аналізом вихідного графіку котирувань). На прикладі валютних котирувань показано виділення історичної хвилі Елліотта на спектрі створеного індикатору. Запропонований індикатор в подальшому пропонується вдосконалити для використання у існуючих методах виділення хвиль Елліотта.

Література

1. *Тарабрина М.Е.* Самоорганизация в сфере экономики // *Вестник Бурятского государственного университета*, 2010,№ 6, с. 77-80.
2. Razaqzada F. The Ultimate Fibonacci Guide— London: Fxstreet, 2010. — 5 с.
3. *C. Tan* Financial Time Series Forecasting Using Improved Wavelet Neural Network: дис. ... канд. комп. наук: 6.050101.— Arhus, Denmark, 2009. — 113 с.

4. *Джозеф Т.* Упрощенный Анализ Волны Эллиота. — Санкт-Петербург: Литера, 2012. — 80 с.
5. *S. B. Achelis.* Technical Analysis from A to Z.— Probus: ProbusPub, 1995. — 80 с.
6. *Калуш Ю. А., Логинов В. М.*Показатель Хёрста и его скрытые свойства. // *Сибирский журнал индустриальной математики,* 2002, №5:4. с. 29-37.
7. *Короновский А. А., Храмов А.Е.*Непрерывный вейвлетный анализ и его приложения.— Москва: Физматлит, 2003. — 158 с.
8. *H. J. Nussbaumer* Fast Fourier Transform and Convolution Algorithms.— London: Springer, 1982. — 240 с.
9. *Roads C.* The computer music tutorial. — Boston: The MIT Press,1996. — 1256 с.
10. *P. Bloomfield* Fourier Analysis of Time Series: An Introduction.— Boston: The MIT Press, 2014. — 288 с.

УДК 681.322

В. В. Логін

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ НАДАННЯ РЕКЛАМНОЇ ІНФОРМАЦІЇ ШЛЯХОМ МАСОВОЇ РОЗСИЛКИ

Науковий керівник: д.т.н., професор Бідюк П.І.

Вступ

Однією з поширених сфер, у якій потрібна оцінка ефективності, підвищення ефективності та порівняння моделей, є прямий маркетинг (з англ. *direct marketing*) – пряма особиста комунікація з отримувачем повідомлення або клієнтом середовища «бізнес для бізнесу», з метою побудови взаємовідношень та отримання прибутку [1, 2]. Головна задача у даній сфері – мінімізація кількості розсилок клієнтам, які не відгукаються на рекламу. Якщо проводити розсилку без попереднього аналізу клієнтської бази, то затрати на таку розсилку можуть виявитись вище доходу, отриманого від продажу товарів, що рекламуються. Зазвичай фірми, що займаються прямим маркетингом, мають так званий список контактів, тобто список клієнтів, яким вже здійснювалась рекламна розсилка. Деякі з клієнтів реагували активно, відгукнувшись, скажемо, на 7 розсилок із 10, інші – менш активно (2-3 відгуки), треті – взагалі ніяк не реагували.

Бажання людини придбати той чи інший товар залежить від багатьох ознак, що описують його як потенціального клієнта. Ними можуть бути вік, стать, дохід, сімейний стан, кількість дітей, наявність чи відсутність власного дому та автомобіля, клімат місцевості проживання та ін. Аналітик може виділити набір таких ознак для кожного виду товарів та побудувати модель, що буде визначати, чи

варто включати в список рекламної розсилки клієнта, що має певний набір значень ознак.

Тому, можна з впевненістю сказати що задача підвищення ефективності масової рекламної розсилки є дуже актуальною, оскільки рекламна розсилка є досить поширеною в наш час, причому її проведення без попереднього аналізу може спричинити збитки для компанії замість отримання прибутку.

Дослідженням сутності та розробкою методик рекламної розсилки займались такі провідні зарубіжні спеціалісти у цій області як А. Розенспен, Д. Берд, Е. Неш, Б. Стоун, Р. Джейкобс [3 – 6]. Вагомий внесок у вирішення питань пошуку клієнтів та їхнього переконання в купівлі товару зробили такі відомі спеціалісти як Д. Кеннеді, З. Фегеле та ін. [7, 8].

Аналіз тематичної літератури різних авторів, котра присвячена поняттю «масової розсилки», свідчить про те що у даному напрямку напрацьована досить велика кількість практичних результатів. Однак варто зазначити, що у різних методичних підходах до прямого маркетингу, процедурою якого є масова рекламна розсилка, наявна велика кількість протиріч. Це значною мірою ускладнює процес вибору певного підходу, хоча більшість основних підходів до проведення прямого маркетингу зазвичай забезпечують позитивний результат.

Оскільки XXI століття – це ера інформаційного суспільства, то останнім часом існує необхідність застосування інформаційних технологій та нових підходів до всіх напрямки діяльності, зокрема до маркетингу. Одним із пріоритетних засобів при проведенні сучасного маркетингу є моделювання процесів, що у ньому відбуваються, із використанням методології Data Mining.

Дослідження, що описується у даній статті, присвячене підвищенню ефективності масової рекламної розсилки. Зокрема наведено опис побудови моделей відгуку клієнтів на рекламу та вибору кращої моделі.

Постановка задачі

На початку дослідження була поставлена задача підвищення ефективності масової рекламної розсилки. Опишемо детальніше поставлену бізнес-задачу. Деяка торгова компанія, що здійснює продаж товарів, має інформацію про своїх клієнтів та їхні покупки. Компанія провела рекламну розсилку своїм клієнтам та отримала певний відсоток відгуків. Необхідно побудувати моделі відгуку та проаналізувати результати, щоб запропонувати шляхи мінімізації витрат на нові поштові розсилки. Для того, щоб розв’язати поставлену задачу, потрібно здійснити такі кроки:

- Виконати загальний огляд методів Data Mining та вибрати ті, які відповідають поставленій проблемі моделювання відгуків клієнтів на рекламну розсилку.
- Побудувати моделі Data Mining за допомогою обраних підходів.
- Порівняти отримані моделі та вибрати найбільш ефективну з точки зору отриманого прибутку.
- За вхідні дані використати статистичні дані деякої торгової компанії, а на виході отримати найбільш ефективну стратегію проведення рекламної розсилки.

Застосування методології Data Mining

За **визначенням**, Data Mining полягає у виявленні в необроблених даних раніше невідомих, нетривіальних, практично

корисних і доступних для інтерпретації знань, що необхідні для прийняття рішень в різних сферах діяльності людини.

Залежності і шаблони, знайдені в процесі застосування методів Data Mining, повинні бути нетривіальними і раніше невідомими, наприклад, відомості про середні продажах такими не є. Отримані знання мають описувати нові зв'язки між властивостями, передбачати значення одних ознак на основі інших.

Data Mining – це не один метод, а сукупність великого числа різних методів виявлення знань. Існує кілька умовних класифікацій задач Data Mining. Варто зупинитись на чотирьох базових класах задач:

1. *Класифікація* – це встановлення залежності дискретної вихідної змінної від вхідних змінних.

2. *Регресія* – це встановлення залежності неперервної вихідної змінної від вхідних змінних.

3. *Кластеризація* – це групування об'єктів (подій, спостережень) на основі даних, що описують властивості об'єктів. Об'єкти, що знаходяться всередині кластера мають бути схожими один на одного і відрізнятися від інших, які увійшли в інші кластери.

4. *Асоціація* – це виявлення закономірностей між пов’язаними подіями, Для прикладу, такою закономірністю є правило, яке вказує, що з події X випливає подія Y. Такі правила називаються асоціативними

Таким чином, як вже зазначалось у вступі, проведення рекламної кампанії без попереднього аналізу може спричинити збитки для компанії замість отримання прибутку. Саме тому аналітик повинен виділити набір певних ознак для кожного виду товарів та побудувати модель, що буде визначати, чи варто включати в список рекламної

розсилки клієнта, що має певний набір значень ознак. Тобто потрібно побудувати моделі відгуку клієнтів на рекламну розсилку.

Оскільки відгук клієнтів приймає тільки два значення – був відгук та не було відгуку, то побудова моделі відгуку зводиться до задачі класифікації. Тому за допомогою методології Data Mining потрібно побудувати моделі, що будуть розв’язувати поставлену задачу класифікації. Після цього потрібно застосувати побудовані моделі до конкретних даних та вибрати модель з найкращою ефективністю. Тим самим ми підвищимо ефективність рекламної розсилки у порівнянні із моделлю «Розіслати всім».

Вибір методів Data Mining для моделювання відгуків клієнтів

На початку дослідження була поставлена задача підвищення ефективності масової рекламної розсилки. Проаналізуємо поставлену задачу, та визначимо шляхи її розв’язання.

Оскільки ефективність розсилки напряму залежить від відгуку клієнта, то логічно, що сформульована вище задача зводиться до бінарної класифікації. Отже, потрібно розв’язати задачу класифікації методами Data Mining. У попередньому розділі розглядалась ця задача. Оберемо деякі методи розв’язання поставленої задачі класифікації.

Так як маємо справу із бінарною класифікацією, то серед статистичних методів варто вибрати логістичну регресію, що призначена саме для таких типів задач. Поміж методів машинного навчання доцільно вибрати дерева рішень та нейронні мережі.

Опис вибраних методів Data Mining для моделювання відгуків клієнтів на рекламну розсилку

Логістична регресія. Логістична регресія – це статистична модель, що використовується для передбачення ймовірності виникнення деякої події. При цьому використовується логістична крива, а самих подій може бути декілька. Досить часто логістичну регресію використовують для розв'язання задач бінарної класифікації. Однак дана модель є чутливою до кореляції між характеристиками, тому в моделі не повинно бути сильно корельованих незалежних змінних.

Дерева рішень. Дерева рішень – популярний метод машинного навчання, що часто використовується при розв'язання задач класифікації. Даний підхід має дві основні проблеми – це проблема вибору критерію розділення у певній вершині, проблема зупинки навчання, а також проблема «обрізання» дерева. До переваг дерев рішень можна віднести наступні особливості: висока точність, отримання правил на «природній» мові, відносно швидкий процес навчання.

Нейронні мережі. Нейронні мережі – це ще один підхід, що заснований на машинному навчанні, який широко використовується для розв'язання задач класифікації. Вони корисні тим, що можуть моделювати практично будь-які, в тому числі нелінійні, залежності. Зокрема вони є досить ефективними у тих випадках, коли алгоритмам регресії та дерев рішень не вдається розділити множину об'єктів на класи чи побудувати регресійну залежність із допустимою помилкою. Але варто зазначити, що як і попередні моделі, нейронні мережі мають свої недоліки. Зокрема, вони не володіють пояснюючою здатністю дерев рішень. Крім того, вибір оптимальної конфігурації нейронної мережі для розв'язання конкретної задачі не завжди є очевидним. Також сам процес навчання мережі може виявитись досить затратним як за об'ємом обчислень, так і за часом.

Вибір інструментальної платформи для здійснення моделювання

Навіть найпотужніші технології знаходження закономірностей і машинного навчання, такі як Data Mining, не становлять особливої цінності без інструментальної підтримки у вигляді відповідного *програмного забезпечення*. Ринок програмних засобів продовжує формуватися донині, однак в цій області вже можна виділити деякі стандарти де-факто.

Ринок програмного забезпечення Data Mining ділиться на кілька сегментів: статистичні пакети з можливостями Data Mining; настільні пакети Data Mining; СУБД з набором алгоритмів Data Mining; аналітичні платформи.

У даному дослідженні, для здійснення моделювання відгуків клієнтів, я використовував аналітичну платформу Deductor, оскільки аналітична платформа дає змогу створювати завершені аналітичні рішення та містить в собі потужні засоби Data Mining. Я зупинився саме на платформі Deductor, тому що вона має інтуїтивно зрозумілий інтерфейс, має вбудовані сучасні методи отримання, візуалізації даних та їхнього подальшого аналізу.

Моделювання масової рекламної розсилки із використанням методології Data Mining

При побудові моделей відгуку клієнтів на рекламну розсилку можна виділити декілька етапів.

Етап 1. Вибір даних для обробки. Набір даних для аналізу є навчальним та був взятий із навчального набору даних, що постачається разом із роботою [9]. Даний набір містить інформацію про 13504 клієнтів, включаючи відомі відгуки на рекламну розсилку.

Всього для аналізу доступно 9 незалежних змінних та 1 залежна змінна. Також наявна інформація щодо затрат, які пов'язані з рекламною кампанією:

- очікуваний дохід від 1-го замовлення $RR = 40 од.;$
- витрати на обслуговування 1-го клієнта $CR = 20 од.;$
- витрати на розсилку для 1-го клієнта $CM = 3 од.$

Етап 2. Розбиття вихідної вибірки на навчальну та тестову. Після того, як вихідні дані були імпортовані до аналітичної платформи, проведемо розділення вихідної вибірки на навчальну та тестову. Спочатку розділимо у співвідношенні 60/40 %, а вже потім, після побудови моделей, повернемось до даного кроку, щоб вибрати оптимальне розбиття. На рис. 1 та на рис. 2 зображено розбиття вихідної вибірки на навчальну та тестову. Як бачимо із цих рисунків вихідна та результуючі множини є достатньо схожими, що необхідно для подальшого моделювання.

Етап 3. Проведення кореляційного аналізу. Поки що будемо працювати із навчальною вибіркою. Спочатку оцінимо вплив вхідних змінних на вихідну. Скористаємось обробником «Кореляційний аналіз» та відкриємо візуалізатор «Матриця кореляції» (рис. 3).

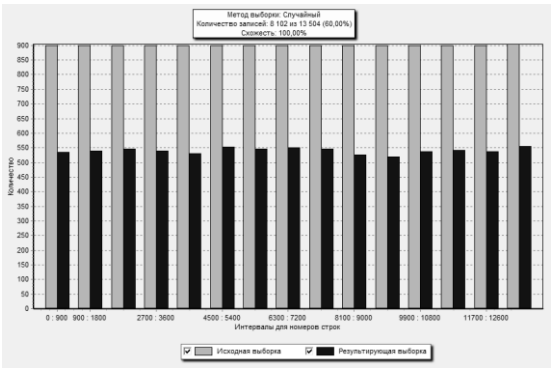


Рис. 1. Порівняння вихідної та навчальної вибірки

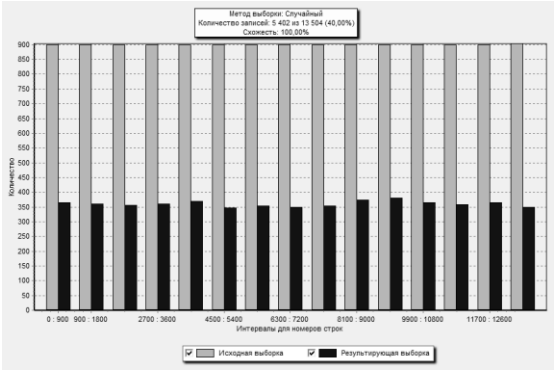


Рис. 2. Порівняння вихідної та тестової вибірки

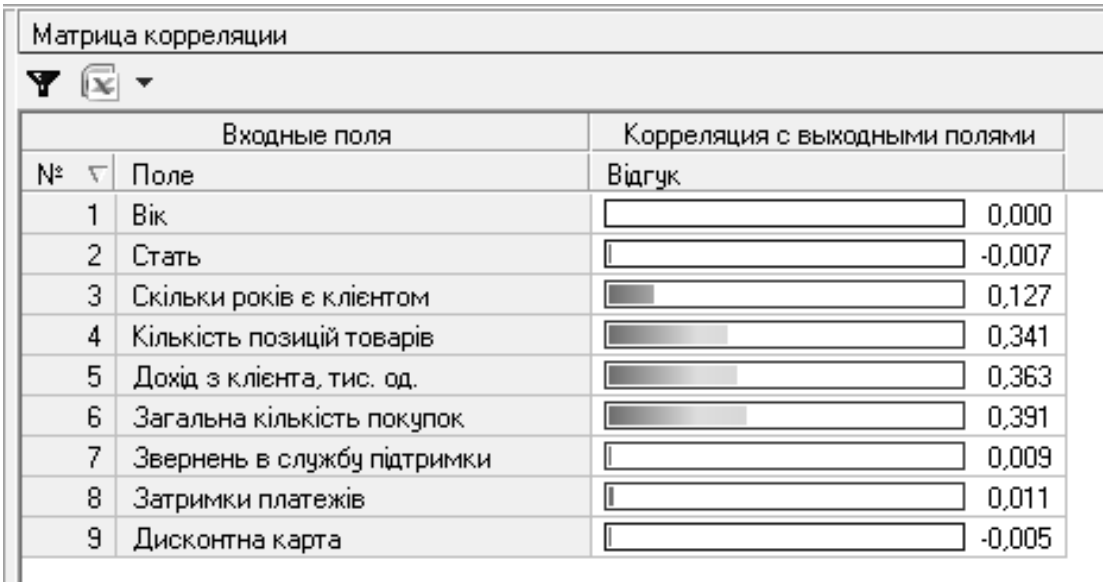


Рис. 3. Кореляція з полем «Відгук»

Як бачимо із останнього рисунка, що поля «Скільки років є клієнтом», «Кількість позицій товарів», «Дохід з клієнта, тис. од.» та «Загальна кількість покупок» сильно впливають на відгук на розсилку та демонструють позитивний зв’язок. Тому їх потрібно включити до моделі. У той же час, усі інші поля мають низьку кореляцію із полем «Відгук», тому їх не потрібно включати до моделі.

Етап 4. Оцінка рівномірності вихідних класів. Після цього приступимо до ще одного важливого етапу. Відкриємо статистичні характеристики для навчальної вибірки (рис. 4).

Рис. 4. Гістограма для атрибута «Відгук» для навчальної вибірки

Як бачимо, доля клієнтів із позитивним відгуком в навчальній множині складає 14,4 %. Тобто розподіл класів у вихідній змінній є нерівномірним. Як відомо, в такому випадку бажано будувати модель не на всій доступній множині прикладів, а рекомендується попередньо врівноважити їх.

Етап 5. Проведення моделювання відгуку клієнтів. Розпочнемо моделювання із побудови логістичної регресії. Як було зазначено раніше, розподіл вихідних класів є нерівномірним, тому здійснимо їхнє врівноваження. В логістичній регресії для збалансування класів потрібно відповідним чином розрахувати поріг відсікання. Для цього будемо використовувати правило Байєса, що є одним із найбільш поширених методів оцінювання оптимального порога. Відомо, що відношення втрат обох типів дорівнює $\frac{C_{FN}}{C_{FP}} = 17/3$. Тоді, відповідно до правила Байєса, поріг відсікання потрібно встановити рівним

$$\frac{1}{1+\left(C_{FN}/C_{FP}\right)} = \frac{1}{1+17/3} = \frac{3}{20} = 0{,}15.$$

Тепер побудуємо модель на основі логістичної регресії на даних навчальної вибірки. При цьому вхідні поля були відібрані при кореляційному аналізі, а вихідне поле – це відгук. У нашому випадку позитивним результатом буде вважатись наявність відгуку. Як було зазначено вище, поріг відсікання встановлюємо рівним 0,15.

В задачі даного типу потрібно також провести аналіз Lift-діаграми, яка дає змогу визначити розмір вибірки, який дасть найкраще співвідношення витрат (кількості розсилок) та прибутку (кількості отриманих відгуків). На рис. 5 зображена отримана Lift-крива. Діагональна лінія відображає роботу некорисного класифікатора, тобто ситуацію, коли списки отримувачів розсилки формуються випадковим чином. Як відомо, чим вище над діагональною лінією проходить графік кривої, тим кращою є модель.

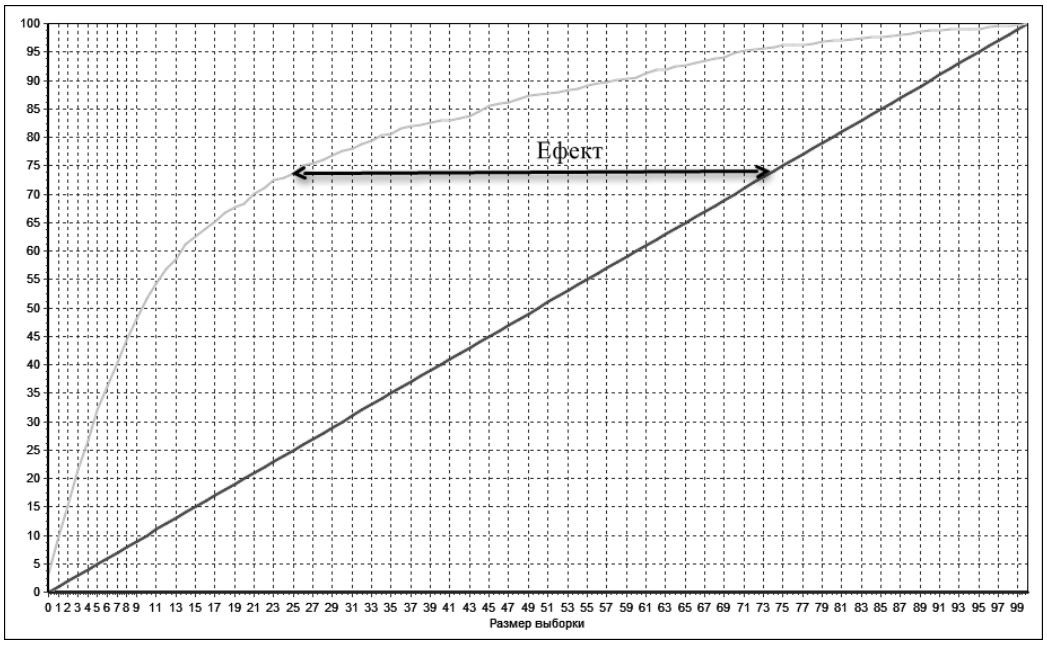


Рис. 5. Lift-крива для моделі відгуку на основі логістичної регресії

Також у даній ситуації корисно припустити, що компанія могла раніше використовувати правило, згідно з яким розсилка здійснювалась в першу чергу тим клієнтам, які принесли найбільші доходи. Тому побудуємо Lift-криву для такого випадку розсилки (рис. 6). Як виявилось, дана Lift-крива є гіршою від кривої для логістичної регресії. Але при невеликих об’ємах розсилки (до 7 %) не поступається їй за ефективністю.

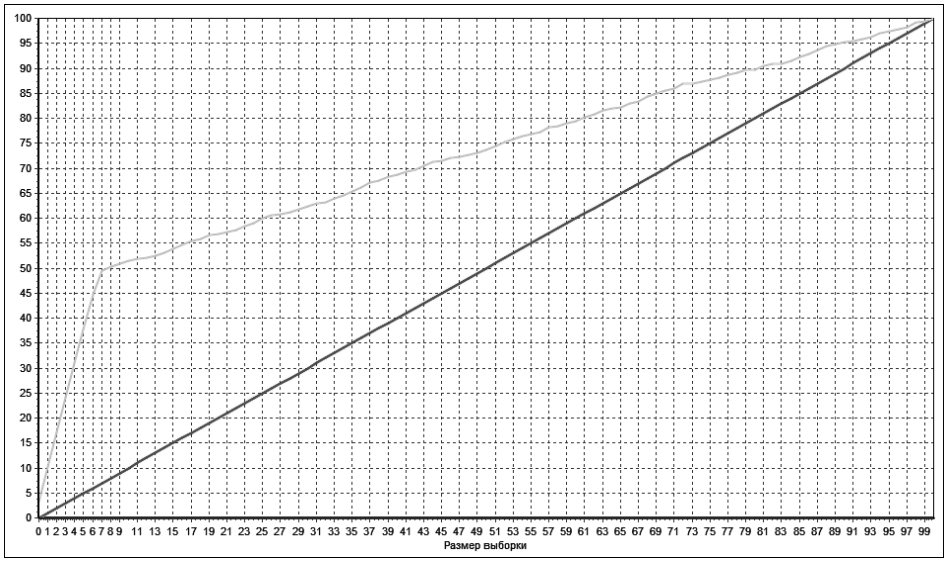


Рис. 6. Lift-крива для випадку першочергової розсилки клієнтам, від яких був найбільший дохід

Тепер побудуємо модель на основі дерева рішень. Як вже було зазначено раніше, класи є незбалансованими, тому проведемо спершу врівноваження. У даному випадку ми використаємо стратегію семплінгу, а саме – undersampling, оскільки маємо досить велику кількість прикладів (13504). Як відомо, відношення втрат від помилкової класифікації дорівнює 17/3. Тоді, відповідно до правила, потрібно пожертвувати приблизно 14/17 прикладів з клієнтами, від яких не було відгуку.

При налаштуванні обробника «Дерево рішень» вхідні та вихідні поля встановлюємо аналогічно із моделлю логістичної регресії. Побудова дерева буде здійснюватися за алгоритмом C4.5.

В результаті роботи алгоритму було виявлено 7 правил; точність класифікації на навчальній вибірці становить 81,83 %. Візуалізатор «Дерево рішень» дозволяє побачити отриманий набір правил в схематичному вигляді, а також виводить показники достовірності та підтримки для кожного вузла (рис. 7). Це і є модель відгуку на основі дерева рішень.

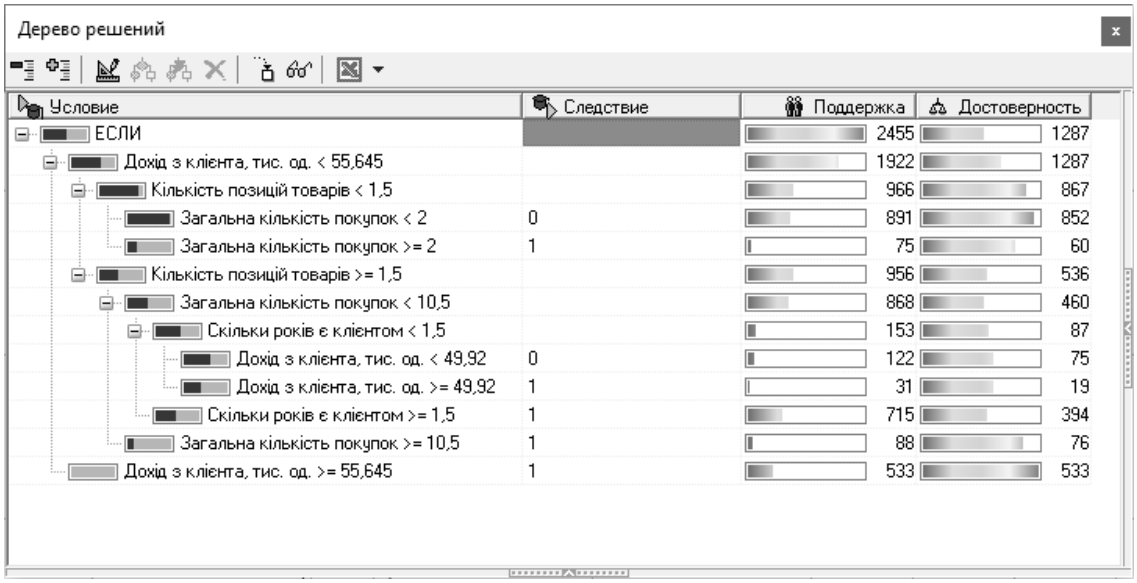


Рис. 7. Дерево рішень для моделі відгуку

Нарешті, приступимо до побудови моделі на основі нейронної мережі. Спершу виконаємо балансування вихідних класів. Як і у

випадку з деревом рішень, використаємо процедуру undersampling. Відповідно до неї потрібно відкинути приблизно 14/17 прикладів з клієнтами, від яких не було відгуку.

При налаштуванні обробника «Нейромережа» вхідні та вихідні поля встановлюємо аналогічно із моделлю логістичної регресії. Як відомо, із усіх архітектур нейронних мереж, багатошаровий перцептрон разом із сигмоїдною функцією активації є базовим для розв'язання задач класифікації та регресії. Також було згадано, що у задачах Data Mining найбільш популярним є тришаровий перцептрон, що включає вхідний шар, вихідний рівень та середній (внутрішній) шар (у нього входять невидимі, тобто приховані нейрони).

Кількість нейронів у внутрішньому рівні, відповідно до деяких обґрунтованих рекомендацій, має дорівнювати половині суми чисел вихідних та вхідних нейронів [10].

Тобто можна зробити висновок, що для нашої задачі достатньо одного прихованого шару із трьома нейронами. Для навчання виберемо алгоритм Back Propagation (BackProp), після навчання – візуалізатор «Граф нейромережі» (рис. 8).

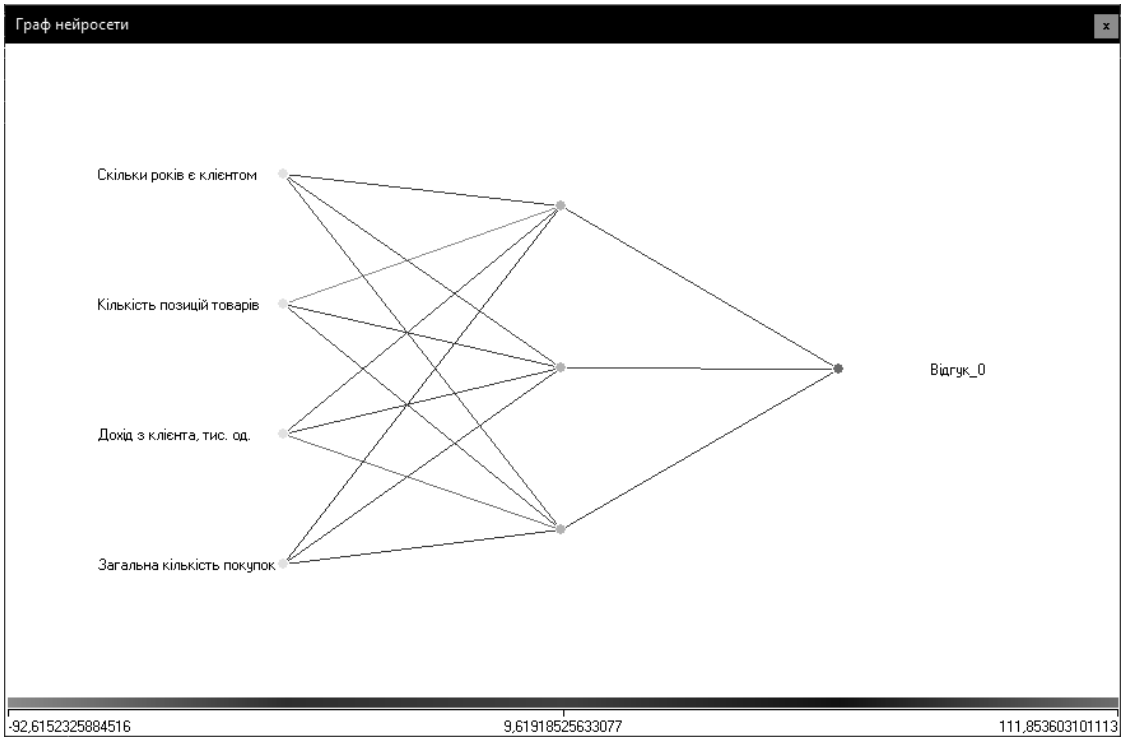


Рис. 8. Граф нейромережі

Етап 6. Вибір оптимального розбиття вихідної вибірки. Після того, як моделі були побудовані, повернемось до вибору оптимального розбиття вихідної вибірки, оскільки тільки зараз ми можемо дізнатись точність класифікації за моделями.

Таким чином, для порівняння варіантів розбиття, здійснимо наступні кроки. Як вже зазначалось раніше, будемо розбивати вихідну множину в таких співвідношеннях: 60/40 %, 70/30 % та 80/20 %. Після цього порівняємо отримані варіанти розбиття за точністю класифікації та виберемо найкращий. Саме з цим варіантом будуть проводитись подальші дослідження.

У таблиці 1 наведені показники якості моделей для варіанту розбиття 60/40 %.

Табл. 1. Показники якості моделей для розбиття 60/40 %

Тип вибірки	Модель	AUC	Gini, %	KS, %	Точність класифікації, %
Навчальна	Логістична регресія	0,84	68,75	55,57	81,03
	Дерево рішень	–	–	–	82,28

	Нейронна мережа	–	–	–	80,52
Тестова	Логістична регресія	0,85	70,5	57,47	82,01
	Дерево рішень	–	–	–	76,47
	Нейронна мережа	–	–	–	87,86

У таблиці 2 наведені показники якості моделей для варіанту розбиття 70/30 %.

Табл. 2. Показники якості моделей для розбиття 70/30 %

Тип вибірки	Модель	AUC	Gini, %	KS, %	Точність класифікації, %
Навчальна	Логістична регресія	0,85	70,1	57,16	81,6
	Дерево рішень	–	–	–	82,67
	Нейронна мережа	–	–	–	80,14
Тестова	Логістична регресія	0,84	68,3	54,51	80,99
	Дерево рішень	–	–	–	74,94
	Нейронна мережа	–	–	–	87,09

У таблиці 3 наведені показники якості моделей для варіанту розбиття 80/20 %.

Табл. 3. Показники якості моделей для розбиття 80/20 %

Тип вибірки	Модель	AUC	Gini, %	KS, %	Точність класифікації, %
Навчальна	Логістична регресія	0,85	69,3	55,75	81,1
	Дерево рішень	–	–	–	81,56
	Нейронна мережа	–	–	–	80,16
Тестова	Логістична регресія	0,85	70,4	57,8	82,71
	Дерево рішень	–	–	–	75,53
	Нейронна мережа	–	–	–	87,97

Отже, після порівняння даних в наведених вище таблицях, можна зробити висновок, що зміна тестової вибірки в межах 20-40 %

незначною мірою впливає на якість класифікації. Проте, серед наведених варіантів розбиття кращим виявився варіант 60/40 %.

Етап 7. Вибір кращої моделі відгуку клієнтів на рекламну розсилку. На даному етапі зведемо разом результати класифікації та прибутки для побудованих моделей на тестовій множині та для моделі «Розіслати всім» (табл. 4).

Таблиця 4. Результати класифікації

Модель	TN	TP	FN	FP	Загальна помилка	Прибуток, од.
Логістична регресія (поріг 0,15)	3859	571	213	759	17,99 %	7430
Дерево рішень (undersampling, 17/3)	3430	701	83	1188	23,53 %	8353
Нейронна мережа (undersampling, 17/3)	4206	540	244	412	12,14 %	7944
«Розіслати всім»	0	784	0	4618	—	–526

Як бачимо із табл. 4, зусилля, що буди прикладені на врівноваження вихідних класів, призвели до того, що на тестовій множині класифікатори частіше помилялись в бік псевдопозитивних випадків, що добре у нашому випадку, оскільки відношення помилок обох типів більше від 1. Позитивність такого результату пояснюється тим, що у нашому випадку, втрати від пропусків потенціального клієнта є більшими, ніж втрати від розсилки тим, хто не реагує на неї.

Також із таблиці 4 видно, що з точки зору отриманого прибутку найкращою моделлю для передбачення відгуку клієнта є дерево рішень, що дає прибуток 8353 од. Хоча варто зазначити, що з точки зору мінімальної похибки кращою є модель «Нейронна мережа».

При використанні моделі, що заснована на дереві рішень, ми можемо спостерігати підвищення ефективності розсилки на $8353 - (-526) = 8879$ од.

Аналіз отриманих результатів

Варто зазначити, що у випадку розсилання реклами всім клієнтам, компанія взагалі буде працювати в збиток. А після моделювання ми отримуємо стратегії розсилки, що дозволяють отримувати прибуток.

Після проведення порівняння моделей, було з’ясовано, що модель відгуку клієнтів на основі дерева рішень є найбільш ефективною з точки зору отриманого прибутку. Тому саме цю модель потрібно застосовувати при проведенні наступних рекламних розсилок.

В результаті дослідження було отримано інтуїтивно передбачуваний результат того, що в результаті розсилки реклами не всім клієнтам, а тільки тим у кого є певний набір ознак, компанія отримає більший прибуток.

Висновки

Отже, в результаті проведення дослідження, для розв'язання поставленої задачі було побудовано 3 моделі з використанням методології Data Mining: логістична регресія; дерево рішень; нейронна мережа.

Після порівняння побудованих моделей за передбачуваним прибутком, виявилось, що найбільш ефективною є модель на основі дерева рішень. При використанні даної моделі спостерігається підвищення ефективності рекламної розсилки, у порівнянні з моделлю «Розіслати всім», на 8879 од.

Таким чином, ми отримали нову стратегію рекламної розсилки, що є найбільш ефективною та повинна приносити прибуток, на відміну від стратегії «Розіслати всім». При наступних рекламних

розсилках рекомендовано використовувати стратегію, що заснована на моделі дерева рішень.

Новизна отриманих результатів полягає у тому, що при проведенні рекламної розсилки пропонується використовувати не стратегію «Розіслати всім» чи першочергової розсилки клієнтам, від яких був найбільший дохід, а стратегію, що заснована на моделі відгуку клієнтів на основі дерева рішень.

Список використаної літератури

1. *Директ-маркетинг* [Електронний ресурс] – Режим доступу до ресурсу: http://www.marketch.ru/marketing_dictionary/marketing_terms_d/direct_marketing/.
2. *Прямий маркетинг* [Електронний ресурс] – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/Прямий_маркетинг.
3. *Розенспен А.* Исповедь одержимого эффективностью / Алан Розенспен., 2005. – 368 с.
4. *Бёрд Д.* Прямой маркетинг. Бизнес здравомыслящих / Драйтон Бёрд., 2004. – 400 с.
5. *Nash E.* Direct Marketing: Strategy, Planning, Execution / Edward Nash., Hardcover, 600 Pages, Published 2000, 4th Edition.
6. *Стоун Б.* Директ-маркетинг: эффективные приемы / Б. Стоун, Р. Джейкобс., 2005. – 616 с.
7. *Кеннеди Д.* Жесткий директ-маркетинг: Заставьте покупателя достать бумажник / Дэн Кеннеди. – Москва: Альпина Паблишер, 2014. – 175 с.
8. *Зигфрид Фегеле.* Директ-маркетинг. 99 практических советов, как найти потребителя. – М. : Интерэксперт, 2001.

9. *Паклин Н. Б., Орешков В. И.* Бизнес-аналитика: от данных к знаниям (+CD): Учебное пособие. 2-е изд., испр. – СПб.: Питер, 2013. – 704 с.: ил.
10. *Ситник В. Ф., Краснюк М. Т.* Інтелектуальний аналіз даних (дейтамайнінг): Навч. посіб. — К.: КНЕУ, 2007. — 376 с.

УДК 004.8

Система реконструкції тривимірної моделі по двовимірним зображенням

Є.І.Морильов, М.В. Дідковська

Вступ

На сьогодні питання комп'ютерного зору та аналізу зображень є дуже актуальним, бо обсяги відеоінформації збільшуються в геометричній прогресії. Вартість камер та електронноїпам'ятізнизилась у кілька разів, але вартістьсканерівтривимірнихоб'єктів все щедужевисока. Крім того, в деяких областях неможливо застосування декількох камер одночасно, наприклад в комп'ютернійтомографії. Цю проблему можнавирішити за допомогоюсистемитривимірноїреконструкції по двовимірнимзображеннямвикористовуючизвичайну камеру, що робить дану роботу особливо актуальною, оскільки дані питання наразі ще не вирішені остаточно.

Завданнямтривимірної реконструкції почали займатися з кінця 70-х років. Великезначеннязробиловведення в 1981 році Лонгета-Хіггинсом так званоїсуттєвоїматриці (essentialmatrix), яку можнаотримати по парінормалізованихзображень і яка міститьінформацію про руходнієїкамерищодоіншої[1]. Пізніше, цепоняттябулорозширеноХартлі до фундаментальноїматриці. Фундаментальна матрицяміститьбагато корисної інформації про пару зображень, і практично всііснуючіалгоритмистереоскопіївикористовують ФМ в своїйроботі[2].

Область застосуваннятривимірноїреконструкціїдосить широка: віднавігаціїроботів до використанняувіртуальномутуризмі[3].

Найочевиднішим прикладом використаннявідновленнясцени є визначенняположення точки оглядущодорізнихоб'єктівнавколокамери - будинків, дерев, дорожніхзнаків, каменів і т.п. Ценаштовхує на думку використовуватитакісистеми для полегшеннякеруванняавтомобілем, визначаючивідстань до перешкод, щооточуютьавтомобіль, відображаючивідповіднуінформаціюводієві.

Недоліком базового алгоритму реконструкціхтривимірних сцен є накопиченняпомилок і швидкістьзнахоженняпозицій та поворотів камер віодносноодноеодної. Також, якістьреконструкції сильно залежитьвідвиборувихідної пари зображень і від порядку подальшихдоповненьпослідовності тих зображень, по котримбудуєтьсяреконструкція об’єкта.

Базовий алгоритм реконструкціїтривимірних сцен по двовимірнимзображеннямскладається з таких крків:

1. Знаходження ключових точок на зображеннях та їх дескрипторів
2. Знайтипопарнеспіввідношеннявихіднимизображеннями по знайденимключовим точкам.
3. Відфільтрувати хибні співвідношення
4. Решить систему уравнений и найти трёхмерную структуру вместе с положениями камер

Знаходженняключовихточоктаїїдекскрипторівможебутиреалізованом етодомSIFT, SURF, та іншими методами знаходження ключових точок, які інваріантні до:

- масштабування
- повороту

– зміщення

– зміни яскравості

По дескриптору ключових точок знаходиться співвідношеннямиж зображеннями. Якщо дескриптори точок на різних зображенях близькі, то можна вважати, що це точка одного і того же фізичного об’єкту.

Фільтрація хибних співвідношень проходить в два етапи:

1. Фільтрація, незалежна від геометрії.
2. Фільтрація по епіполярнихобженнях

Фільтрація, незалежнавідгеометрії описана в методіSIFT[5]. Найкраща фільтрація – це фільтрація з використанням епіполярного обмеження. Її геометричний сенс в тому, що для кожної точки одного зображення відповідна їй точка на іншому зображенні знаходиться на деякій лінії, яка не залежить від справжніх тривимірних координат точки (рисунок 1).

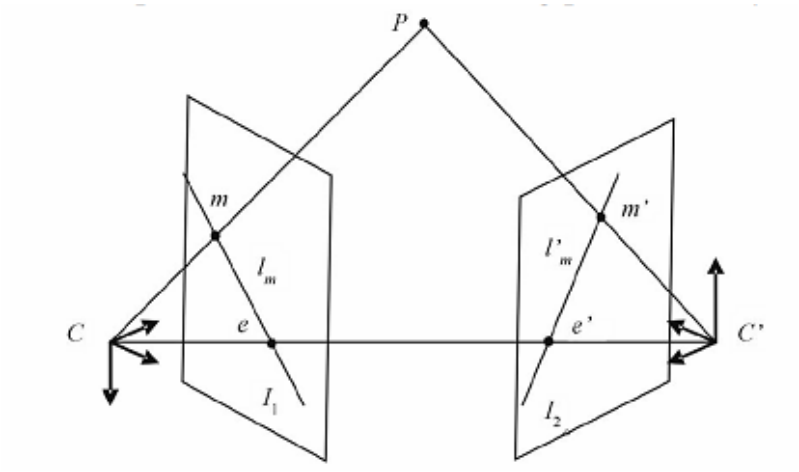


Рисунок 1 – Епіполярна геометрія. C і C' - оптичніцентри камер.

Якщо точка P на першомузображенніспроектованавm, тоді на іншомузображенніїїпроекціюпотрібношукати на прямийl'_m.

Математично ця властивість виражається рівнянням:

$$m^T F m' = 0 \qquad (1)$$

Тут $m = (u, v, 1)^T$ - однорідні координати ключової точки на зображенні, а F - фундаментальна матриця. Дане співвідношення (1) не залежить від тривимірної структури сцени, що реконструюється.

Оскільки фундаментальна матриця невідома, то для фільтрації помилкових співвідношень використовується наступний метод:

1. Вибирається випадково декілька точок і рахується по ним фундаментальна матриця (метод обчислення описаний в [5], розділ 14.3.2) .

2. Обчислюється кількість точок, які задовольняють умові $m^T F m' = 0$ із заданою точністю.

3. Якщо кількість точок досить велика, то припиняємо цикл, інакше йдемо на пункт 1.

Пошуктривимірноїреконструкціїшукаєтьсяпроективнимперетворення м.Геометричніпараметрикамерискладаються з двох частин:

– Зовнішні параметри, що визначають положення камери в просторі

– Внутрішні параметри – масштабування по кожній з осей, кут між осями і коефіцієнти спотворення більш високих порядків.

Зв’язок між тривимірними координатами $(X,Y,Z)^T$ точки та її положенням на зображенні $(u,v)^T$ має вигляд:

$$\begin{pmatrix} u \\ v \end{pmatrix} = k\left\{(R)\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + t\right\} = M\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + kt \tag{2}$$

Потрібно підібрати зовнішні та внутрішні параметри всіх камер та тривимірні координати точок так, щоб співвідношення (2) виконувалося якомога точніше для всіх точок.

В [5] це робиться наступним чином: Переносяться всі крапки на всіх зображеннях так, щоб центр мас знаходився на початку координат. Це дозволяє повністю виключити з рівнянь вектор зсуву.

З проекції всіх знайдених точок знаходиться матриця спостереження:

$$W = \begin{pmatrix} u_1 & \cdots & u_n \\ v_1 & \cdots & v_n \\ \vdots & \ddots & \vdots \\ u_1' & \cdots & u_n' \\ v_1' & \cdots & v_n' \end{pmatrix}$$

Тоді, записав всі матриці камер в M, а всітривимірні точки в матрицюX отримуємо рівняння .

Знаходимо сингулярне розкладання $W = UDV^T$. Тоді матриці камер отримаємо з перших трьох стовпців U, перемножених на сингулярне значення. Тривимірну афінну структуру отримуємо із перших стовпців матриці .

Постановка задачі

Метою роботи є створення системи реконструкції тривимірної моделі по двовимірним зображенням та модифікація методу реконструювання.

Об'єктом дослідження є інтелектуальні системи обробки та розпізнавання зображень. Предмет дослідження –методи та алгоритми реконструкції тривимірних об’єктів по двовимірним зображенням.

Для досягнення мети булипоставлені наступнізадачі:

1. Дослідження існуючих способів реконтрукції тривимірних об’єктів.
2. Розробка системи реконструкції тривимірної моделі по двовимірним зображенням.
3. Реалізація запропонованої архітектури комп’ютерної системи.
4. Аналіз результатів роботи програми на реальних прикладах та обґрунтування переваг запропонованого методу.

Модифікація базового алгоритму реконструкції

Алгоритм системискладається з наступнихкроків:

1. Введення зображень.
2. Знаходження ключових точок.
3. Співставлення зображень та відсів похибок.
4. Мінімізація похибки відносних поворотів камер.
5. Мінімізація похибки відносних позицій камер.
6. Триангуляція тривимірних точок.
7. Застосування алгоритму Bundle Adjustment дляоптимізації положення камер.
8. Отримання остаточної реконструкції та запис її у файл у точковому форматі PCL.

Основнамодифікаціябулазроблена в попереднійоцінціположень та поворотіввсіх камер в сцені, навідімінувідмінімізаціїпохибкипозицій камер

після кожного додаваннязображення до послідовностізображень,якіпідлягаютьреконструюванню, як булозроблено в базовому алгоритмі.

За фундаментальною матрицею, знаючивнутрішніпараметриK та K' камер, булоотриманоістотнуматрицюSVD-розкладаннямE в $E = UDV^T$.Тоді одна з матриць$[UWV^T|U_3], [UWV^T|-U_3], [UW^TV^T|U_3], [UW^TV^T|-U_3]$буде матрицею зсуву другоїкамери, причому при тріангуляції, тільки для однієї з цихматриць точка сцени, яку ми отрумуємо, буде знаходитися перед обома камерами.

$$W = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (1)$$

Нехай у нас є матрицяперетвореннядругоїсистеми координат щодопершоїM і матриця

$$I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad (2)$$

Для того, щобпідібрати масштаб загальноїсценищодопершоїмінімізувалася величина

$$\sum |KM_n P_{n-} - KIP_{ts}|, \quad (3)$$

де M_n – матриця, отримана з M заміною останнього стовпчика стовпцем, помноженим на n.

Цейкритеріймінімізували методом Макварда-Левенберга в результатічоговиходиламатрицяM_n переносу з першоїсценив другу.

Для кожної точки першогозображенняшукалися точки структури, відновленітільки по зображеннях з першоїкамериL_1 і по зображеннях з обох камер L_{12}. Ці точки через матрицюM_n'переносилися на друге місце з урахуванняммасштабів. Потім для першої точки шукаласявідповідна точка r на другому зображенні. Якщо вона знаходилась, для неїшукалася точка сцениR_1. В якості точки узагальненоїсценибрався центр мас$R_1, M_n, L_1, M_n', L_{12}$. Якщо одна або дві точки не були знайдені, то в якості точки сцени бралася або середина відрізка, якийз'єднує точки (коли булознайдено 2 точки), або сама ця точка.

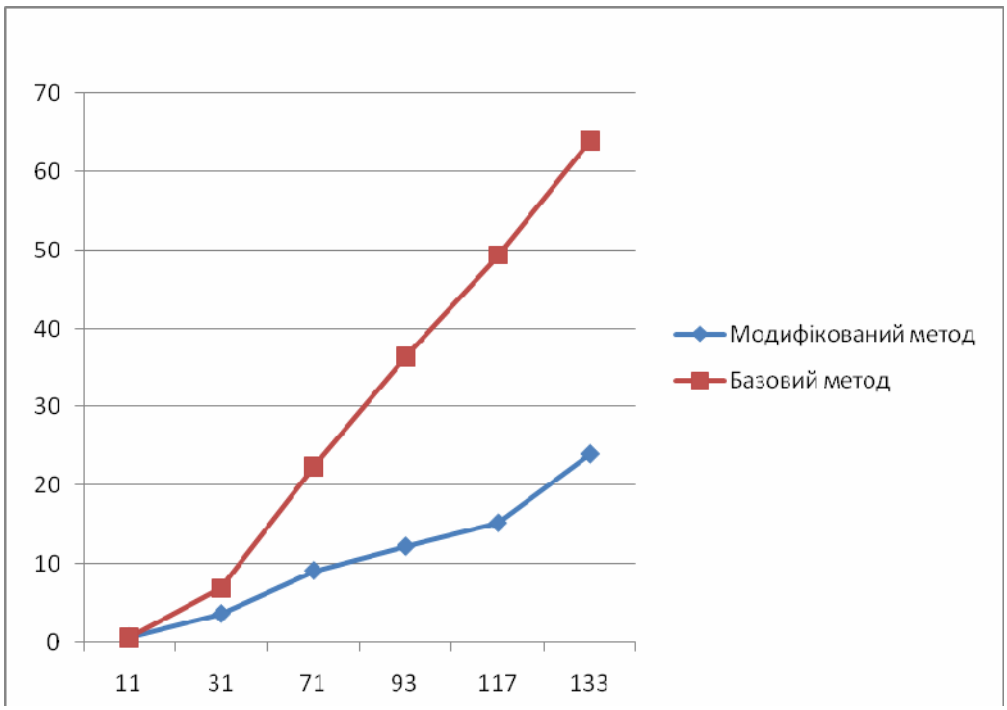
Критерії якості реконструкції

Для оцінки якості роботи системи використовувалося порівняння отрманних тривимірних моделей по таким параметрам:

– Похибка проекцій
– Кількість розпізнаних камер
– Кількість реконструйованих точок
– Середня кількість камер, з яких видно точку на кількість точок
– Час роботи алгоритму в залежності від кількості вхідних зображень

Результати застосування методики

На графіку представлено загальний час роботи базового та модифікованого алгоритму



У таблиці1 приведено швидкістьроботикожногоетапубазового алгоритму.

Таблиця 1 – Час роботиетапів для базового алгоритму

Кількістьзображень	Часроботи ВА (с)	Частриангуляції (с)	Часоцінкипозицій	Кількість ключових точок	Загальнийчас
11	0.4	0.1	0.01	1250	0.6
31	5.7	0.9	0.08	5586	6.9
71	18.5	3.2	0.2	14256	22.4
93	29.4	5.8	0.34	16760	36.4
117	40.2	7.4	0.46	21560	49.4

133	52.2	9.6	0.5	24547	64.3

У таблиці 2 приведено швидкістьроботи для кожногоетапумодифікованого алгоритму. Як ми можемопобачити, модифікований алгоритм дає кращі результати на всіх етапах. Основний приріст швидкості досягається на етапі роботи алгоритму BundleAdjustment. Кількість знайдених точок майже не відрізняється як у модифікованому, так і у базовому алгоритмі.

Таблиця2 – швидкістьроботиетапівмодифікованого алгоритму

Кількістьзображень	Час роботи BA (с)	Час триангуляції (с)	Часоцінкипозицій	Кількістьключовихточок	Загальнийчас
11	0.4	0.17	0.04	1159	0.63
31	2.64	0.82	0.18	5583	3.6
71	5.9	2.24	0.844	14211	9.11
93	7.8	2.8	1.5	16723	12.2
117	9.2	3.4	2.4	21663	15.2
133	14	4.7	4.4	24485	24.1

Аналіз отриманих результатів

Кількість реконструйованих точок для базового метода - 24596. Для модифікованого - 24674

Результати порівняння бавзового та модифікованого алгоритма:

1. Різниця повороту камер

Середнє значення: 0.000315

Медіана: 0.000264

2. Різниця положень камер

Середнє значення: 0.001643

Медіана:0.001336

Висновки

Уроботібулоотриманонаступнірезультати:

1. Представлена інтелектуальна система тривимірної реконструкції по двовимірним зображенням, що використовує модифікований алгоритм відновлення тривимірної сцени.

2. Запропоновано алгоритм, який дозволяє реконструювати тривимірну модель по двовимірним зображенням використовуючи звичайну камеру та базується на попередній оцінці всіх позицій та поворотів камер, триангуляції та використанні алгоритму bundleadjustment.

3. Проаналізовано та обгрунтовано преваги та недоліки представленої системи. Основним недоліком є більша чутливість до поганих зображень, але швидкість роботи алгоритму зросла майже у два рази.

4. Проведено аналіз результатів реконструкції, порівняння швидкості роботи та якості реконструкції. Результати показали, що час реконструювання сцени модифікованим алгоритмом для повного набору зображень склав 24.1с, що в 2.67 рази швидше за час реконструювання базовим алгоритмом. На більшій вибірці модифікований алгоритм дає ще кращі результати, оскільки базовий алгоритм робить переоцінку та

мінімізацію похибок положень для кожного наступного кадру, що поступає на вхід алгоритму.

Література:

1. Ж. Понс, Д. Форсайт, “Компьютерное зрение. Современный подход”- М.Издательский дом “Вильямс”,2004. – 427с.

2. D. J. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. “Discrete-continuous optimization for large-scale structure from motion”. In CVPR, 2011. – 18p.

3. M. Arie-Nachimson, S. Z. Kovalsky, I. KemelmacherShlizerman, A. Singer, and R. Basri. “Global motion estimation from point matches”. In 3DIMPVT, 2012.–14p.

4. J. Fredriksson and C. Olsson. “Simultaneous multiple rotation averaging using lagrangian duality”. In ACCV, 2012.–19p.

5. R. Hartley, A. Zisserman, “Multiple View Geometry in Computer Vision”– Cambridge University Press, 2000. – 551c.

УДК 519.226; 336.717.06

І.В. Пудло, П.І. Бідюк

ПРОГНОЗУВАННЯ ДОХІДНОСТІ БАНКІВСЬКИХ ПРОДУКТІВ З ВИКОРИСТАННЯМ СКОРИНГОВОГО ПІДХОДУ

Вступ

Ефективне функціонування банківських установ в умовах ринкової економіки та в умовах кризових явищ важливе місце повинна зайняти стратегія управління ризиком. Стратегія управління ризиком повинна поєднувати внутрішньобанківську діяльність із механізмами з боку Національного банку України та інших державних структур [1].

Кредитування є найбільш прибутковим і одночасно ризикованим видом банківської діяльності. Висока частка проблемного кредитного портфелю може спричинити банкрутство банку, а через його положення в економіці, до цілого ряду банкрутств, пов'язаних з ним суб’єктів господарювання. Тому управління кредитним ризиком є необхідною частиною стратегії і тактики розвитку будь-якого банку.

Із збільшенням об'ємів кредитування актуалізуються і задачі управління кредитним ризиком банку. В зв'язку з цим розробка методів оцінки і механізму регулювання кредитних ризиків забезпечує зміцнення фінансового положення банку.

Кредитний ризик є одним із найбільш вивчених видів ризику, для оцінки характеристик якого розроблено багато методів (серед існуючих методів оцінки кредитного ризику є скоринг, методика Монте-Карло, методика НБУ, методика Базельського комітету тощо).

Базовим та одним із найбільш поширеним елементом стратегії управління ризиком є скоринг.

Скоринг є математичною або статистичною моделлю, за допомогою якої на основі кредитної історії «минулих» клієнтів банк намагається визначити, наскільки велика вірогідність, що конкретний потенційний позичальник поверне кредит в строк [2].

Його суть полягає в тому, що кожному параметру, що характеризує позичальника, надається реальна оцінка в балах. Таким чином, у спрощеному вигляді, скорингову модель можна представити як зважену суму визначених характеристик позичальника: вік, сімейний стан, місце роботи, дохід та багато інших факторів. Така методика є знеособленою і може застосовуватися як для фізичних, так і для юридичних осіб.

Скоринг є класифікаційною задачею, де виходячи з наявної інформації необхідно одержати функцію, що найточніше розділятиме вибірку клієнтів на «поганих» і «хороших». Скорингові моделі є первинним індикатором кредитоспроможності потенційного позичальника. На їх основі експерт ухвалює остаточне рішення про видачу кредиту [3].

Можна виділити два основні типи кредитного скорингу, незалежно від того яка саме цільова змінна прогнозується – це:

– Аплікаційний скоринг – оцінка клієнта без використання інформації про його попередню поведінку в компанії (базуючись на соціо-демографічних характеристиках клієнта). Найчастіше модель будується на прогнозуванні ймовірності дефолту (недотримання зобов’язань) клієнта в перші n місяці життя кредиту. Основне завдання відділити на портфельному

рівні групи клієнтів з високою ймовірністю дефолту та запобігти видачі кредиту таким клієнтам.

– Поведінковий скоринг – оцінка клієнта з використанням інформації про його попередню поведінку в компанії. Найчастіше використовуються для застосування тих чи інших дій до клієнта вже в процесі обслуговування ним боргу. Діапазон прогнозування набагато ширший ніж в аплікаційному скоринзі: ймовірність дефолту, прогнозування повернення після дефолту,
ймовірність відтоку клієнта з банку (закриття договору), ймовірність переплати по кредиту, ймовірність додзвону тощо.

Статистичне прогнозування в кредитному скоринзі базується не на багаторазовому повторенні експерименту для одного клієнта (що технічно є неможливим), а в групуванні клієнтів в подібні групи по демографічних, бізнесових, поведінкових характеристиках та в знаходженні взаємозв'язку між характеристиками, що визначають ці групи, та цільовою змінною.

Прогнозоване значення цільової змінної розбивається для зручності на сегменти, – рейтингові класи, і стратегії всіх подальші дій з клієнтом та його кредитами приймаються на рівні рейтингу [4].

Основний статистичний метод, який в останні 20-30 років асоціюється в банках з кредитним скорингом – це логістична регресія [5]. Поряд з іншими більш витонченими та складнішими моделями вона має дві суттєві переваги:

– Простота інтерпретації: Вагові коефіцієнти отримані в результаті розрахунків легко переводяться в скорингові бали та карти, які в свою чергу інтерпретуються просто: «краща група

клієнтів – більший бал». Ризик-менеджмент більшості європейських банків не сприймає моделі на кшталт «чорний ящик», в яких логіка виведення рішення по конкретному запису(клієнту) настільки складна, що не може бути доволі просто, а то і взагалі відображена аналітично, тобто формулами на листку паперу. Прикладами таких моделей є моделі нейронних мереж, SVM, метод k-найближчих сусідів, Random forests тощо.

– Поряд з іншими статистичними методами, які можуть давати несуттєвий приріст до предикативної сили моделі, логістична регресія є однією з найбільш стабільних, тобто такою що зберігає свою силу протягом довгого періоду часу і не потребує частої перебудови. Крім того наявність класичної скоркарти дозволяє якісно «моніторити» модель та давати відповіді на питання, які характерстики моделі (фактично її складові частинки) дають збої в предикативності та зсуви в популяції.

Постановка задачі

Починаючи з 2006 року банки в Україні почали приділяти значно більшу увагу кредитним ризикам. З того часу були розвинуті знання у цій сфері та побудовано безліч моделей для мінімізації та контролю кредитного ризику. Попри це існує ще один ризик, який може нанести банку не менші збитки від кредитного. Він пов'язаний із клієнтами які користуються продуктами із грейсовим (пільговим) періодом або достроково закривають кредит.

Даний вид ризику полягає у тому, що банк, залучивши ресурси під певний відсоток (за певну ціну), не отримує прибутку від подальшої видачі цих ресурсів клієнту і цим самим несе збиток.

Запропонована модель дозволить провести сегментацію клієнтів і запобігти втратам, що можуть призвести до значного зниження кінцевого фінансового результату.

Таким чином, залежно від кредитної політики, яку проводить банківська установа, можна здійснювати більш якісний контроль дохідності від даного типу продукту.

Задача даної роботи полягає у розробці моделі для револьверних карток із грейсовим (пільговим) періодом, яка дозволить контролювати дохідність даного продукту.

Револьверний кредит (англ. revolving credit) — автоматично поновлюваний (від лат. revolve — обертатись) кредит, який широко використовується у світовій практиці на ринку позичкового капіталу [6].

Грейсовий період (англ. grace period) — пільговий період сплати відсотків за кредитом. За використання пільгового періоду відсотки за користування грошовими засобами не нараховуються взагалі або їх величина є значно нижчою за базову ставку.

Побудова моделі

Аплікаційна карта розрахована на прогноз дефолту по кредиту протягом певного фіксованого часу після його видачі (найчастіше 12 місяців). Дана модель оцінює популяцію, що заходить в банк (як видані так і відмовлені кредити) та визначає ймовірність дефолту кожного із таких кредитів, на основі чого зокрема і приймається рішення про видачу/невидачу кредиту. Відповідно і вибірка, історія дефолтів по якій вже відома, формується в такий же спосіб.

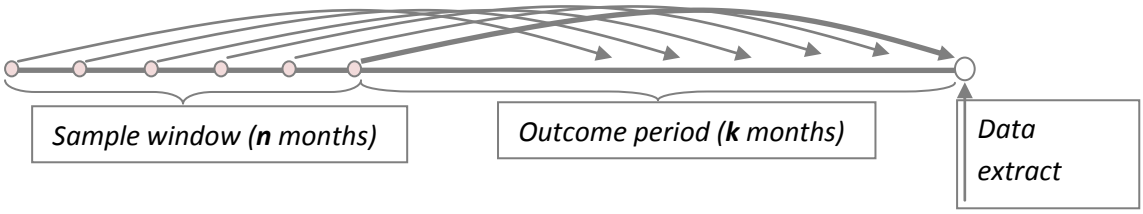


Рис.1 Формування вибірки

Беруться всі кредити видані за певний період в минулому, найближчі до поточної дати, так, щоб в кредитів був час стати «добрими» чи «поганими». Цей фіксований час (k) називається Outcome періодом. Сама вибірка виданих кредитів (за n місяців) називається Sample Window.

Найчастіше розмір Sample Window сягає від 6 до 12 місяців (в нашому випадку – 6 місяців). Вибір розміру такого вікна повинен враховувати сезонності при видачах кредитів та досягнення мінімального обсягу вибірки для побудови.

У випадку якщо найбільш остання історія не є найкращим представником для найближчого майбутнього (наприклад історія кризового періоду для побудови моделі в антикризовий період) можна брати «старіше» Sample window. Основним мотивом в такому випадку є максимальне наближення популяції, на якій будується модель до популяції, на якій вона буде застосовуватись. Таким чином для побудови було обрано період 01.07.2012 – 31.12.2012. Саме цей період можна прийняти як до кризовий в банківській системі України.

Outcome період k найчастіше вибирається як 6-12 місяців (в нашому випадку 6 місяців).

Як для кредитів виданих k місяців тому так і для кредитів виданих $k+1,\ k+2,\ldots$ місяці тому дається однаково час k для того щоб кредит попав у дефолт. Іншими словами, якщо кредит не стає дефолтом протягом перших k місяців – він визначається як «хороший».

Цільова змінна Bad визначається як 1 або 0 якщо кредит став, або відповідно не став дефолтним протягом періоду k.

Дефолтом вважається досягнення протягом періоду k хоча б один раз певного рівня прострочки.

В нашому випадку модель відрізняється від "класичної". Її задача полягає в побудові скорингової карти з неперервною прогнозованою змінною. Результуючою змінною є частка неповернених відсотків по револьверній карті з грейсовим періодом протягом деякого фіксованого «outcome» періоду (6 місяців). Кожен запис по кредиту дублюється 100 разів. Таким чином, якщо рівень повернення відсотків по кредиту складає 73% то 73 записи будуть з "0", а 27 з "1".

Навчальна та тестова вибірки

З метою уникнення так званого "перенавчання" моделі (Overfitting), коли побудована модель показує високі показники предикативності на вибірці, на якій вона будувалася, проте на новій популяції ці показники екстремально падають, робиться розбиття базової вибірки (Sample Window) на:

- Навчальну вибірку (70-80%) – використовується для побудови моделі;
- Тестову вибірку (20-30%) – використовується для перевірки предикативної здатності моделі.

Записи було розбито на дві групи випадковим чином (генерація випадкового числа від 0 до 1 і розділення вибірки по значенню 0.3); можна проводити розбивання більш витонченим методом таким як стратифікація.

Стратифікація полягає у розбитті вибірки на однорідні сегменти по вибраних змінних (стратах). Такими змінними найчастіше виступають: сама результуюча змінна, вік, тип продукту, сума

кредиту тощо. В середині кожної групи випадковим чином вибирається 70% вибірки з ознакою «train sample», решта 30% отримують ознаку «test sample». В такий спосіб після розбиття на тестову та навчальну вибірку зберігаються розподіли кожної із страт.

Предикативна сила

Предикативна сила характеристики визначається з допомогою показника IV (informative value):

$$IV = \sum_{i=1}^{k} (\%good_i - \%bad_i) \circ \ln(\frac{\%good_i}{\%bad_i}),$$

де k – кількість атрибутів відповідної характеристики, $\%good_i$ та $\%bad_i$ – відповідно відсотки добрих та поганих клієнтів, які попали в атрибут «i».

В залежності від значення IV можна зробити висновок про предикативну здатність характеристики:

- менша ніж 0,02 – непредикативна (тобто розподіляється однаково по різних сегментах результуючої змінної);
- від 0,02 до 0,1 – слабка предикативна здатність;
- від 0,1 до 0,3 – середня предикативна здатність;
- 0,3 – висока предикативна здатність

Варто звернути увагу на те, що при екстремально високих значеннях IV (більше 1 для аплікаційних характеристик та більше 3-4 для поведінкових) включення такої характеристики в модель може призвести до overfitting'у (англ. перенавчання) скорингової моделі.

Часто при побудові аплікаційних карт у випадку слабої предикативної здатності аплікаційних змінних варто будувати так звані крос-змінні, які позволяють знаходити нестандартні зв'язки між змінними та посилювати нелінійність моделі.

Для прикладу, при побудові змінна «Стать» та «Освіта» показали слабку предикативну здатність. Проте при побудові крос-змінної вдалося провести значно кращу сегментацію і включити її у модель:

Vars	*Categ*	*Good*	*Bad*	*WOE*	*Badrate, %*	*IV*
APPSEX_APPBIRTHDATE	APPSEX = 0 & APPBIRTHDATE <= 25	25 737	8 712	-0,16819	25,3	0,0012
APPSEX_APPBIRTHDATE	APPSEX = 0 & APPBIRTHDATE <= 40	152 759	40 579	0,07420	21	0,0013
APPSEX_APPBIRTHDATE	APPSEX = 0 & APPBIRTHDATE <= 50	93 184	21 178	0,23021	18,5	0,0069
APPSEX_APPBIRTHDATE	APPSEX = 0 & APPBIRTHDATE > 50	87 462	17 426	0,36182	16,6	0,0149
APPSEX_APPBIRTHDATE	APPSEX = 1 & APPBIRTHDATE <= 47	218 943	75 975	-0,19300	25,8	0,0140
APPSEX_APPBIRTHDATE	APPSEX = 1 & APPBIRTHDATE <= 52	30 259	9 706	-0,11436	24,3	0,0007
APPSEX_APPBIRTHDATE	APPSEX = 1 & APPBIRTHDATE > 52	35 396	10 599	-0,04558	23	0,0001
APPSEX_APPBIRTHDATE	total	643 740	184 175	0,00000	22,2	0,0390

Таблиця 1. Крос-змінна «Стать» та «Освіта»

Крім того показник IV (informative value) для цієї змінної набагато більший ніж сума показників IV кожної із змінних. Такий ефект називається синергетичним ефектом.

Сегментація характеристик

Сегментація (розбиття на атрибути) характеристик, що входитимуть в скорингову модель, відіграє важливу роль для побудови стабільної та предикативної скорингової моделі. Правильна сегментація сприяє збільшенню предикативної сили та стабільності характеристики, а отже і скорингової моделі.

Сегментація проводиться як для неперервних характеристик при розбитті їх на категоріальні так і для категоріальних при наявності

атрибутів з малою часткою в вибірці та атрибутів подібних між собою.

Показник WOE (weight of evidence) для кожного атрибуту характеристики розраховується за формулою

$$WOE_i = \ln(\frac{\%good_i}{\%bad_i}),$$

де $\%good_i$ та $\%bad_i$ – відповідно відсотки добрих та поганих клієнтів, які попали в атрибут «i».

Класична таблиця після розбиття характеристики на атрибути має наступний вигляд:

Vars	*Categ*	*Good*	*Bad*	*%, good*	*%, bad*	*WOE*	*Badrate, %*	*IV*
APPBIRTHDATE	<= 25	45 930	16 020	7%	9%	-0,19812	25,9	0,0031
APPBIRTHDATE	(25;30]	114 782	36 349	18%	20%	-0,10154	24,1	0,0019
APPBIRTHDATE	(30;45]	272 153	80 166	42%	44%	-0,02914	22,8	0,0004
APPBIRTHDATE	(45;50]	76 978	20 350	12%	11%	0,07903	20,9	0,0007
APPBIRTHDATE	> 50	133 897	31 290	21%	17%	0,20236	18,9	0,0077

Таблиця 2. Розбиття характеристики на атрибути

Від'ємні показники WOE вказують на те, що атрибут «поганий» (чим менше – тим гірший), тобто гірший ніж середній показник в портфелі, додатні – що «хороший».

Основні принципи, яких потрібно дотримуватись при сегментації характеристик:

- Проводити сегментацію в такий спосіб, щоб IV кожної характеристики було максимальним (максимізація статистики Хі-квадрат) та зберігалась логічність такої сегментації. Логічність в даному випадку передбачає:
  - Монотонність для неперервних характеристик;

- Бізнес-логіку – відповідність очікуваним результатам (пр. ми очікували що старші клієнти кращі);
- Групування в одні атрибути сегментів з близьким значенням WOE.

Наприклад: Якщо сегмент «вдівець/вдова» має близьке значення WOE до сегменту «неодружений/на», то їх варто об’єднувати в один сегмент, якщо сумарний показник IV характеристики при цьому зменшується несуттєво.

- Намагатись уникати створення малих сегментів (<5% популяції);
- Об’єднувати дуже дрібні сегменти з сегментом, значення WOE якого найближче до нуля;
- Об’єднувати сегмент порожніх значень з сегментом, значення WOE якого найближче до нуля; У випадку якщо такий сегмент значний (> 2-3%) і його значення WOE суттєво відрізняється від нуля – це означає що порожні значення утворюються не випадково. Якщо сегмент таких відмінних порожніх значень можна пояснити і надалі очікується утворення таких порожніх значень, – лише тоді можна виділяти сегмент порожніх значень як окремий атрибут.

В результаті сегментації для побудови моделі було обрано такі характеристики:

- APPBIRTHDATE - вік клієнта у заявці
- APPFAMILYSTATUS - сімейний стан
- APPPOSTCATEG - займана позиція
- APPSEX_APPBIRTHDATE - схрещений параметр статі та віку
- APPSEX_APPEDUCATION - схрещений параметр статі та освіти

– DEP_NUMACT_DEP_SUM - схрещений параметр кількості активних депозитів та їх суми
– APPWPERIODG - робочий стаж
– CHAR10 - кількість місяців до планованої дати закриття
– COMPPO - форма влаштування
– CHR4 - кількість активних депозитів за 2 роки
– FLAT - наявність житла

Кореляція та бізнес-логіка

Включення/не включення змінних, що корелюють між собою є доволі суперечливим питанням серед спеціалістів по скорингу та статистиці, хоча більшість все-таки схиляється до того, щоб забороняти входження в одну модель змінних, кореляція між якими висока (> 0.7).

З одного боку включення змінних що корелюють між собою в одну модель не впливає негативно на її предикативну здатність (при різних видах моделювання: логіт-регресія, класифікаційне дерево, нейронна мережа, тощо), проте поява високої мультиколінеарності між змінними в моделі може призводити до нелогічного (з точки зору бізнес-логіки) розрахунку вагових коефіцієнтів. Приклад: для гірших атрибутів змінної (по bad rate) проставляються вищі бали скоринговою картою ніж для кращих атрибутів.

При побудові моделей логіт-регресії, рекомендовано виключати слабшу по IV з двох характеристик, кореляція між якими перевищує 70%. При дуже великих наборах змінних (більше 100) ще до розрахунку кореляції можна відкидати змінні з меншим IV з груп однотипних змінних.

Результати

У результаті подальшого присвоєння скорингово балу для кожного значення параметру та подальшої оцінки її якості, було визначено що побудована математична модель з індексом GINI 24%.

Динаміка індексу GINI коливається в межах допустимого рівня протягом всього розглянутого періоду, що свідчить про стабільність обраних параметрів.

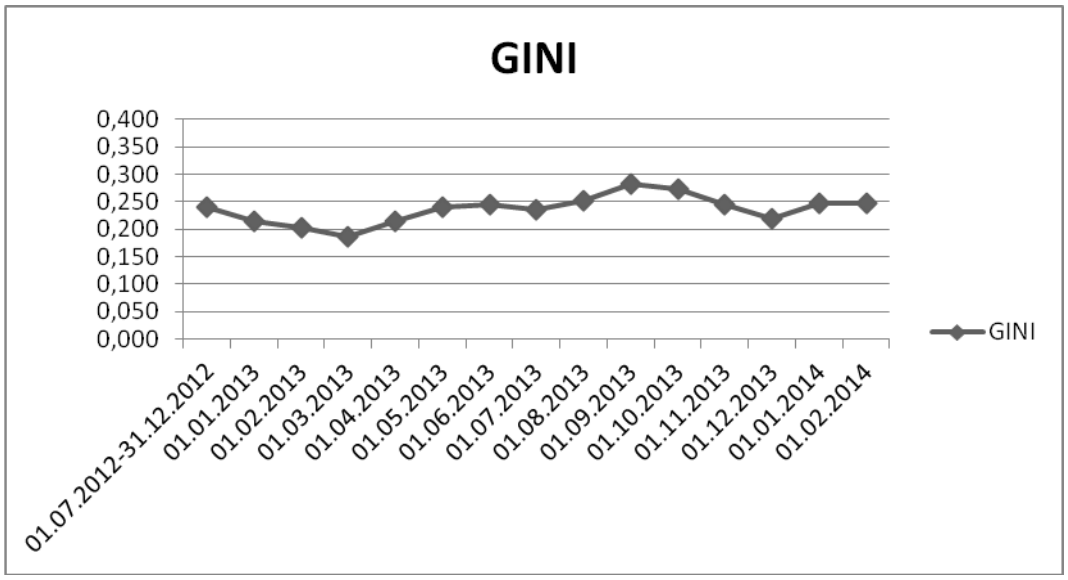


Рис. 2 Динаміка індексу GINI

Згідно зі створеною моделлю, граничні значення скорингового балу можуть коливатися від -92 до 205 (рис. 3). Концентрація популяції у скоринговому балі (розподіл з кроком 20) наведена також на рис. 3.

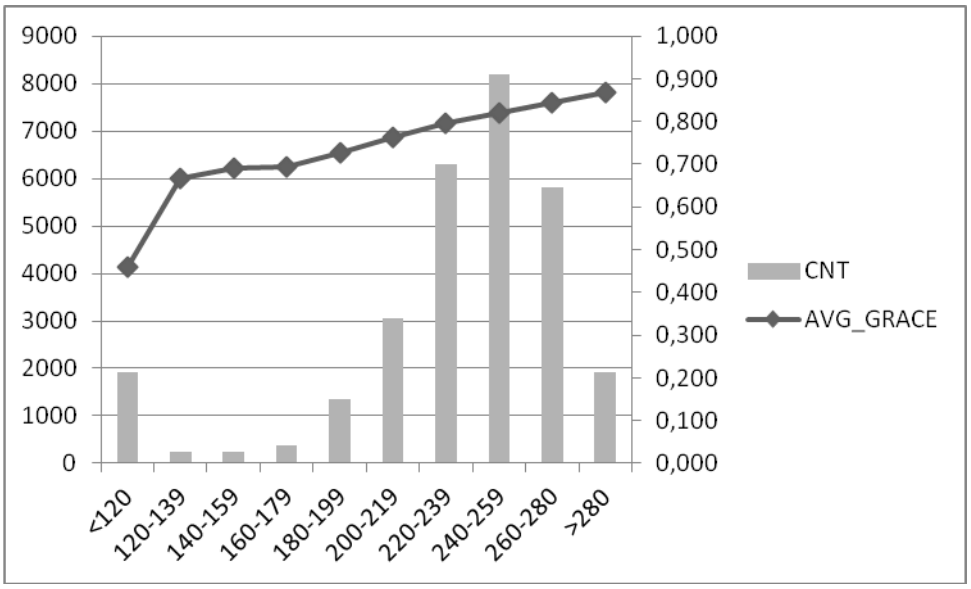


Рис. 3 Концентрація популяції у скоринговому балі

Дану сегментацію можна практично застосовувати починаючи із етапу розробки продукту, закінчуючи прогнозуванням дохідності револьверної карти. Також, коригуючи пункти відтину, можна здійснювати контроль видачі тому чи іншому сегменту популяції і тим самим запобігти втратам, що можуть призвести до значного зниження кінцевого фінансового результату.

Висновки

В роботі розглянуто проблему оцінювання кредитного ризику, визначені причини його виникнення та підходи до обчислення. Проаналізовано супутні ризики в процесі кредитування на прикладі револьверних карток з грейсовим (пільговим) періодом. Розглянуто основні етапи побудови скорингової моделі: формування вибірки, сегментації характеристик тощо.

Результатом роботи стала побудована аплікаційна скорингова модель яка на основі визначених параметрів дає можливість провести сегментацію клієнтів таким чином, щоб банк не зазнав збитків від недоотримання прибутку.

Для побудованої моделі індекс GINI мав значення 24%, що цілком прийнятно для економіки перехідного періоду. Динаміка індексу GINI коливається в межах допустимого рівня протягом всього розглянутого періоду, що свідчить про стійкість побудованої моделі до змін у даних.

Література

1. Особливості мінімізації кредитного ризику банківської установи / І.В. Єлейко, О.В. Сідак // Науковий вісник НЛТУ. – 2011. − Вип. 21.8.
2. *Камінський А.* Експертна модель кредитного скорингу позичальника банку // Банківська справа. – 2006. – № 1. – С. 75–81.
3. *Бондаренко В.* Скоринг-оценка кредитоспособности заемщика // Финансовая консультація. – 2005. – №1-2. – С. 13-16. 4. Бугель Ю. Основні шляхи вдосконалення сучасних методів оцінки кредитоспроможності позичальника // Банківська справа. – 2007. – № 4. – С. 54-59.
4. Рекомендації з приводу оцінки комерційними банками кредитоспроможності і фінансової стабільності позичальника / Національний банк України. № 23011/79 від 02.06.94 р. – [Електронний ресурс]. – Режим доступу: http://www.bank.gov.ua.
5. Бідюк П.І., Матрос Є.О. Моделі оцінки ризиків кредитування фізичних осіб // Кібернетика та обчислювальна техніка. – 2007. – № 153. – С. 87–95.
6. *Міщенко В. І.* Банківські операції: Підручник / В. І. Міщенко, Н. Г. Слав'янська, О. Г. Коренєва. – 2-е вид., переробл. і доп. – К.: Знання, 2007. – С. 280–283.

УДК 004

Аналіз фінансово економічних процесів за допомогою СППР

Вертилецький В.Г.

Науковий керівник: д.т.н., проф. Бідюк П.І

Вступ

Дослідження фінансово-економічних процесів, безсумнівно, вимагає все більшої уваги від наукового суспільства. Зокрема, все актуальнішими питаннями стають моделювання та прогнозування цих процесів, що у Україні, за умов перехідної економіки, набувають нових особливостей. Існуючі методи не завжди дають можливість отримати достатню точності прогнозу. Тому пошук нових методів та комбінації існуючих для досягнення більшої точності та адекватності моделей, що дозволить формувати ефективні управлінські рішення, є вкрай актуальним[1].

Кожний суб’єкт підприємницької діяльності ставить певну низку завдань. Чільне місце в цьому списку займає управління ризиками. Їх поява обумовлена специфікою та особливостями ринкового механізму, зокрема, свободою дій, яка надається кожному суб'єкту господарювання [2].

Основною метою роботи є побудова системи підтримки прийняття рішень (СППР), дослідження, аналіз існуючих та нових моделей прогнозування для макроекономічних та нелінійних фінансових процесів та порівняння методів.

Прогноз повинен визначати можливості, в рамках яких можуть ставитися реалістичні задачі планування розвитку економіки або роботи підприємства[3]. У прогнозуванні використовуються математичні методи, часом досить складні і різні для конкретних завдань. Поширення комп'ютерів та ефективних програмних

продуктів робить прогнозування більш швидким і досить ефективним механізмом аналізу фінансово-економічних процесів.

Одним з вирішальних елементів ефективної організації управління окремими господарюючими суб'єктами та економічними спільнотами є прогнозування внаслідок того, що якість прийнятих рішень великою мірою визначається якістю прогнозування їх наслідків. Тому рішення, прийняті сьогодні, повинні спиратися на достовірні оцінки можливого розвитку досліджуваних явищ та подій в майбутньому[4,5].

Тому в даній роботі розглядалися методи моделювання процесів, які дають можливість прогнозувати нелінійні та інші процеси. На їх основі реалізована СППР, із допомогою якої можливо визначати найкращі моделі з мінімальними на це затратами ресурсів. Також передбачено інтуїтивно зрозумілий інтерфейс для роботи з програмним продуктом

Постановка задачі

Виконати аналіз існуючих та нових моделей для прогнозування нелінійних макроекономічних та фінансових процесів і порівняння методів. Створення СППР для прогнозування та моделювання нелінійних процесів довільної природи на основі емпіричної вибірки даних. Порівняти розроблену СППР з вже існуючими комерційними продуктами-аналогами.

В роботі наведено результати прогнозування обраних процесів за допомогою таких методів, як побудова різнотипових регресійних моделей, МГУА. На основі цього дослідження здійснено порівняльний аналіз методів між собою, досліджено ефективність роботи розробленої СППР у порівнянні з комерційними продуктами-аналогами.

Методика прогнозування та побудови математичних моделей

Множинна регресія

Множинна регресія широко використовується при рішенні питань попиту, доходності акцій, при вивченні витрат виробництва, у макроекономічних розрахунках і тощо.

Загальна множинна регресійна модель має наступний вигляд:

$$y = f(x_1, x_2, ...x_p) + \varepsilon, \quad (1)$$

де y - залежна змінна;

$x_1, x_2, ...x_p$ - фактори (незалежні змінні).

Якщо множинна регресійна модель є лінійною (ЛМР), то вона подається у вигляді:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p + \varepsilon. \quad (2)$$

Позначимо і-е спостереження змінної y через y_i, а факторів - $x_{i1}, x_{i2}, ...x_{ip}$. Відтоді модель (1) можна подати у вигляді:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + \varepsilon_i, i = \overline{1,n}, \quad (3)$$

або у матричній формі:

$$y = Xb + \varepsilon, \quad (4)$$

де $y = [y_1, y_2, ...y_n]^T$ - вектор (матриця-стовпець) значень залежної змінної;

$b = [b_1, b_2, ...b_p]^T$ - вектор (матриця-стовпець) коефіцієнтів регресійної моделі;

$\varepsilon = [\varepsilon_1, \varepsilon_2, ...\varepsilon_n]^T$ - вектор (матриця-стовпець) похибок;

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$ - матриця значень факторів.

Відзначимо основні припущення регресійного аналізу:

В моделі похибка ε_i (або залежна змінна y_i) є випадковою величиною, а фактори x_{ip} невипадкові величини ($i=\overline{1,n}$).

Математичне сподівання похибки ε_i дорівнює нулю:

$$M\left[\varepsilon_i\right]=0, i=\overline{1,n}. \quad (5)$$

Дисперсія похибки ε_i (або залежної змінної y_i) постійна дія будь-якої:

$$D\left[\varepsilon_i\right]=\sigma^2, \quad (6)$$

тобто виконується умова гомоскедастичності.

Похибки ε_i та ε_j не корельовані:

$$M\left[\varepsilon_i\varepsilon_j\right]=0, i\neq j. \quad (7)$$

Похибка ε_i (або залежна змінна y_i) є нормально розподіленою випадковою величиною.

Матриця значень факторів невироджена, тобто її ранг дорівнює $p+1$

$$rangX=p+1<n. \quad (8)$$

Модель, для якої виконуються припущення, називається класичною нормальною лінійною моделлю множинної регресії (CNLMR-model).

Оцінкою цієї моделі за вибіркою є рівняння регресії:

$$y=b_0+b_1x_1+b_2x_2+...+b_px_p, \quad (9)$$

де y - оцінка математичного сподівання залежної змінної $M_x\left[y\right]$;

$b_i(i=\overline{0,p})$ - оцінка коефіцієнтів регресійної моделі (або коефіцієнти регресії). Як і раніше, для оцінки коефіцієнтів CNLMR-model використовують МНК:

$$S(b_0,b_1...b_p)=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-...-b_px_{ip})\rightarrow \min. \quad (10)$$

Після розв'язання системи нормальних рівнянь:

$$\begin{cases} \dfrac{ds}{db_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip}) = 0, \\ \dfrac{ds}{db_1} = -2\sum_{i=1}^{n} x_{i1}(y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip}) = 0, \\ \dfrac{ds}{db_p} = -2\sum_{i=1}^{n} x_{ip}(y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip}) = 0, \end{cases} \quad (11)$$

отримаємо значення коефіцієнтів рівняння регресії, які в матричній формі мають вигляд:

$$b = (X^T X)^{-1} X^T y, \quad (12)$$

де $b = \left[b_1, b_2, ..., b_p\right]^T$ - вектор (матриця-стовпець) коефіцієнтів рівняння регресії.

Оцінки b_j є незміщеними, обґрунтованими та ефективними.

Оцінка дисперсії похибок (13)є незміщеною.

$$S^2 = \frac{\sum_{i}^{n}(y_i - y_i)^2}{n - p - 1}. \quad (13)$$

Коефіцієнт (індекс) множинної кореляції R використовується для оцінки тісноти спільного впливу факторів на залежну змінну:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \quad (14)$$

Властивості коефіцієнта множинної кореляції R:

1. Коефіцієнт множинної кореляції приймає значення на відрізку$[0,1]$, тобто $0 \le R \le 1$.

Чим ближче R до одиниці, тим тісніший зв'язок між залежною змінної y та факторами $x_1, x_2, ...x_p$

2. При $R = 1$ кореляційний зв’язок є лінійною функціональною залежністю.

3. При $R = 0$ лінійний кореляційний зв’язок відсутній.

Щодо оцінки ступеня взаємозв’язку можна керуватись аналогічними емпіричними правилами, як і для випадку ЛПР.

Побудова СППР

Система підтримки прийняття рішень або СППР — це комп'ютерна система, яка шляхом збору і аналізу великої кількості інформації може впливати на процес ухвалення рішень організаційного плану в бізнесі чи підприємництві. Інтерактивні системи дозволяють керівникам отримати корисну інформацію з першоджерел, проаналізувати її, а також виявити існуючі бізнес-моделі для вирішення певних завдань.

Реалізована СППР є простою з погляду архітектури, тому її впровадження буде доцільним в організаціях, що не ставлять перед собою глобальних завдань, але орієнтуються на високий рівень розвитку інформаційних технологій.

Архітектура створеної СППР налічує наступні рівні:

1) завантаження і обробка даних;
2) аналіз даних;
3) побудова та вибір кращої моделі;
4) прогнозування.

Розглянемо кожен із перерахованих рівнів:

- перший рівень надає можливість завантаження даних шляхом імпорту із текстового файлу чи ручним вводом. Після чого можливе перетворення даних, з метою усунення їх надлишковості, та підготовка даних до аналізу;
- другий рівень забезпечує можливість візуальної оцінки даних, проведення статистичного та кореляційного аналізу.

- третій рівень надає засоби для побудови моделей авто регресії ковзкого середнього. Надає можливість оцінки параметрів якості моделі для вибору кращої.

- четвертий рівень реалізує динамічне та статистичне прогнозування на базі створеної авторегресії з ковзним середнім (АРКС) моделі.

Рисунок 1 – Рівні архітектури створеної ІСППР

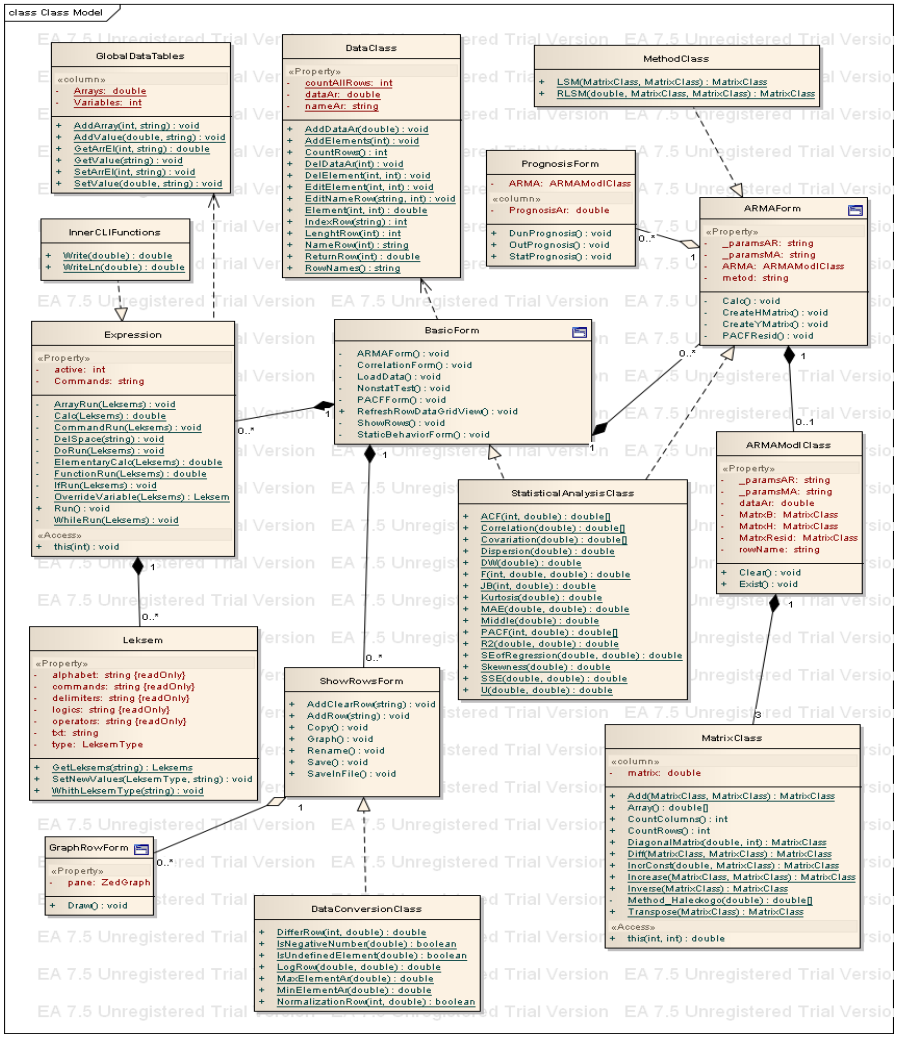


Рисунок 2 – Діаграма класів створеної СППР

Результати дослідження

Для аналізу та функціонування СППР було обрано часовий ряд Export — загальна кількість значень становила 60.

Export (Експорт) — вивіз із митної території країни за кордон товарів і послуг без зобов'язання їхнього зворотного повернення. Факт експорту фіксується в момент перетину товаром митного кордону, надання послуг та ін.

Розглянемо грошовий агрегат EXPORT, графік якого представлено на рис. 3

Рисунок 3 – Графік показника EXPORT

Обчислимо і представимо графічно АКФ та ЧАКФ заданих рядів та їх статистичні характеристики EXPORT (рис. 4-5):

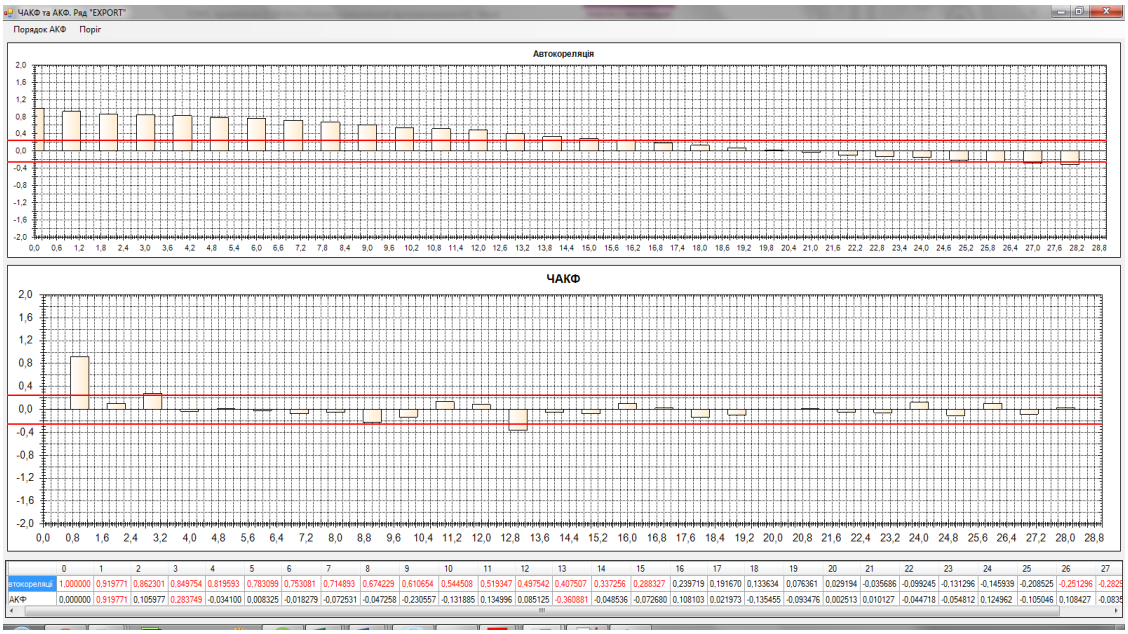


Рисунок 4 – Графік АКФ і ЧАКФ для EXPORT

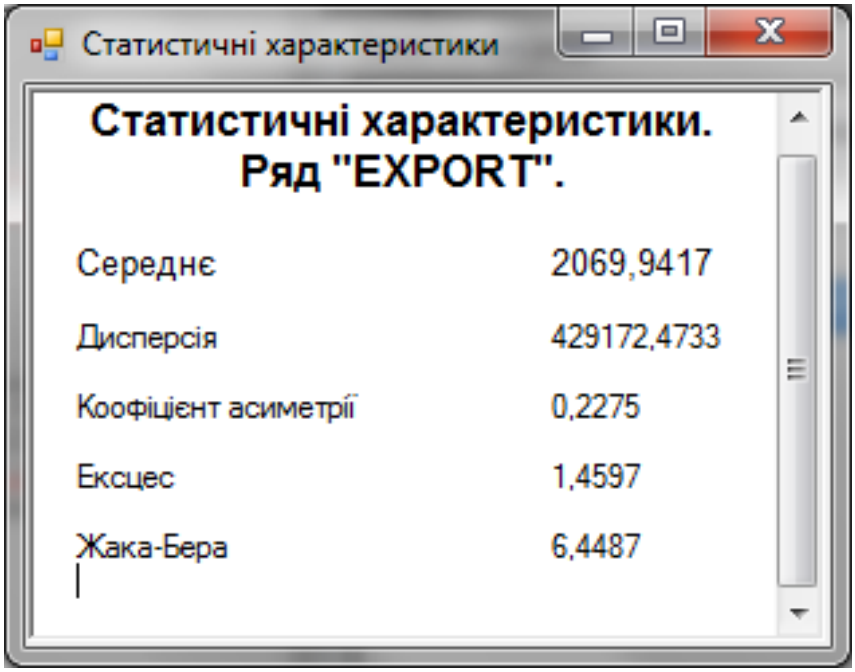


Рисунок 5 – Статистичні характеристики EXPORT

Побудуємо АР першого порядку для ряду EXPORT:

$$y(k) = 113,6688 + 0,9604\,y(k-1) + e(k).$$

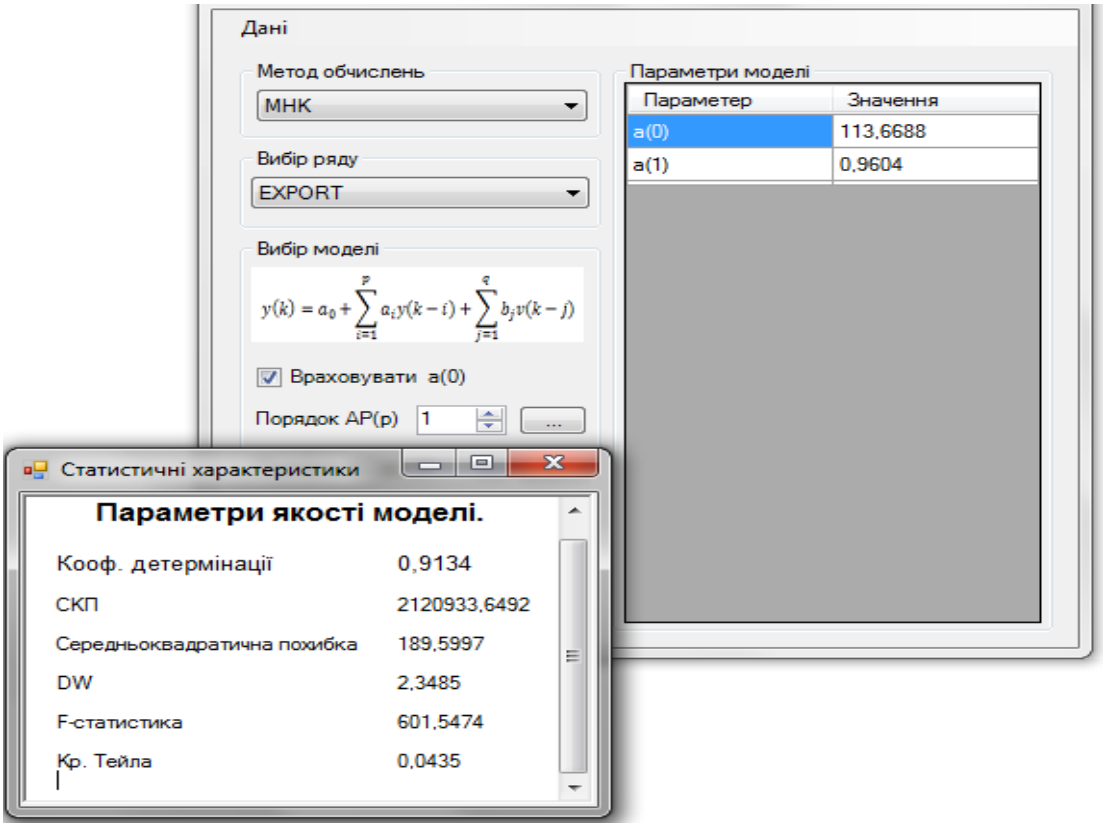


Рисунок 6 – Оцінка для АР 1-ого порядку

Спрогнозуємо на 4 кроки (рис. 7):

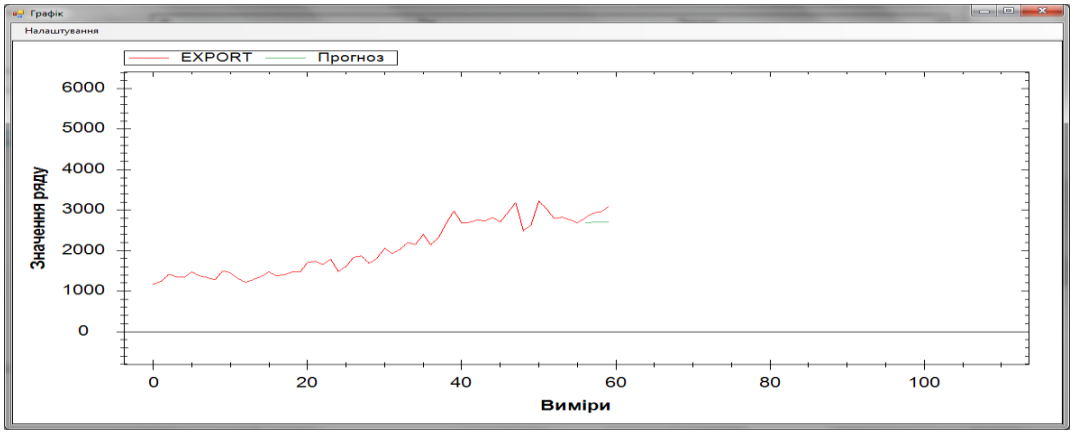


Рисунок 7 – Графік прогнозування для АР 1-ого порядку

У даному випадку критерії якості прогнозів приймають такі значення:

$СеКП = 259,9427; САП = 243,01276; САПП = 8,1569309; U = 0,04602.$

Побудуємо АР 3-ого порядку (рис. 8).

$$y(k) = 64,6731 + 0,6844 * y(k-1) - 0,2133 * y(k-2) + 0,5241 - y(k-3) + e(k).$$

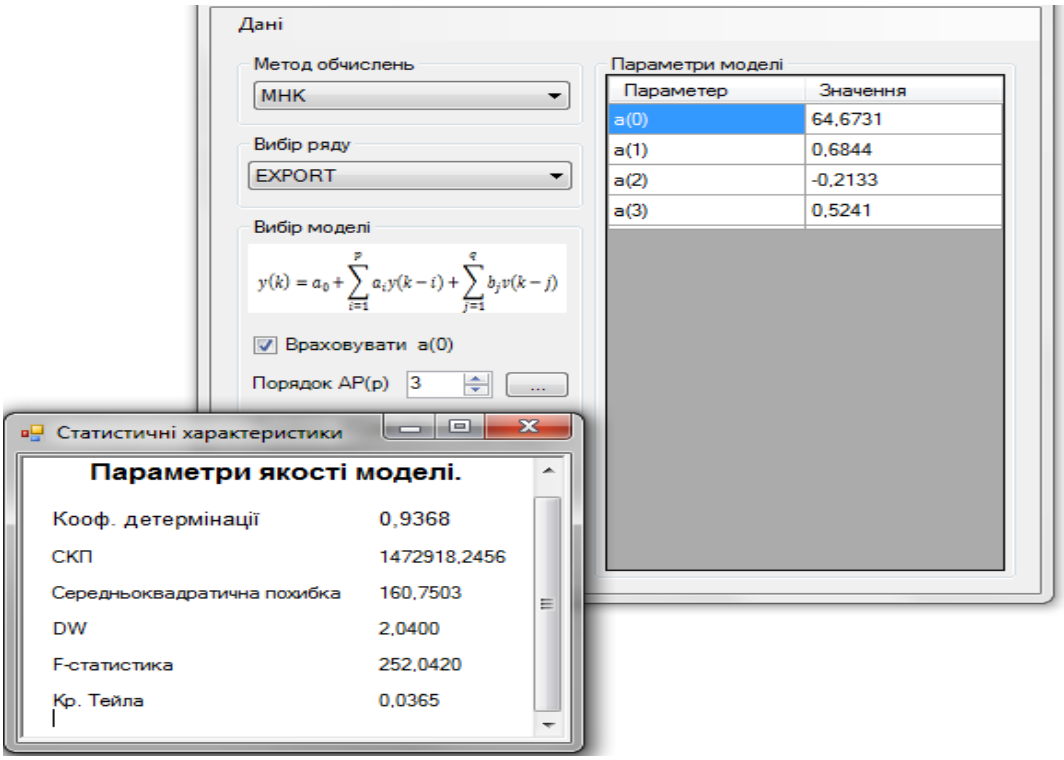


Рисунок 8 – Оцінка для АР 3-ого порядку

Обчислимо оцінки прогнозів на 4 кроки (рис. 9).

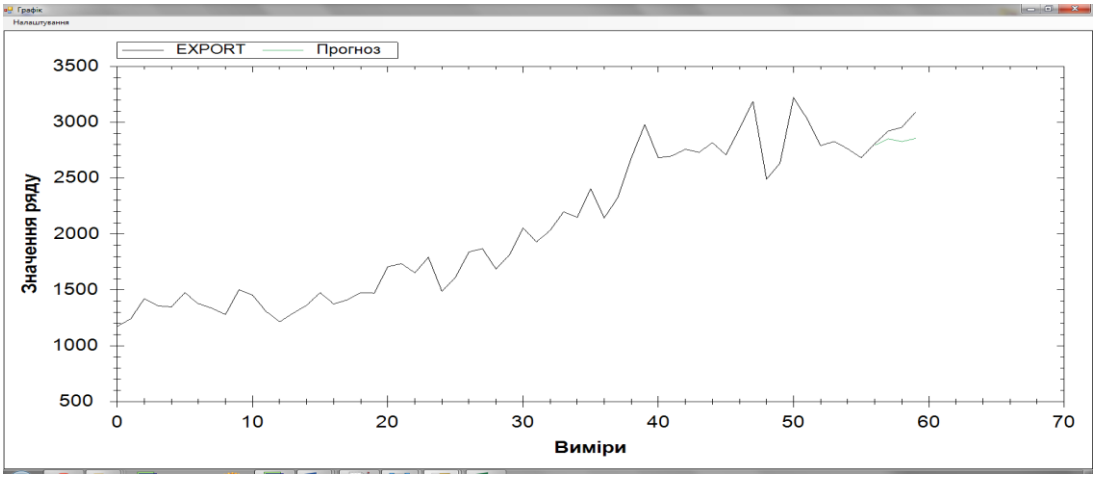


Рисунок 9 – Графік прогнозування для АР-3-ого порядку

Критерії якості оцінок прогнозів :

$$СеКП = 137,7216;\ САП = 111,42509;\ САПП = 3,69746;\ U = 0,02383.$$

Побудуємо АР 13-ого порядку (рис.10).

$$y(k) = 161,0277 + 0,5960 * y(k-1) - 0,1379 * y(k-2) + 0,4599 * y(k-3)$$
$$-0,0518 * y(k-4) + 0,2384 * y(k-5) + 0,0896 * y(k-6)$$
$$-0,0078 * y(k-7) - 0,0005 * y(k-8) - 0,2408 * y(k-9)$$
$$-0,1509 * y(k-10) - 0,0912 * y(k-11) + 0,6740 * y(k-12)$$
$$-0,4341 * y(k-13) + e(k).$$

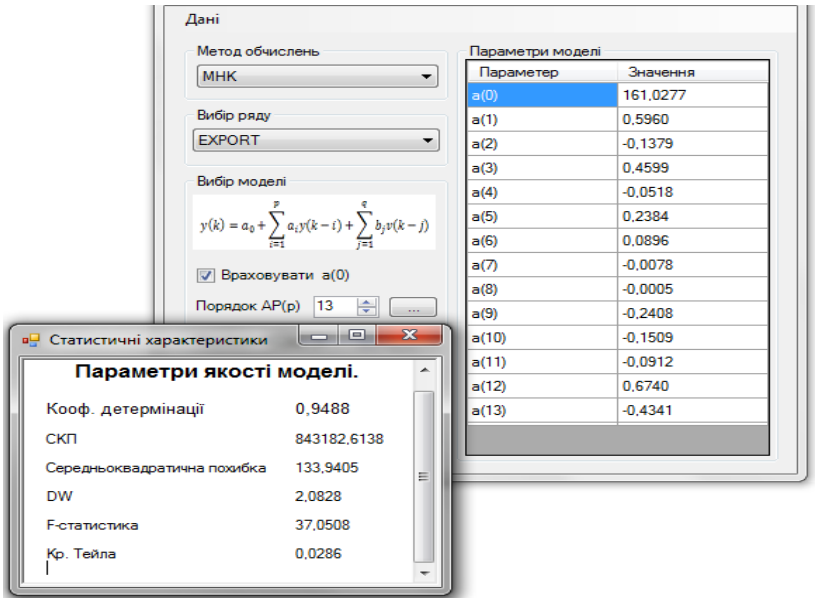


Рисунок 10 – Оцінка для АР 13-ого порядку

Побудуємо прогноз на 4 кроки (рис. 11).

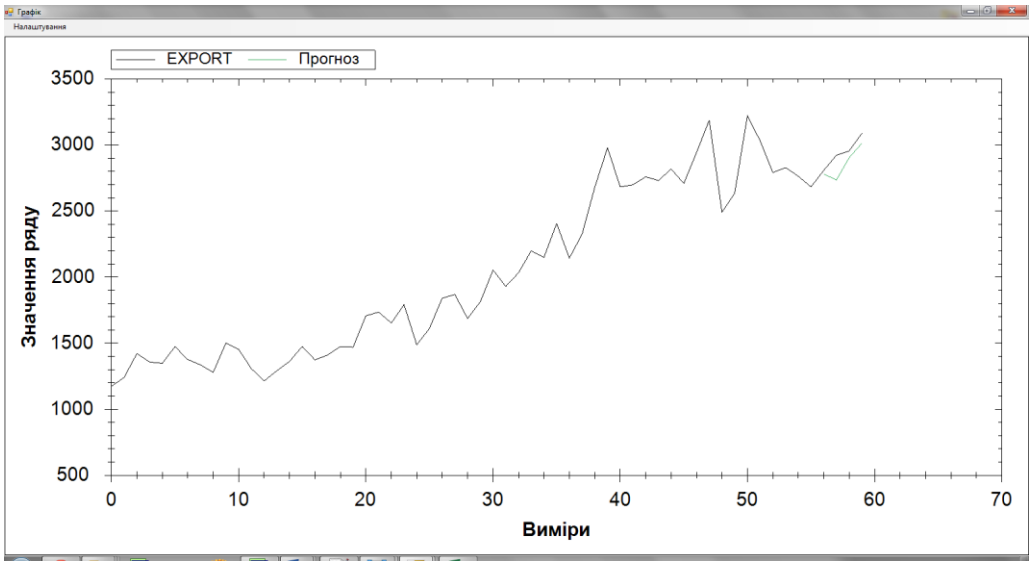


Рисунок 11 – Графік прогнозування для АР 13-ого порядку

Якість оцінок прогнозів характеризуються критеріями:

$СеКП = 105,0212$; $САП = 85,51131$; $САПП = 2,895727$; $U = 0,018085$.

Як можемо зазначити, показники моделі і прогнозу ще покращились, про це свідчить зокрема зменшення середньої абсолютної похибки у %, яка покращилась з $3,697462304\%$ до $2,895727\%$

Спробуємо видалити незначущі коефіцієнти, а саме $a(2)$, $a(4) - a(12)$.

$$y(k) = 145,6002 + 0,5731 * y(k-1) + 0,4327 * y(k-3) - 0,0487 * y(k-4) + e(k).$$

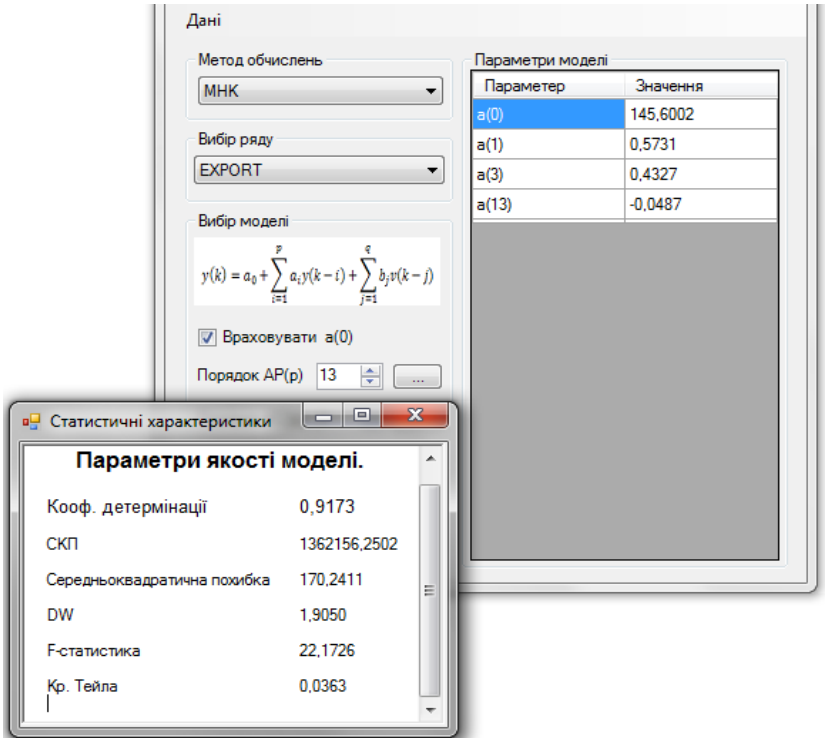


Рисунок 12 – Оцінка для АР 13-ого порядку з видаленням незначущих параметрів

Параметри якості погіршились, спрогнозуємо на 4 кроки (рис. 13).

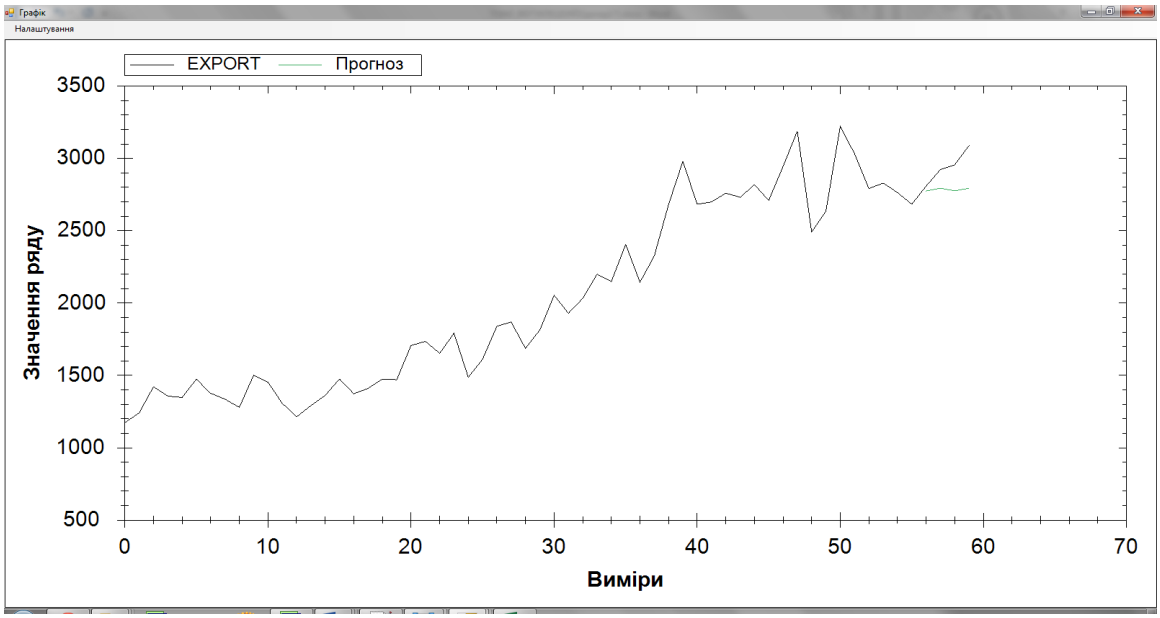


Рисунок 13 – Графік прогнозування для АР 13-ого порядку з видаленням незначущих параметрів

$СеКП = 185,35439$; $САП = 159,65316$; $САПП = 5,3206726$; $U = 0,0323435$.

Показник якості лише погіршились, отже видалення коефіцієнтів не потрібне .

В таблиці 1 зведені характеристики математичних моделей, побудованих для часового ряду EXPORT, і характеристики прогнозів на 4 кроки, обчислених на основі цих моделей. Ця таблиця дає можливість оперативно порівняти результати моделювання та прогнозування, а також встановити можливість практичного використання результатів.

Таблиця 1 - Результати моделювання і прогнозування на 4 кроки часового ряду EXPORT

Тип моделі	Характеристики моделі			Характеристики прогнозу			
	R^2	$\sum e^2(k)$	DW	$СеКП$	$САП$	$САПП$	Коеф-т Тейла
АР(1)	0,9134	2120933,6	2,348	259,9428	243,0128	8,156931	0,046028
АР(3)	0,936	1472918,2	2,040	137,7216	111,4251	3,697462	0,023831
АР(13)	**0,9488**	**843182,61**	**2.082**	**105,0212**	**85,51131**	**2,895727**	**0,018085**
АР(13)w DEL	0,9173	1362156,2	1.905	185,3544	159,6532	5,320673	0,032344

Результати моделювання, наведені в таблиці 1, свідчать про те, що всі моделі, побудовані для EXPORT, є придатними для прогнозування, оскільки коефіцієнт Тейла вимірюється сотими долями. Характеристики прогнозу на 4 кроки для моделей різного порядку відрізняються суттєво. Найкращі характеристики щодо прогнозування має модель АР(13) (САПП = 2,89%).

Застосування МГУА

Спробуємо проаналізувати та спрогнозувати часовий ряд EXPORT використовуючи додатково грошовий агрегат М3 за допомогою МГУА.

Отримаємо модель:

$$Y1 = 1,69514e^{-10} - 7,74775e^{-7} - 1,66909 + x_1[t-6]*(-3,4706) + " x_1[t-6],$$

$$cubert"* 2077,11.$$

Характеристики моделі наведені в таблиці 2.

Таблиця 2 - Характеристика моделі

Результати підготовки даних	Навчання	Прогноз
Число спостережень	56	4
Макс. негативне відхилення	-414,043	-355,053
Макс. позитивне відхилення	370,994	44,834
Середній модуль похибки	171,783	166,514
Середньоквадратичне відхилення	208,309	203,605
Сума відхилень	-5,45697E-12	-576,388
Стандартне відхилення залишків	208,309	143,844
Коефіцієнт детермінації (R^2)	0,48713	-3,15875
Кореляція	0,697947	-0,980357

Зобразимо графік прогнозу на 4 кроки (рис. 14).

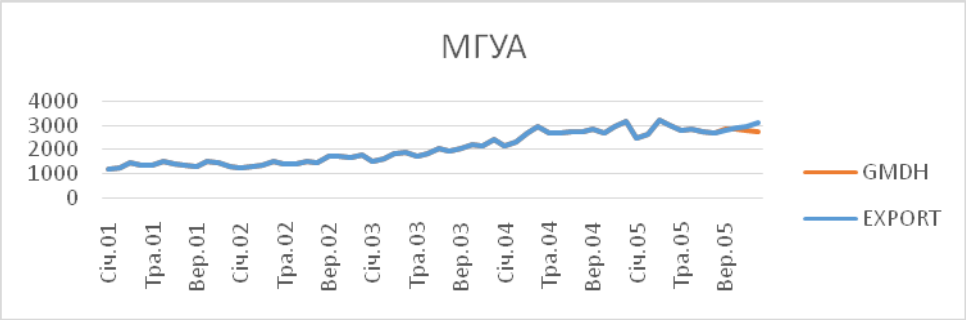


Рисунок 14 – Застосування МГУА для ряду EXPORT

$$СеКП = 203,6049; САП = 166,5138; САПП = 5,5328; U = 0,0354.$$

Як бачимо застосування МГУА для даного часового ряду дає дещо гірші значення ніж АР(13). САПП, коефіцієнт Тейла мають прийнятні

показники, що дає нам зрозуміти про придатність даної моделі до подальшого прогнозу. Але в той же час характеристики моделі мають посередні значення наприклад $R^2 = 0,48$. Маючи дані за 60 періодів, був зроблений прогноз на 4 кроки. Прогнозування велось багатьма методами. Такими як, АР, АРКС, АР з відхиленням від середнього, метод групового врахування аргументів Були зроблені певні висновки щодо побудованих моделей. Але щоб краще оцінити прогнози зроблені вище, зобразимо їх всі в таблиці 3 для часового ряду EXPORT.

Зобразимо тепер зведені результати для ряду EXPORT.

Таблиця 3 - Зведена таблиця прогнозів для часового ряду EXPORT

Тип моделі	Характеристики моделі			Характеристики прогнозу			
	R^2	$\sum e^2(k)$	DW	$СеКП$	$САП$	$САПП$	Коеф-т Тейла
АР(1)	0,9134	2120933,64	2,348	259,942	243,0128	8,156	0,046028
АР(3)	0,936	1472918,24	**2,040**	137,721	111,4251	3,697	0,023831
АР(13)	**0,9488**	**843182,613**	2.082	**105,021**	**85,51131**	**2,895**	**0,018085**
АР(13)w DEL	0,9173	1362156,25	1.905	185,354	159,6532	5,320	0,032344
МГУА	0,48713	—	—	203,604	166,513	5,532	0,03542

З таблиці 3 слідує що все ж таки найкращим для часового ряду EXPORT є авторегресія 13-ого порядку. Показники САПП и коефіцієнти Тейла показує що всі запропоновані моделі можна в подальшому використовувати для прогнозування.

ВИСНОВКИ

У даномудослідженні виконано загальний огляд вибраних методів моделювання і прогнозування стаціонарних та нестаціонарних процесів, виконано аналіз їх переваг та недоліків. В рамках розробленої СППР реалізовані такі методи: побудова авторегресій різних порядків та авторегресій з ковзним середнім

Моделі АР та АРКС мають чіткий теоретично обґрунтований алгоритм, за яким здійснюється побудова якісної моделі. До недоліків можна віднести нечіткість і складність вибору порядку моделі, що ускладнює моделювання їх на ЕОМ. Вдалим виходом із ситуації є застосування експертних граничних значень параметрів.

Виконано аналіз впливу різних порядків регресійних моделей на якість оцінок прогнозів і зроблено порівняння з МГУА.

Серед методів моделювання і прогнозування фінансово-економічних процесів можна виділити метод групового врахування аргументів, але через громіздкість його використання не завжди є доцільним. Розроблені авторегресійні моделі виявилися достатньо точними при досягненні високих порядків. Для часового ряду EXPORT найкращі результати показала АР високого порядку, а саме АР(13). При цьому МГУА дав не найкращі результати. З цього можна зробити висновок, що не існує універсального метода для прогнозування фінансово-економічних процесів.

У порівнянні з програмними продуктами-аналогами (зокрема, програмою Eviews) розроблена СППР засвідчила високу точність отриманих результатів прогнозування. Тобто можна зробити висновок про доцільність створення власних систем для моделювання і прогнозування.

Література

1. Небава М.І. Теорія макроекономіки: Навч. посіб. — Київ: Слово, 2005. – 536 с.

2. Вугальтер А. Л. Фундаментальна економія. Динаміка. — Москва : Экономика, 2007. — 371 с.

3. Бідюк П.І. Коршевнюк Л.О. Проектування комп'ютерних інформаційних систем підтримки прийняття рішень – Київ: - 2010. – 340 с.

4. Кондратьев Н. Д. Проблемыэкономическойдинамики. – Москва: Экономика, 1989. – 526 с.;

5. Матвєєва Т. Ю. Введення в макроекономіку.— "Видавничий дім ГУ ВШЕ": 2007. — 511 с.

6. Добробуш Р., ФішерС., Макроекономіка. — Київ :Основи, 1996. — 93 с.

7. Лугінін О.Є. Економетрія: Навч.посібник 2-ге видання перероб. та доп. —Київ: Центр учбової літератури, 2008. — 278 с.

8. Бідюк П. І., Романенко В.Д., Тимощук О.Л. Аналіз часових рядів: Навч.посіб. – Київ: Політехніка, 2010. – 317с.

9. Молчанов И. Н., Герасимова И. А. Компьютерный практикум по начальному курсу эконометрики (реализация на Eviews): Практикум – Ростов-н/Д.: –2001. – 58 с.

УДК 004.8

Використання алгоритму CLOPEдля структурування доменної моделі адаптивної системи навчання

О.О.Крюков, К.В.Козлов, Д.Г. Діденко

Вступ

Актуальним завданням в комп'ютерному навчанні є вивчення та розробка механізмів адаптації до індивідуальних особливостей учнів. Реалізація таких механізмів - головне призначення адаптивних та інтелектуальних навчальних систем. До їх складу входять моделі предметної області, системи контролю, якого навчають адаптації фахівця (експерта, вчителя)[1].

Оверлейні моделі[2,3]ґрунтуються на структурі предметної області, яка розбивається на елементарні одиниці знань - концептиабо теги. Модель виконує накладення знань користувача на структуру предметної області. Для кожного користувача створюється індивідуальна оверлейна модель, що зберігає числові характеристики за всіма концепту. Характеристикою може бути оцінка, ймовірність або булевське значення, виставлені в процесі тестування або іншої форми опитування за знання кожного концепту. Залежно від ієрархії предметної області характеристики можуть бути нерівнозначними. Зазвичай їх ділять на загальні (інтегровані) і поточні (відображають засвоєння поточного матеріалу)[2]. Таким чином, зміст моделі виходить багаторівневим. На вищому рівні можна подивитися глобальні характеристики знань учня, наприклад, рівень знань по всьому курсу.

В даній роботі пропонується метод автоматизації побудови такої структури предметної області, яка розбита на елементарні одиниці знань та

згрупована на рівномірно розділені частини, за допомогою категоріального методу кластеризації CLOPE.

Постановка задачі

Метою роботи є створення алгоритму структурування доменної моделіадаптивної системи навчання.

Для досягнення мети було поставлено наступні задачі:

1. Запропонувати модифікований варіант алгоритму CLOPEдля генерації структури доменної моделісистеми адаптивного навчання.
2. Провести розбиття набору питань за допомогою базового та алгоритму CLOPE.
3. Провести розбиття питань за допомогою модифікованого алгоритму CLOPE.
4. Аналіз результатів роботи базового та модифікованого алгоритмівCLOPEта обґрунтування переваг запропонованого методу.

Кластеризація питань і тем на модулі

Після створення питань, правил та маркування їх тегами постає наступна складна задача, яка потребує автоматизації - розбиття всіх матеріалів на модулі та структурування на основі цих модулів процесу навчання студентів.

Вручну розбити навіть декілька сотень питань на модулі по складності є дуже трудомісткою задачею з точки зору часу та якості. Модулі повинні бути рівнозначними по розміру та впорядковані за змістом. Складність цього процесу зростає експоненційно в залежності від кількості питань з якими доводиться працювати[4].

Для розбиття множини питань, яка вже помічена тегами, на модулі питань за тематиками пропонується використати методи кластеризації якісних характеристик.

Задача розбиття питань на кластери за схожістю між собою, тобто на модулі, є задачею кластеризації категоріальних даних.

Для початку, формалізуємо розглянуту задачу кластеризації для категоріальних даних. У нас в наявності є база транзакційних даних, за допомогою алгоритму CLOPEможна розбивати на кластери будь-які категоріальні масиви, працюючи з ними як з транзакційними.

Під терміном транзакція тут розуміється деякий довільний набір об'єктів, будь це список ключових слів статті, товари, куплені в супермаркеті, множина симптомів пацієнта, характерні фрагменти зображення і так далі, в нашому випадку це питання з набором тегів. Завдання кластеризації транзакційних даних полягає в отриманні такого розбиття всієї множини транзакцій, щоб схожі транзакції опинилися в одному кластері, а ті що відрізняються один від одного - у різних кластерах.

В основі алгоритму кластеризаціїCLOPEлежить ідея максимізації глобальної функції вартості, яка підвищує близькість транзакцій в кластерах за допомогою збільшення параметру кластерної гістограми.

Застосування цього алгоритму дозволить автоматичне розбиття всього матеріалу для кожного з розділівна модулі(кластери) за схожістю та близькістю тематик(тегів), які в ньому вивчаються, та контролювати їх розмір за допомогою коефіцієнта відштовхування (repulsion)[5].

У міру того як збільшується кількість контенту та кількість відповідей на питання, ми повинні динамічнооновлювати розбиття матеріалу на модулі (їх кількість може збільшуватись - тобто з одного більш абстрактного модуля можуть з’явитися декілька більш детальних по тематикам, які в ньому вивчаються).

Модифікація базовогоалгоритмуCLOPE

Так як базовий алгоритм не передбачає завідомо визначену кількість кластерів для розбиття (в рамках даної задачі - тем), необхідно створити механізм для попередження ситуацій створення розріджених кластерів та кластерів з відносно великою кількістю питань та зробити так, щоб елементи в кластерах розподілялися рівномірно. Для вирішення даної задачі пропонується модифікувати базову функцію корисності таким чином, щоб значення функції корисності в залежності від кількості елементів в кластері зростала нелінійно і при збільшенні кількості питань більше деякого порогу, сповільнювала свій зріст.

В якості функції, що буде обмежувати зріст глобальної функції корисності в залежності від кількості питань в кластері було розглянуто$\log_b N$, де b - параметр, що визначає швидкість росту корисності в залежності від кількості питань, N - кількість питань в кластері.

Загальний вигляд модифікованої функцій корисності:

$$Profit(C) = \frac{\sum_{i=1}^{k} \frac{S(C_i)}{W(C_i)^r} \times \log_b(|C_i| + 1)}{\sum_{i=1}^{k} |C_i|}.$$

Дана модифікація дозволяє обмежити верхню границю кількості елементів в кластері. Для введення нижньої границі кількості елементів в кластері вводиться параметр minClusterSize, що задає мінімальну кількість питань в окремому кластері. В базовий алгоритм додається ще один етап в рамках якого розріджені кластери, кількість елементів в яких менше границі minClusterSize перерозподіляються між іншими кластерами.

Загальний вигляд додаткового етапу алгоритму:

1. Сортування кластерів за кількістю елементів за зростанням.
2. Вибираємо перший кластер із сортованого списку. Якщо кількість елементів в ньому менше заданого порогу minClusterSize, то на крок 3, інакше - завершення алгоритму.

3. Ітеративний перерозподіл елементів кластера в інші кластери (для кожного елемента вибираємо кластер, який дає найбільший приріст функції корисності). Видаляємо кластер зі списку. Переходимо на крок 1.

Критерії якості кластеризації

Алгоритм CLOPE не передбачає розрахунок відстані між окремими об’єктами в кластері та міжкластерної відстані, а використовує глобальну функцію корисності. Враховуючи це, а також факт, що розбиття повинно мати особливі ознаки для вирішення конкретної задачі, для оцінки було вибрано зовнішні критерії якості: індексиRand, JaccardtaFM.

Дані метрики дозволяють виявити наскільки отримане розбиття відповідає еталонному. Нехай нам дана деяка навчальна вибірка X, для якої правильна кластеризація C відома. За допомогою деякого алгоритму отримана кластеризація K. Необхідно перевірити наскільки K співпадає з C. Підрахуємо кількість пар елементів (x_i, x_j), для яких:

- елементи належать одному кластеру з C і одному кластеру з K : SS;
- елементи належать одному кластеру з C і різним кластерам з K : SD;
- елементи належать різним кластерам з C і одному кластеру з K : DS;
- елементи належать різним кластерам з K і різним кластерам з C : DD.

Тоді індекси відповідності розраховуються як:(ss + dd) / (ss + sd + ds + dd).

1. Rand:

$$RI = \frac{SS + DD}{SS + SD + DS + DD}.$$

2. Jaccard:

$$J = \frac{SS}{SS + SD + DS}.$$

3. FM (Folkes and Mallows index):

$$FM = \sqrt{\frac{SS}{SS + SD} \times \frac{SS}{SS + DS}}.$$

Приклад застосування методики

В якості вхідного масиву для кластеризації маємо 100 питань. Кожне питання у вибірці протеговане і має від 2 до 9 тегів. Питання попередньо розбиті на теми експертом на 4 модулі. Кількість питань в кожному модулі24-26. Формат запису у вхідному файлі:

[question_id, tags: {tag_1, tag_2, .. , tag_i, ..tag_N}],

де***question_id*** - унікальний ідентифікатор питання,

tags - набір тегів, що відповідає питанню в даному записі.

В якості критеріїв оцінки розбиття будемо використовувати зовнішні метрики оцінки схожості кластерів, такі як: Rand, Jaccard та FMіндекси. Еталонним розбиттям, що використовується для обчислення метрик, виступає розбиття, реалізоване експертом.

Табл. 1. Еталонне розбиття вибірки питань

Номер кластера	**Ідентифікатори питань**
1	571, 595, 574, 578, 527, 605, 566, 586, 554, 526, 542, 547, 577, 589, 518, 611, 560, 614, 533, 583, 569, 543, 557, 607
2	524, 581, 519, 558, 573, 516, 515, 530, 563, 592, 613, 598, 561, 546, 567, 551, 575, 570, 609, 591, 594, 584, 562, 564, 601
3	588, 517, 565, 555, 541, 531, 520, 587, 576, 544, 556, 608, 545, 599, 602, 549, 540, 539, 538, 548, 550, 597, 612, 593, 536
4	596, 528, 522, 582, 521, 529, 600, 525, 590, 537, 579, 523, 534, 559, 553, 532, 610, 585, 568, 552, 580, 535, 572, 606, 604, 603

Було проведено ряд експериментів з використанням базового алгоритму зі зміною коефіцієнту r від 1 до 3х з кроком 0,2.

Табл.2. Оцінки кластеризації базовим алгоритмом

Rand	**Jaccard**	**FM**	**r**
0,7124752475247524	0,28678678678678678	0,3223576893280606	1,0
0,7318811881188119	0,33052464228934816	0,3748995740211979	

			1,2
0,7592079207920792	0,22769010043041606	0,266662252818048	1,4
0,7730693069306931	0,20328638497652583	0,2634688552355421	1,6
0,7776237623762376	0,18101472995090016	0,2607792715617362	1,8
0,7772277227722773	0,1771123872026251	0,2562992519112617	2,0
0,7724752475247525	0,15432098765432099	0,2129253836069806	2,2
0,7726732673267327	0,154365733113673806	0,2142687747481673	2,4
0,7712871287128713	0,14702970297029703	0,1994568222714615	2,6
0,7712871287128713	0,14702970297029703	0,1994568222714615	2,8

Найкращим виявилося розбиття при r = 1,2.

Табл.3. Розбиття вибірки питань базовим алгоритмом при r = 1,2

Номер кластера	**Ідентифікатори питань**
1	522, 590, 515, 558, 530, 552, 563, 531, 516, 581
2	555, 597, 550, 546, 598, 561, 551, 578, 541, 601, 584, 562, 604, 594, 608, 576, 613, 549, 603, 564
3	523, 534, 610, 537, 579, 525, 600, 571, 524, 596
4	582, 565, 559, 519, 573, 545, 567, 553, 557, 556, 588, 532, 528, 607, 592
5	518, 614, 533, 538, 587, 542, 574, 602, 527, 529, 566, 583, 554, 526, 547, 569, 605, 586, 577, 611, 540, 599, 535, 520, 521, 589, 548, 544, 595, 612, 539, 517, 543, 560
6	570, 591, 609, 575, 606, 572, 585, 536, 593, 580, 568

Далі проведемо ряд експериментів для модифікованогоалгоритму зі значенням r = 1,2, minClusterSize = 15, змінюючипараметр b (від 2 до 10 з кроком 0,5).

Табл.4. Оцінки кластеризації модифікованим алгоритмом

Rand	**Jaccard**	**FM**	**r**	**b**
0,56	0,24957784532252617	0,426569506627921	1,2	2,0
0,5257425742574258	0,20826446280991737	0,36697101792046627	1,2	2,5

0,7087128712871287	0,2604323780794369	0,41360584507340314	1,2	3,0
0,7176237623762376	0,27906976744186046	0,4369231776814638	1,2	3,5
0,6415841584158416	0,34983560356975106	0,5606669393266894	1,2	4,0
0,9001980198019802	0,6536082474226804	0,7905247832556728	1,2	4,5
0,9001980198019802	0,6536082474226804	0,7905247832556728	1,2	5,0
0,740990099009901	0,3130252100840336	0,47716675771109424	1,2	5,5
0,740990099009901	0,3130252100840336	0,47716675771109424	1,2	6,0
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	6,5
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	7,0
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	7,5
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	8,0
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	8,5
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	9,0
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	9,5
0,6875247524752475	0,2211253701875617	0,36232000066321546	1,2	10,0

Найкращим виявилося розбиття prib = 4,5.

Табл.5. Розбиття вибірки питань модифікованим алгоритмом при b = 4,5

Номер кластера	Ідентифікатори питань
1	571, 595, 574, 578, 527, 605, 566, 586, 554, 526, 542, 547, 577, 589, 518, 551, 611, 560, 614, 533, 583, 569, 543
2	524, 581, 519, 558, 573, 516, 515, 530, 563, 592, 613, 550, 598, 546, 590, 567, 575, 570, 609, 591, 594, 584, 562, 564, 601, 604, 603
3	588, 517, 565, 555, 541, 531, 520, 587, 576, 544, 597, 535, 612, 556, 561, 608, 545, 557, 607, 599, 602, 549, 540, 539, 538, 548
4	596, 528, 522, 582, 521, 529, 600, 525, 537, 579, 523, 534, 559, 553, 532, 610, 585, 568, 552, 580, 593, 536, 572, 606

Аналіз отриманих результатів

Із отриманих результатів видно, що найкращий результат кластеризації ми отримали при використанні модифікованого алгоритму CLOPE. При використанні базового алгоритму розбиття вийшло нерівномірним, в найкращому випадку діапазон елементів в кластерах був від 10 до 36, при використанні модифікованого – від 23 до 27. Тобто була досягнута мета позбутися розріджених кластерів та зменшити відхилення розміру кожного кластера від середнього значення. Такий результат досягається при правильно підібраних параметрах (в даному випадку b = 4,5, minClusterSize = 15). Проте гнучкість налаштування алгоритму дозволяє розбивати вибірку на кластери з різною середньою кількістю елементів.

Висновки

Запропоновано модифікований алгоритм кластеризації категорійних даних CLOPE, що дозволяє задавати параметри, які обмежують верхню та нижню границю кількості елементів у кластері. Виконано розбиття наданої вибірки питань на кластери за допомогою базового та модифікованого алгоритму CLOPE з різними значеннями параметрів rта b. Як видно з отриманих оцінок найкращим розбиттям базовим алгоритмом є розбиття zr =1,2 і має результати для індексів Rand = 0,73, Jaccard = 0,33 і FM = 0,37.Найкращим

розбиттям модифікованим алгоритмом є розбиття zr =1,2, b = 4,5 і має результати для індексів Rand = 0,9, Jaccard = 0,65 і FM = 0,79.

Таким чином, з отриманих результатів можна сказати, що використання модифікованого алгоритму CLOPEдобре підходить для задач, коли необхідно враховувати такі характеристики як розмір кластеру. Слід зазначити, що такий підхід достатньо спрощує задачу побудови структури доменної моделі систем адаптивного навчання, тому що набагато зменшує втручання людини, а також дозволяє рівномірно розбити існуючий матеріал на модулі, що за допомогою базового алгоритму CLOPEзробити майже не можливо.

Література:

1. Brusilovsky, Peter (2003). "Adaptive and Intelligent Web-based Educational Systems". International Journal of Artificial Intelligence in Education 13 (2–4): 159–172.
2. Learner Model in Adaptive Learning. Loc Nguyen, Phung Do. World Academy of Science, Engineering and Technology 45 - 2008.
3. Martins, A. C., Faria, L., Vaz de Carvalho, C., & Carrapatoso, E. (2008). User Modeling in Adaptive Hypermedia Educational Systems. Educational Technology & Society, 11(1), 194-207.
4. Система персоналізованого адаптивного навчання на основі онтологій та методів машинного навчання. Дідковська М., Івінський А., Козлов К. - Електроніка та зв’язок – 2016.
5. CLOPE:A Fast and Effective Clustering Algorithm for Transactional Data Yiling Yang, Xudong Guan, Jinyuan You, Dept.of Computer Science & Engineering,Shanghai JiaoTong University, Shanghai, P.R.China.

УДК 004.67

Використання комбінованих оцінокпрогнозу фармацевтичного ринку України у агентій моделі

Б.С. Журке, Д.Г. Діденко

Вступ.Фармацевтичний ринок – це один із потужних промислових секторів, який входить до лідерів найбільш прибуткових галузей світового господарського комплексу. Ця галузь посідає значне місце в економіці України. На даний момент часу, фармацевтичний ринок України переживає скрутні часи. Так, наприклад, обсяг проданих упаковок препаратів всіх категорій скоротився на 16,7 %, а в доларовому вираженні загальний обсяг продажів «аптечного кошика» зменшився на 40,2% [1]. З огляду на такі тенденції, необхідно створити ефективну політку управління та планування. Для розробки ефективних рішень застосовують прогнозування.

Однією з популярних сучасних тенденцій в області моделей прогнозування є створення комбінованих моделей і методів. Подібний підхід дає можливість компенсувати недоліки одних моделей за допомогою інших і спрямований на підвищення точності прогнозування, як одного з головних критеріїв ефективності моделі. Саме використання комбінованого підходу до прогнозування в агентній моделі буде розглянуто в даній статті.

Постановка задачі.Метою роботи є розробка та дослідження ефективності використання комбінованого методу прогнозування фармацевтичного ринку України в агентній моделі.

Для досягнення зазначеної мети поставлені наступні завдання:

1. Побудуватипрогноз за допомогою інтегрованої моделі авторегресії ковзного середнього.

2. Побудувати прогноз за допомогою багатошарової нейронної мережі з оберненим поширенням похибки.
3. Побудувати прогноз за допомогою комбінованого методу прогнозування реалізованого в агентній моделі.
4. Дослідити якість прогнозу отриманого за допомогою агентної моделі.
5. Аналіз отриманих результатів.

Викладення теорії. Агентне моделювання – метод імітаційного моделювання, який досліджує поведінку децентралізованих агентів і те, як така поведінка визначає поведінку всієї системи в цілому. Основним поняттям агентної моделі – це агент. Виділяють два основних типів агентів[2], інтелектуальні та реактивні. Інтелектуальні агенти володіють добре розвиненою моделлю зовнішнього світу, що досягається завдяки наявності у них бази знань, механізмів вирішення і аналізу дій. У повноцінного інтелектуального агента обов'язково повинні бути присутніми як мінімум чотири з перерахованих функції: когнітивна, регулятивна, комунікативна і ресурсна. У той же час реактивні агенти не мають ні скільки-небудь розвиненого уявлення зовнішнього середовища, ні механізму багатокрокових міркувань, ні достатньої кількості власних ресурсів. Звідси випливає ще одна істотна відмінність між інтелектуальними і реактивними агентами, пов'язана з можливостями прогнозування змін зовнішнього середовища і, як наслідок, свого майбутнього. Саме за допомогою реактивних агентів буде побудована агентна модель для прогнозування фармацевтичного ринку України. Серед них можна виділити два типи агентів, агент-менеджер, агент-виконавець. Агент-менеджер буде роздавати завдання іншим агентам і опрацьовувати отримані результати іншими агентами, та виводити результати. Результат представляє собою комбінований прогноз.

Застосування комбінованих моделей є напрямом, який при коректному підході дозволяє підвищити точність прогнозування. Головним недоліком комбінованих моделей є складність таресурсоємність їх розробки: потрібно розробити моделі таким чином, щоб компенсувати недоліки кожної з них, не втративши їхніх переваг. Спосіб комбінування окремих прогнозів, як правило, полягає в тому, щоб представити комбінований прогноз у вигляді зваженої суми окремих прогнозів:

$$\hat{\mathbf{y}} = \sum_{\mathbf{i=1}}^{\mathbf{M}} w\widehat{\mathbf{y}_{\mathbf{it}}}, \qquad (1)$$

де $\widehat{y_{it}}$ - окремий прогноз, отриманий для моменту часу t; M-кількість комбінованих прогнозів; w_i - вагові коефіцієнти окремих прогнозів. 0 <w<1. Сума всіх вагових коефіцієнтів повинна давати одиницю $\sum_{i=1}^{M} w_i = 1$, окремі ваги повинні знаходитися в інтервалі [0, 1].

Похибка двох прогнозів які ми будемо використовувати для побудови комбінованого прогнозу розраховується наступним чином:

$$\text{var}\left[\frac{e_1(k)+e_2(k)}{2}\right] = E\left[\frac{e_1(k)+e_2(k)}{2}\right]^2 = \frac{1}{4}E\left[e_1^2(k)+2e_1(k)e_2(k)+e_2^2(k)\right] =$$

$$= \frac{1}{4}\left\{E\left[e_1^2(k)\right]+2E\left[e_1(k)e_2(k)\right]+E\left[e_2^2(k)\right]\right\} = \qquad (2)$$

$$= \frac{1}{4}\left[\sigma_1^2 + 2\frac{E[e_1(k)e_2(k)]}{\sigma_1\sigma_2}\sigma_1\sigma_2 + \sigma_2^2\right] = \frac{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}{4}.$$

Дисперсія комбінованого прогнозу:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4}, \qquad (3)$$

де, ρ – коефіцієнт кореляції між похибками прогнозу. Враховуючи, що похибки за двома моделями незалежні, тобто $\rho = 0$, то формула (3), матиме вигляд:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2}{4}. \qquad (4)$$

Вагові коефіцієнти можна визначити за допомогою похибок прогнозів. Зазвичай похибки прогнозів для конкретних моделей і процесів відомі, або їх можна визначити.Оскільки для моделей, у яких суми квадратів похибок прогнозів не великі, генерують більш якісніші прогнози. Тому логічно прийняти цю міру за основу для визначення вагових коефіцієнтів. Позначимо суму квадратів похибок прогнозування (для історичного прогнозу) через

$$sse = \sum_{k=1}^{N} e^2(k). \qquad (5)$$

Тепер можна записати вирази для вагових коефіцієнтів окремих прогнозів [3]:

$$w_1 = \frac{1/sse_1}{1/sse_1 + 1/sse_2}, \qquad (6)$$

$$w_2 = \frac{1/sse_2}{1/sse_1 + 1/sse_2}, \qquad (7)$$

де sse_1, sse_2 – суми квадратів похибок для кожного з методів, що використовуються в даному випадку [3]. Критеріями якості прогнозу виступають наступні величини [4] середньоквадратична похибка, корінь з середньоквадратичної похибки, середня абсолютна похибка, середня абсолютна похибка у відсотках. Формули перелічених критеріїв наведені нижче:

$$MSE = \frac{\sum_{i=n-m+1}^{n} (y_i - \hat{y}_i)^2}{m}, \qquad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=n-m+1}^{n} (y_i - \hat{y}_i)^2}{m}}, \qquad (9)$$

$$MAE = \frac{\sum_{i=n-m+1}^{n} |y_i - \hat{y}_i|}{m}, \qquad (10)$$

$$MAPE = \sum_{i=n-m+1}^{n} \frac{100|y_i + \hat{y}_i|}{m|y_i|}. \qquad (11)$$

Приклад застосування методики. Початкові даніобсягу продажу біологічно активних добавок за місяць та прогноз, отриманий за допомогою моделі ARIMAнаведені в таблиці 1.

Табл. 1. Обсяг продажу біологічно активних добавок (01.2012-03.2015 рр.)

Період	Сума (млн.)	Період	Сума (млн.)	Період	Сума (млн.)	Період	Сума (млн.)
2012/M01	112,9551	2013/M01	149,4604	2014/M01	157,4691	2015/M01	203,595
2012/M02	123,3385	2013/M02	156,7737	2014/M02	174,2648	2015/M02	205,1441
2012/M03	144,5633	2013/M03	166,1693	2014/M03	194,6393	2015/M03	200,2705
2012/M04	124,0609	2013/M04	164,9727	2014/M04	226,7557		
2012/M05	129,1109	2013/M05	151,4843	2014/M05	200,9408		
2012/M06	119,054	2013/M06	138,963	2014/M06	190,431		
2012/M07	124,5956	2013/M07	152,2765	2014/M07	193,8262		
2012/M08	126,6937	2013/M08	138,879	2014/M08	182,0701		
2012/M09	125,4146	2013/M09	138,2881	2014/M09	193,062		
2012/M10	141,3878	2013/M10	149,3924	2014/M10	210,1404		
2012/M11	145,5003	2013/M11	157,8316	2014/M11	203,2838		
2012/M12	156,9459	2013/M12	179,1458	2014/M12	224,8522		

В якості навчальної вибірки будемо використовувати дані з січня 2004 р. по грудень 2014 рр., прогноз виконується на січень, лютий, березень 2015 р. Побудова прогнозу обсягу продажу за допомогою моделі ARIMA. Для побудови прогнозу завантажуємо дані за попередні періоди [5]:

```
> data=scan("C:\\1.txt")
Read 132 items
> datats<-ts(data,frequency=12,start=c(2004,1))
```

Далі для побудови моделі на основі наших даних виконуємо наступні дії:

1. Завантажуємо бібліотеку для побудови моделі та прогнозу.

2. Викликаємо функцію для автоматичного підбору коефіцієнтів для нашої моделі.
3. Будуємо прогноз на 3 кроки:

```
> library("forecast")
> model<-auto.arima(datats)
> res<-predict(model,n.ahead=3)
```

Результати якості прогнозу наведені в таблиці 11, а результат прогнозування на один, три і п’ять кроків наведені в таблицях 2,3 та 4.

Табл. 2. Результати прогнозування на один крок

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники
2015/03	213,7650788	218,532782

Табл. 3. Результат прогнозування на три кроки

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники
2015/01	213,7631156	199,393855
2015/02	232,3278308	204,971447
2015/03	213,7650788	218,532782

Табл. 4. Результати прогнозування на п’ять кроків

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники
2014/11	213,4009933	203,283756
2014/12	230,9603885	224,852172
2015/01	212,3344333	199,393855
2015/02	226,8405618	204,971447
2015/03	245,1552515	218,532782

Далі побудуємо прогноз за допомогою нейронної мережі з оберненим поширенням похибки. Для цього була написана програма мовою програмування Java. Нейрона мережа має наступну структуру: 4 вхідних нейрона, 9 нейронів в прихованому шарі, 1 вихідний нейрон. Максимальна кількість ітерацій – 1000. Активаційна функція сигмоїдальна. В якості навчальної вибірки виберемо показники обсягу продаж біологічно активних добавок з січня 2004 р. по

грудень 2010 р. Перевірочна вибірка з січня 2011р по грудень 2014 р. У таблиці 11 наведені показники якості прогнозу. У таблицях 5, 6, 7 наведені результати прогнозування на один, три та п’ять кроків.

Табл. 5. Результати прогнозування обсягів продаж на один крок

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники (млн.)
2015/03	212,5639871	218,532782

Табл. 6. Результат прогнозування обсягів продаж на три кроки

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники (млн.)
2015/01	203,5950191	199,393855
2015/02	205,144155	204,971447
2015/03	200,2705886	218,532782

Табл. 7. Результат прогнозування обсягів продаж на п’ять кроків

Період	Прогнозоване значення обсягу продажу (млн.)	Реальні показники(млн.)
2014/11	209,9662955	203,283756
2014/12	208,265637	224,852172
2015/01	210,8092987	199,393855
2015/02	212,2963666	204,971447
2015/03	212,5639871	218,532782

Тепер побудуємо прогноз використовуючи агентну модель для прогнозування.В агентній моделі реалізований комбінований підхід до побудови прогнозу. Один з агентів це агент – менеджер який роздає завдання іншим агентам – підрядникам. Які виконують функцію побудови прогнозів на основі статистичної моделі та за допомогою нейронних мереж. Отримавши завдання агенти підрядники будують прогноз на кількість кроків, яку задає користувач. Коли агенти отримали результати прогнозу, вони надсилають прогнозовані дані агенту менеджеру, який в свою чергу будує комбінований

прогноз. Було проведено десять експериментів за допомогою агентної моделі. Результат побудови прогнозу, а також оцінки прогнозу агентної моделі наведені в таблиці 8, 9, 10.

Табл. 8. Оцінки прогнозів та результати прогнозу на один крок

Період	**Прогнозоване значення обсягу продажу (млн.)**	**Реальні показники**
2015/03	211,0095985	218,532782
	RMSE	0,236243634
	MAPE	0,231297307

Табл. 9. Оцінки прогнозів та результати прогнозу на три кроки

Період	**Прогнозоване значення обсягу продажу (млн.)**	**Реальні показники**
2015/01	204,3957522	199,393855
2015/02	205,1452384	204,971447
2015/03	212,904018	218,532782
	RMSE	18,91072107
	MAPE	1,723015455

Табл. 10. Оцінки прогнозів та результати прогнозу на п’ять кроків

Період	**Прогнозоване значення обсягу продажу (млн.)**	**Реальні показники**
2014/11	211,0095985	203,283756
2014/12	228,2501304	224,852172
2015/01	211,4767375	199,393855
2015/02	213,7634511	204,971447
2015/03	214,1238212	204,971447
	RMSE	75,65922092
	MAPE	4,025219895

Аналіз отриманих результатів. В таблиці 11 наведені оцінки прогнозів для кожної моделі.

Табл. 11. Оцінки моделей в залежності від кількості кроків

Кількість	**MAPE Модель**	**MAPE Нейронна мережа**	**MAPE Агентна модель**

кроків	ARIMA (0,1,1)	backpropagation	
1	2,2	1,7	0,23
3	7,57	3,51	1,72
5	7,4	4,54	3,54

Із отриманих результатів видно, що найкращий результат прогнозованих значень досягається при використані агентної моделі, яка використовує комбінований підхід для побудови прогнозу. Оскільки при такому підході похибка прогнозу незалежно від кількості завжди менша ніж при використанні інших методів для прогнозування.

Висновки

Виконано побудову прогнозу за допомогою інтегрованої моделі авторегресії ковзного середнього, багатошарової нейронної мережі з оберненим поширенням похибки та агентної моделі. Як видно з отриманих результатів, середня абсолютна похибка при використанні інтегрованої моделі авто регресії ковзного середнього, в залежності від кількості кроків складала MAPE=2,2%, MAPE=7,57%, MAPE=7,4%. При використанні нейронної мережі з оберненим поширенням в залежності від кількості кроків абсолютна похибка складала MAPE=1,7, MAPE=3,51, MAPE=4,54. Це свідчить про те що побудовані моделі є адекватними та дають високоякісний прогноз, оскільки похибки не перевищують значення 10%. Але застосувавши ті ж самі моделі для побудови комбінованого прогнозу в агентній моделі дало набагато кращі результати, а саме абсолютна похибка в залежності від кроків складала MAPE=0,23, MAPE=1,72, MAPE=3,54. Такі результати свідчать про те що побудована агентна модель є адекватною і надає високу точність при побудові прогнозу.

Таким чином, з отриманих результатів можна сказати, що використання комбінованого методу підвищує якість прогнозу в агентній моделі. Необхідно також зазначити, що використання агентної моделі спрощує задачу побудови комбінованого прогнозу, оскільки вказаний підхід, надає системі більшу

гнучкість, зменшує навантаження на систему в якій проводяться експерименти, також дозволяє прорахувати стан моделі через певний інтервал часу та підвищити якість прогнозу додавши складові які раніше не враховувались. Шляхом додавання нових агентів в модель.

Література

1. Аптечный рынок Украины по итогам I полугодия 2015 г.: HelicopterView. [Електронний ресурс]. – Режим доступу: http://www.apteka.ua/article/339254.
2. Классификация агентов. [Електронний ресурс]. – Режим доступу: http://www.aiportal.ru/articles/multiagent-systems/agent-classification.html.
3. *Бідюк П.І., Романенко В.Д., Тимощук О.Л.* Аналіз часових рядів. – Київ:Політехніка, 2013. – 607 с.
4. *Присенко Г.В., Равікович Є.І.* Прогнозування соціально-економічних процесів – 2005р. – 382 с.
5. *Мастицкий С.Е., Шитиков В.К.* Статистический анализ и визуализация данных с помощью R. – 2014р. – 401 с.

УДК 007:681.3.06

Передбачення сонячної активності на основі аналізу радіовипромінювання Сонця

В.С. Варава, В.М. Подладчіков

Вступ

Сонячна активність - сукупність явищ, які періодично виникають в сонячній атмосфері. Прояв сонячної активності тісно пов'язаний з магнітними властивостями сонячної плазми. Виникнення активної області починається з поступового збільшення магнітного потоку в деякій області фотосфери. У відповідних місцях хромосфери після цього спостерігається збільшення яскравості.

Прогнозування особливостей сонячного циклу допоможе передбачати появу сонячних плям і виникнення сонячних бурь, а також мінімізувати проблеми, що виникають при їх впливі на нашу планету.

Оглядове радіоспостереження Сонця почалося в Канаді в листопаді 1946 року, коли Ковінгтон використовував 4-футовий рефлектор від радіолокаційної станції гарматного наведення Type IIIC, щоб почати запис сонячного потоку на довжині хвилі 3 см.

У 1947 році вимірювання Ковінгтона переросли в регулярну програму спостережень на довжині хвилі 10,7 см. Щоденні вимірювання потоку відбувалися в Алгонкінсьій радіо обсерваторії, Оттава, до червня 1991 року, потім програма була перенесена в Домініонську астрофізичну обсерваторію, Пентіктон, що на три часових пояси західніше, де вона триває і сьогодні.

Протягом 1947 року, оглядові вимірювання почалися і в Сіднеї, але їх не продовжували після 1950-х років. В кінці 1951 року регулярні спостереження почалися в Японії на 3750 МГц (8,0 см), а в 1956 р вони були доповнені спостереженнями на 1000, 2000 і 9400

МГц (30; 15 і 3,2 см, відповідно). Ці дані продовжують безперервно записуватися до наших днів. До 1958 року кілька інших обсерваторій почали оглядові програми, і їх кількість збільшувалася в багато разів протягом 1960-х років.

Сонячна радіо обсерваторія Сагамор Хілл почала працювати на частотах 8800, 4995, 2695, 1415 і 609 МГц в 1966 році, до яких додали 15400 МГц в 1967 році, 242 МГц на початку 1969, і 410 МГц на початку 1971 року.

У середині 1960-х років стало очевидним, що існують серйозні розбіжності в вимірах потоку, зроблених в різних обсерваторіях, в зв'язку з цим була зібрана комісія URSI для досягнення рівномірного калібрування. Робоча група, на чолі з Х. Танака, погодилася, що пірамідальна рупорна антена в поєднанні з фоновими вимірами забезпечать точне і стабільне калібрування. Спостерігачі в Тойокава прийняли цю техніку в 1965 році, і в більш ранніх вимірах були виправлені похибки близько 5%. Згодом, більшість обсерваторій мали розбіжності менш ніж на ± 1%.

Індекс F10,7 - потік сонячного радіовипромінювання на хвилі довжиною 10,7 см (частота 2800 МГц), виражений в сонячних одиницях потоку (с.о.п.), причому $1 с.о.п. = 10^{-22} \frac{Вт}{м^2 Гц}$.

Було встановлено, що індекс F10,7 добре корелює з багатьма сонячними показниками і властивостями. Серед них:

Кількість сонячних плям: використовуючи вимірювання за 40 років, Теппінг отримав наступне співвідношення між 10,7 см потоком і кількістю сонячних плям: $SN = 1.147 F_{10,7} - 73,21$ [1].

Площа сонячних плям: Денисе (1948) виявив 76% кореляцію між 10,7 см потоком і площею сонячних плям [2].

Іонізація Е-шару: Денисе і Кунду (1957) виявили хорошу кореляцію між індексом іонізації Е-шару Землі і 10,7 см потоком [2].

Виявлена залежність між 10,7 см потоком та даними, отриманими за допомогою двох бортових магнітографів: MDI на SOHO (Solar & Heliospheric Observatory), яка обертається навколо точки Лагранжа L1 (тут сила земного тяжіння рівна силі сонячного тяжіння) на відстані приблизно 1,5 млн. км від Землі, та HMI на SDO (Solar Dynamics Observatory), яка знаходиться на геосинхронній орбіті Землі на висоті 36000 км[3].

Досліджувалась залежність між сонячними випромінюваннями на п’яти різних довжинах хвилі (30; 15; 10,7; 8 та 3,2 см). Виявилося, що найкращий компроміс між складністю моделі та кореляцією виходить при комбінації двох довжин хвиль. Хороші результати отримані при використанні 10,7 см з 30 см потоків [4].

Побудовані алгоритми передбачення сонячної активності (індекс F10,7) за зображеннями Сонця [5].

Але всі ці результати отримані для величезних вибірок, довжиною не менше одного сонячного циклу (11 років). Для таких вибірок часто використовують не денні значення активності сонця, а середньомісячні. В таких масштабах дійсно проглядається гарна залежність між різними сонячними показниками. Але при прогнозуванні сонячної активності на 1 день, ці залежності несуттєві, і часто внесення їх до моделі погіршують результати.

Постановка задачі: розробити ефективну методику прогнозування індексу F10,7 (потік випромінювання Сонця на довжині хвилі 10,7 см) на 1, 2 та 3 дні.

Тому в даній статті запропоновано новий підхід до прогнозування сонячної активності.

Фільтр Калмана є дуже потужною технікою, що включає фільтрацію і екстраполяцію для оцінки поточних і майбутніх станів, досить точний в умовах неповної інформації про досліджувані процеси. Для застосування фільтра Калмана необхідно представити модель досліджуваного процесу в просторі станів.

Опис моделі

Для прогнозування індексу F10,7 формуємо модель системи, що описує стан і вимір потоку радіовипромінювання в просторі стану. Рівняння стану, що описує динаміку потоку радіовипромінювання

$$f_{k+1} = f_k + w_{k+1} + q_{k+1}. \quad (1)$$

Тут f_k – значення потоку радіовипромінювання в k-й день, w_{k+1} – шум моделі, що описує похибки моделі і непередбачуваність динаміки в майбутньому, q_{k+1} – систематична складова зміни потоку радіовипромінювання за один день.

Припускається, що доступні спостереження потоку радіовипромінювання в присутності шуму.

Тоді рівняння вимірювання потоку радіовипромінювання має вигляд

$$z_k = f_k + v_k. \quad (2)$$

де z_k – спостережуване значення потоку радіовипромінювання в k-й день, v_k – послідовність вимірювальних шумів.

Оптимальний для моделі (1,2) прогноз потоку радіовипромінювання може бути отриманий на основі фільтра Калмана.Для того, щоб використовувати рівняння фільтра Калмана ми вважаємо w_{k+1} та v_k некорельованими та незміщенимиздисперсіями Q та R відповідно.

Фільтр Калмана

Фільтр Калмана представлений рекурентним чином і його алгоритм включає в себе фільтрацію і прогнозування для досить точної оцінки поточного і майбутнього станів процесу (2).

Для того, щоб почати процедуру фільтра Калмана, ми визначаємо початкові оцінки потоку радіовипромінювання f_0^f та дисперсії похибки цієї оцінки P_0^f.

Алгоритм включає в себе послідовне виконання двох операцій - передбачення і фільтрації.

Передбачення на один крок

$$f_{k+1}^p = f_k^f + q_{k+1}, \qquad (k = 1,2,...,n). \qquad (3)$$

Тут f_{k+1}^p – екстрапольоване на $k+1$ день значення потоку радіовипромінювання, f_k^f – фільтроване значення потоку радіовипромінювання, отримане за даними спостереження на інтервалі часу $(1,2,...,k)$.

Дисперсія похибки передбачення P_{k+1}^p визначається наступним чином

$$P_{k+1}^p = P_k^f + Q. \qquad (4)$$

Фільтрація

На цьому кроці виконується корекція передбаченої на попередньому $k-1$ кроці оцінки f_k^p по даним спостережуваного на k-му кроці потоку радіовипромінювання z_k

$$f_k^f = f_k^p + K_k(z_k - f_k^p). \qquad (5)$$

Тут K_k – коефіцієнт посилення фільтра, який визначається наступним чином

$$K_k = \frac{P_k^p}{P_k^p + R}. \qquad (6)$$

Після фільтрації, потоки які спостерігаються z_k перетворюються в нові відфільтровані значення f_k^f для яких точність оцінки характеризується дисперсією похибки фільтрації, яка визначається наступним чином

$$P_k^f = (1 - K_k) P_k^p. \qquad (7)$$

Фільтрація Калмана забезпечує оптимальну оцінку для моделі (1,2), проте потрібна апріорна інформація про параметри моделі Q та R, а також q_k – змінне у часі зміщення потоку радіовипромінювання за один день.

Тому для практичної реалізації алгоритму необхідна його попереднє налаштування.

Адаптивне налаштування фільтра Калмана за даними спостережень

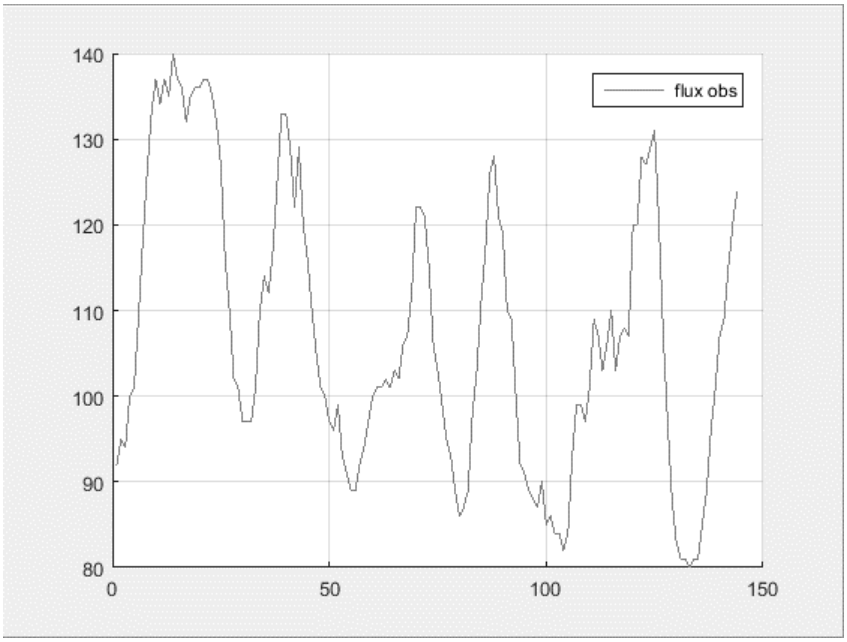


Рис. 1.Зміна потоку радіовипромінювання з червня 2015 по жовтень 2015: по осі абсцис кількість днів; по осі ординат – значення індексу F10,7

На рис.1 показані вимірювання потоку радіовипромінювання з червня 2015 по жовтень 2015. З рис. 1 видно, що динаміка потоку радіовипромінювання має циклічний характер. Період коливань приблизно дорівнює 26 дням.

Для прогнозування потоків відповідно до виразу (3) необхідно оцінити прогнозоване значення q_{k+1} – систематична складова зміни потоку радіовипромінювання за один день.

Зміщення потоку радіовипромінювання за один день $q_k = z_k - z_{k-1}$ змінюється від кроку до кроку, демонструючи, проте, певну залежність від попереднього зміщення q_{k-1}. Виходячи з даних спостережень потоку до k-го дня, припускаємо, що наступна зміна потоку q_{k+1} залежить від попередніх змін потоку радіовипромінювання, причому в більшій мірі залежить від останнього q_k.

Тому, для оцінки q_{k+1} використовуваного для прогнозу потоку радіовипромінювання на $k+1$ день, пропонується застосувати експоненційне згладжування часового ряду q_k $(k = 2,3,...,n)$. При обчисленні експоненційного середнього всі попередні значення зміщення $q_{k-1}, q_{k-2},..., q_1$ враховуються з вагами, що зменшуються по експоненційному закону.

Нехай $\hat{q}_{k+1}$ – прогнозоване значення зміни потоку на $k+1$ крок, отримане в результаті експоненційного згладжування, яке в своїй найпростішій формі задається

$\hat{q}_{k+1} = \hat{q}_k + \alpha(q_k - \hat{q}_k)$.

Тут α – експоненційна вага $(0 < \alpha < 1)$.

Експоненційне згладжування можна розглядати як фільтр, в якому вхідні дані представляють собою зміщення потоку радіовипромінювання за один день $q_k = z_k - z_{k-1}$, а вихідні оцінки є

прогнозованими значеннями $\hat{q}_{k+1}$ зміщення потоку радіовипромінювання на наступний день.

Коли α маленьке, реакція на зміну динаміки є повільною. Зі збільшенням α зміни динаміки захоплюються більш ефективно, але прогнозоване значення зміщення на наступний день наближується до зміщення потоку радіовипромінювання за поточний день. Тому при виборі α потрібно знайти компроміс між цими двома характеристиками. Такі компромісні значення $\alpha = 0.82$, а також параметрів моделі $Q = 4$ та $R = 3$ вибрані на основі даних спостереження потоків радіовипромінювання.

Найбільші проблеми при прогнозуванні виникають в точках, які відповідають максимальним і мінімальним пікам потоку радіовипромінювання, їх видно на рис. 1. У цих точках зміщення q_k змінює знак на протилежний і використання попередніх значень зміщення для прогнозу спотворює результати прогнозу. При виявленні різких стрибків динаміки потоків доцільно мінімізувати інерційність моделі. Тому, при виявленні уповільнення зростання або падіння потоку, тобто при виконанні умови $\frac{q_k}{\hat{q}_k} < 1$ в якості прогнозованого зміщення використовується величина $\hat{q}_{k+1} = \frac{q_k}{2}$.

Аналіз результатів моделювання

Розглянутий алгоритм застосований для прогнозування 10,7 см потоку радіовипромінювання з червня 2015 по жовтень 2015.

Для того, щоб почати процедуру фільтра Калмана, ми задалися початковими оцінками потоку радіовипромінювання $f_0^f = z_0$ та дисперсії похибки цієї оцінки $P_0^f = 100$.

Прогноз на 1 день виконувався відповідно до виразу (3). При прогнозуванні потоків на 2 і 3 дня передбачалося, що зміна потоків протягом інтервалу прогнозування залишається постійною. Тому прогноз на 2 дня визначався таким чином

$$f_{k+2}^p = f_k^f + 2q_{k+1}, \qquad (k = 1,2,...,n).$$

А на 3 дні у відповідності з наступним виразом

$$f_{k+3}^p = f_k^f + 3q_{k+1}, \qquad (k = 1,2,...,n).$$

У табл. 1 наведені середньоквадратичні похибки прогнозу на розглянутому інтервалі.

Таблиця 1. Результати прогнозу

Інтервал прогнозування	1 день	2 дні	3 дні
СКП	4,43	7,79	9,64

Традиційний алгоритм прогнозування потоків на 1 день, заснований на припущенні, що приріст потоку на наступний день збігається з поточним приростом, можна розглядати як окремий випадок розглянутого підходу при $R = 0$ та $\alpha = 1$.

Середньоквадратичні похибки прогнозу на 1 день, заснованого на традиційному підході, для розглянутої вибірки спостережень потоків, рівні 5,14, що на 16% перевищують середньоквадратичні похибки прогнозу, наведені в таблиці.

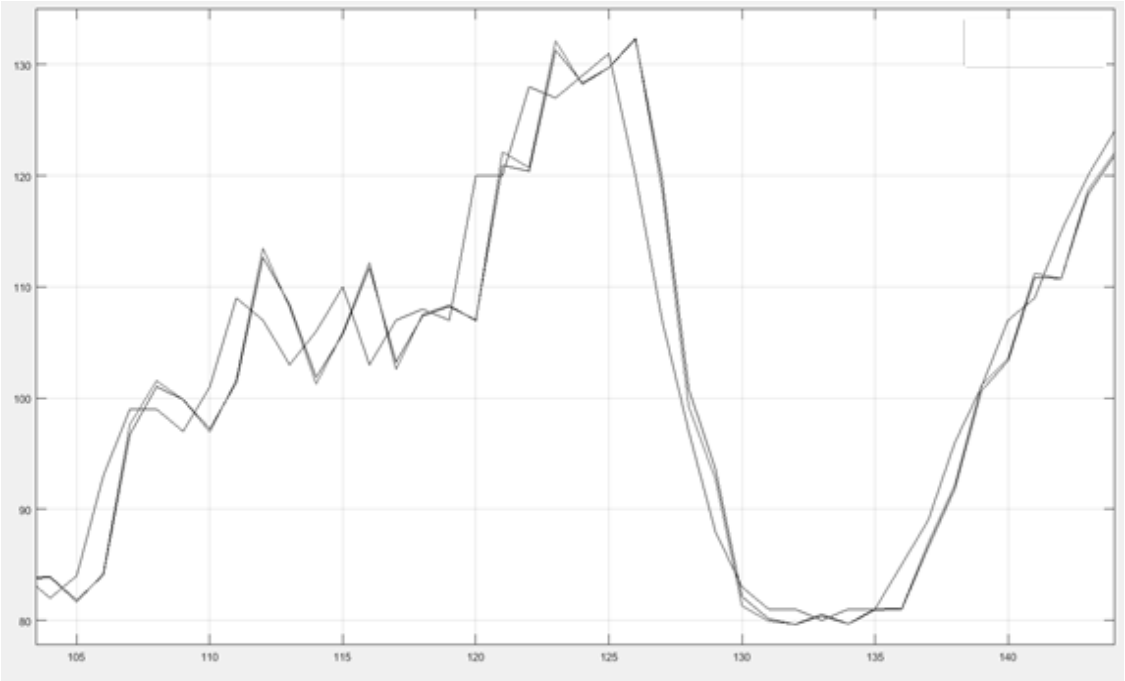


Рис. 2.Результатипрогнозування індексу F10,7з вересня 2015 по жовтень 2015: по осі абсцис кількість днів; по осі ординат – значення індексу F10,7

На рис. 2 приведені спостережувані значення потоків на інтервалі часу з вересня 2015 до жовтня 2015 року, прогнозовані значення потоків на 1 день, отримані на основі запропонованого алгоритму та прогнозовані значення потоків на 1 день, отримані з використанням традиційного підходу.

Як видно з малюнка, згладжування спостережень в запропонованому алгоритмі прогнозу дозволяє зменшити величину стрибкоподібних відхилень прогнозованих потоків від спостережуваних.

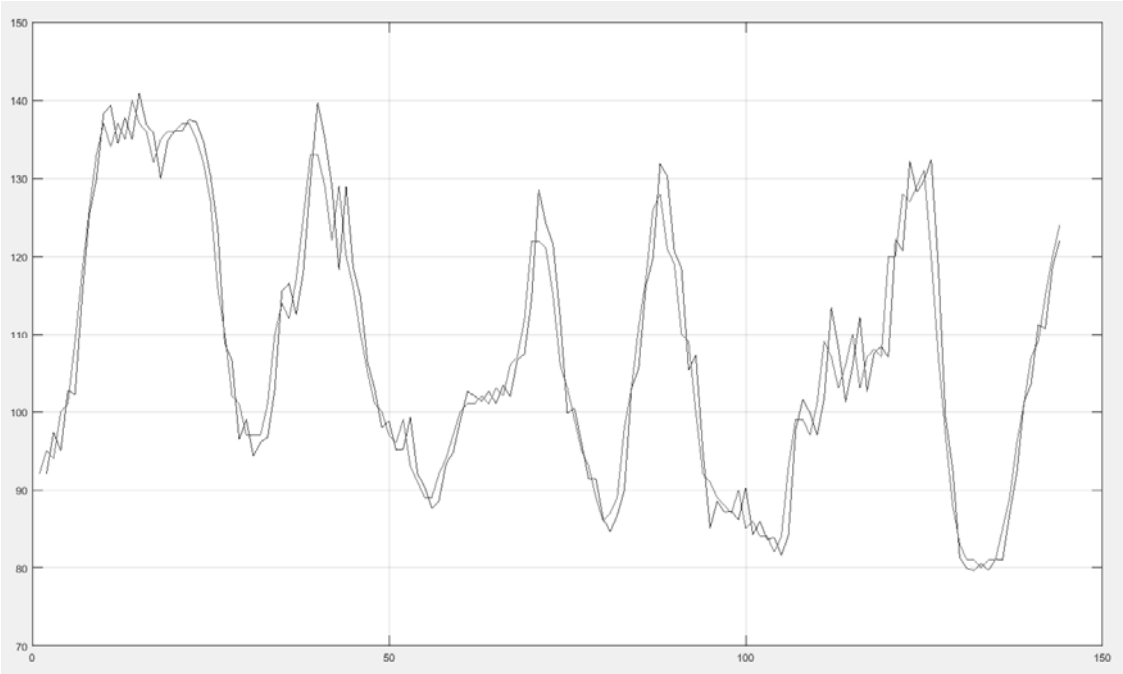


Рис. 3.Прогнозіндексу F10,7 на 1 крок: по осі абсцис кількість днів; по осі ординат – значення індексу F10,7

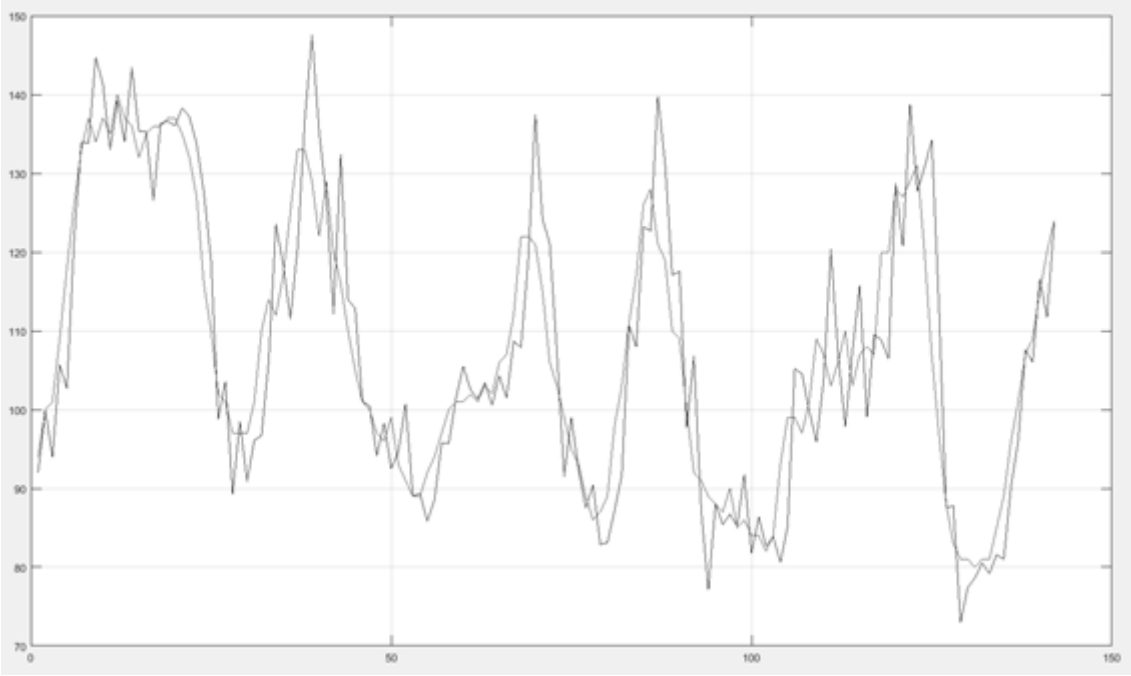


Рис. 4.Прогноз індексу F10,7 на 2 кроки: по осі абсцис кількість днів; по осі ординат – значення індексу F10,7

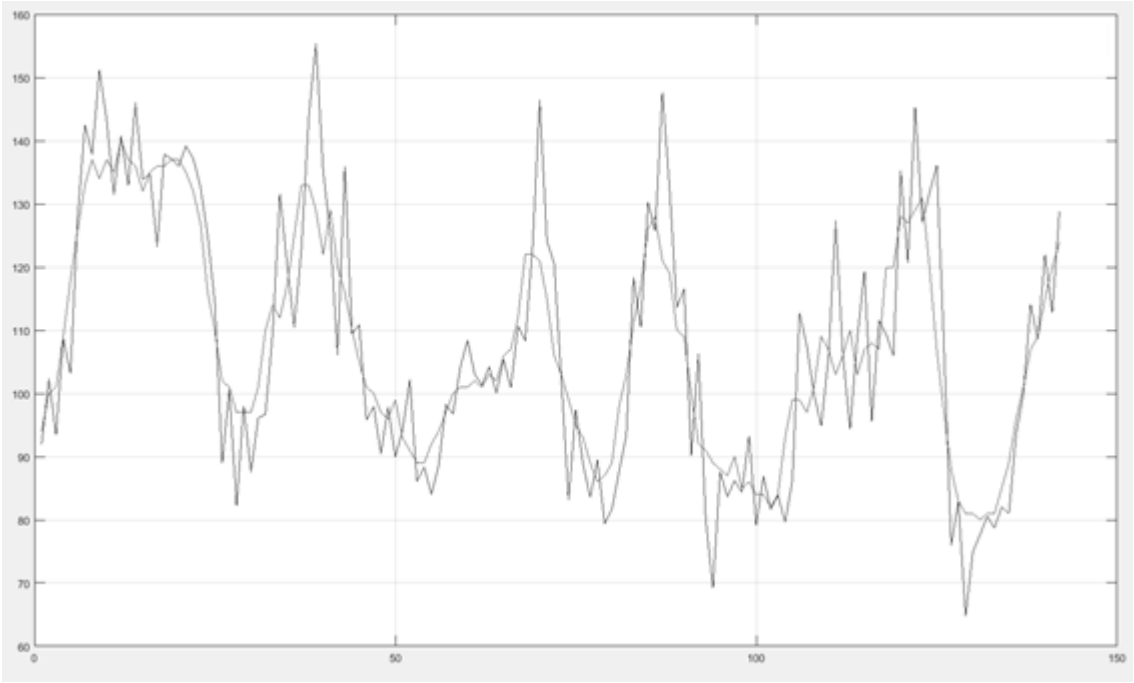


Рис. 5.Прогноз індексу F10,7на 3 кроки: по осі абсцис кількість днів; по осі ординат – значення індексу F10,7

На рис. 3-5 показані спостережувані і прогнозовані значення потоків на 1 день (рис. 3), на 2 дні (рис. 4) і на 3 дні (рис. 5).

З рис. 3-5 видно, що в точках зміни напрямку розвитку процесу, похибки прогнозу різко зростають, досягаючи 8 і навіть 10 одиниць в порівнянні з інтервалами монотонної зміни потоку. Середньоквадратичні похибки прогнозу на 2 або 3 дні збільшуються також в основному за рахунок різкого зростання помилок прогнозу в точках зміни напрямку.

Висновки

Таким чином, в статті запропоновано методику для прогнозування сонячної активності на 1, 2 та 3 дні. Представлено модель досліджуваного процесу в просторі станів та адаптовано до неї фільтр Калмана. Порівняння цього підходу із традиційним алгоритмом показали, що ця методика є на 16% ефективнішою.

В рамках виключно статистичних методів неможливо ефективно прогнозувати кут зміни напряму розвитку процесу. Тому поліпшення прогнозу можна досягти шляхом додаткового прогнозування зміни напрямку потоку з використанням прогнозу корельованих з потоком подій.

Литература

1. *Tapping K. F.* Recent solar radio astronomy at centimeter wavelengths: The temporal variability of the 10.7-cm flux //Journal of Geophysical Research: Atmospheres. – 1987. – p. 829-838.
2. *Schmahl E. J., Kundu M. R.* Synoptic radio observations //Synoptic Solar Physics. – 1998.– p. 387-400.
3. *Svalgaard L., Sun X.*F10.7 Microwave Flux Matches the Total Disk Unsigned Magnetic Flux from MDI and HMI// *Technical Report*, Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, 2016. – http://hmi.stanford.edu/hminuggets/?p=1510
4. *Dudok deWit T., Bruinsma S., Shibasaki K.* Synoptic radio observations as proxies for upper atmosphere modelling //Journal of Space Weather and Space Climate. – 2014.– p.13-26.
5. *Schonfeld S. J. et al.* Coronal Sources of the Solar F10. 7 Radio Flux //The Astrophysical Journal. – 2015.– p. 29-39.
6. *Brown R.G.* Smoothing forecasting and prediction if discrete time series.–New York: Courier Corporation, 1963. – 468p.
7. *Мирошниченко Л.И.* Солнечная активность и Земля. –М.:Наука, 1981. – 145 с.
8. *Чуев Ю. В., Михайлов Ю. Б., Кузьмин В. И.* Прогнозирование количественных характеристик процессов. – М.: Сов. радио,1975. – 205 с.
9. *Лукашин Ю.П.* Адаптивные методы краткосрочного прогнозирования временных рядов.–М.: Финансы и статистика, 2003. – 413 с.
10. *Брайсон А., Хо Юши*. Прикладная теория оптимального управления. – М.: Мир, 1972. – 544 с.

УДК 519.246.8

Гур'янов О.І.

Прогнозування результатів експертного оцінювання рівня знань іноземної мови за допомогою сум випадкових величин

Науковий керівник: к.ф-м.н., доцентКаніовська І.Ю.

Вступ

Побудова ефективних методів прогнозування результатів оцінювання знань студентів є надзвичайно важливою та актуальною задачею. Наприклад, таке прогнозування дозволяє студенту оцінити свою готовність до зовнішнього іспиту, викладачеві виявити проблемні місця та скоригувати навчальний процес, навчальному закладу оцінити рівень підготовки студентів та ефективність викладання, кількість студентів, яких буде відраховано, спрогнозувати суму, необхідну для виплат академічних стипендій.

Методи, що використовуються для прогнозування академічної успішності студентів з певної навчальної дисципліни та їхніх оцінок за екзамен, є досить різноманітними та залежать від постановки задачі і наявних даних про студентів. Лі, Ува та Амін [1] застосовують для прогнозування успішності студентів у курсі елементів статистики множинну регресію з такими факторами, як результати тестів GPA, SAT та ACT, шкільні оцінки з алгебри тощо. Ці фактори є зовнішніми по відношенню до курсу статистики. Данко, Дюк та Франц [2], вивчаючи успішність студентів у курсі бухгалтерії, також визначали фактори, які є найбільш пов'язаними з оцінкою за перший семестр курсу, використовуючи регресійний та дискримінантний аналіз. Фанг і Лю[3] застосовують для прогнозування оцінки на заключному екзамені з інженерного курсу дерева рішень.Хуанг [4] використовував для тієї ж задачі інші методи машинного навчання:MLR, MLP, RBF,

SVM. При цьому в якості можливих факторів розглядалися різні оцінки студента.

Таким чином, найбільш поширеним методом для прогнозування результатів оцінювання студента є регресійні моделі, крім згаданих робіт, вони використовуються, наприклад, у [5], [6].

У даній роботі розглядається побудова математичної моделі експертного оцінювання рівня знань студента на прикладі оцінювання знань іноземної мови. За цією моделлю обчислюються імовірності отримання студентом на екзамені кожної з можливих оцінок. При цьому розглядається конкретний вид завдання, яке виконує студент.Його параметри, що впливають на кінцевий результат, визначаються на основі попереднього досвіду виконання ним завдань такого типу. Робиться припущення, що оцінка студента на екзамені визначається кількістю і грубістю допущених ним помилок. У моделі враховується, що допущення помилок є випадковим, а не детермінованим процесом, тому для його прогнозування недоцільно застосовувати регресію. Модель, що пропонується, повністю базується на математичному апараті теорії ймовірностей та математичної статистики.

Постановка задачі

Нехай для перевіркирівня знань іноземної мови студенту пропонується виконати певне завдання. Для прикладу, це може бути написання твору для перевірки граматики та орфографії, читання тексту для перевірки вимови або усна відповідь для перевіркивільностіволодіння мовою.

Будемо вважати, що невід’ємнадійснозначна випадкова величина (ВВ) S з функцією розподілу (ФР)$F_s(x)$у певному сенсі визначає результат виконання завдання студентом, причому

враховується кількість та грубість допущених помилок. Побудова ВВ S буде розглянута далі.

За виконання завдання на основі значенняs, прийнятого ВВS, студенту виставляється одна з можливих оцінок, яка є узгодженою оцінкою експертної комісії, що оцінює рівень знань з іноземної мови. Формування консолідованої експертної оцінки розглянуте в наступному пункті.

Задача даної роботи: на основі деякої апріорноїінформації про студента знайти імовірності отримання ним кожної з можливих оцінок, таким чином прогнозуючи результат перевіркирівняйого знань.

Задача формування експертної оцінки

Нехай до оцінюваннярівня знань студента на основіякості виконання завдання залучена комісія зm експертівE_i, $i = \overline{1,m}$. При цьому думки експертів можуть бути як рівноважливими, так і враховуватися з певним коефіцієнтом, наприклад, думка голови комісії може бути важливішою за думки іншихчленівкомісії. Задача експертів– виробити консолідовануоцінкурівня знань студента.

Нехай результатом оцінюваннярівня знань студента є одна з k можливих оцінокG_j, $j = \overline{1,k}$. При цьому оцінки впорядковано в порядку від найвищої (G_1) до найнижчої (G_k). Наприклад, у випадку Болонської системи, впровадженої увищих навчальних закладах України, $k = 6$,$G_1 \coloneqq A$, … , $G_k \coloneqq F$.

Будемо вважати, що кожному експерту E_i відповідають$k-1$ порогових значень T_{ij} : , якіподіляють область значень ВВ S на напівінтервали, що, на думку експерта, відповідаютьоцінкам G_j, $j = \overline{1,k}$. Умовно припишемо$T_{i0} \coloneqq 0$, $T_{ik} \coloneqq +\infty$. Тоді якщо $T_{ij-1} < S \leq T_{ij}$, то експерт E_i вважає за належне поставити студенту оцінкуG_j .

Припишемо кожнійоцінціG_j значення j , $j = \overline{1,k}$. Вважатимемо, що за консолідовануоцінку приймається оцінкаG_{j^*}, що відповідає середньому арифметичному значень оцінок, поставлених всіма експертами, округленому до найближчого цілого числа ($j + 0{,}5$ округлюється доj , на користь студента).

Впорядкуємо порогові значення всіх експертівT_{ij} за зростанням і введемо єдину нумерацію: (будемо вважати, що всіпорогові значення різні). Розглянемо напівінтервали, що утворилися на додатнійпівосі.

$$0 < S \le T_1$$

Якщо , це означає, що всі m експертів поставлять оцінку G_1, сума значень їх оцінокдорівнюєm ,$G_{j^*} = G_1$. Якщо $T_1 < S \le T_2$, це означає, що один з експертів поставить оцінку G_2, оскільки його перше порогове значення перевищено, сума значень оцінок експертів дорівнює $m + 1$, $G_{j^*} = G_{\left[\frac{m+1}{m}\right]}$ (під операцією $[\cdot]$мається на увазі округлення до найближчого цілого числа). Аналогічно, якщо $T_2 < S \le T_3$, то $G_{j^*} = G_{\left[\frac{m+2}{m}\right]}$ і т. д. Якщо $T_{m(k-1)} < S$, то всі m експертів поставлять оцінку G_k, $G_{j^*} = G_k$.

Таким чином, враховуючи правила округлення, імовірність того, що студент отримає оцінкуG_{j^*} визначається наступним чином:

$$P(G_{j^*} = G_j) = \begin{cases} P\left(S \le T_{\left[1+\frac{m}{2}\right]}\right), & j = 1 \\ P\left(T_{\left[1+mj-\frac{m}{2}\right]} < S \le T_{\left[1+mj-\frac{3m}{2}\right]}\right), & 1 < j < k \\ P\left(S > T_{\left[1+mk-\frac{3m}{2}\right]}\right), & j = k \end{cases}$$

або в термінах ФР ВВ S :

$$P(G_{j^*} = G_j) = \begin{cases} F_S\left(T_{\lfloor 1+\frac{m}{2} \rfloor}\right), & j = 1 \\ F_S\left(T_{\lfloor 1+mj-\frac{m}{2} \rfloor}\right) - F_S\left(T_{\lfloor 1+mj-\frac{3m}{2} \rfloor}\right), & 1 < j < k \\ 1 - F_S\left(T_{\lfloor 1+mk-\frac{3m}{2} \rfloor}\right), & j = k \end{cases} \quad (1)$$

У випадку можливої рівності оцінок $0 \le T_1 \le T_2 \le \cdots \le T_{m(k-1)}$ формула (1) лишається справедливою.

Якщо думки експертів є нерівнозначними, то кожному експерту E_i приписується вага його думки $w_i > 0$, $i = \overline{1,m}$, з якою значення його оцінки враховується в сумі значень оцінок всіх експертів. Без втрати загальності можна вважати, що $\sum_{i=1}^{m} w_i = 1$ (цього можна досягти нормуванням). Тоді аналогічно до випадку, коли думки експертів є рівнозначними:

$$P(G_{j^*} = G_j) = \begin{cases} F_S(T_{r^*}), & j = 1 \\ F_S(T_{r^*}) - F_S(T_{l^*}), & 1 < j < k \\ 1 - F_S(T_{l^*}), & j = k \end{cases} \quad (2)$$

де $l^* = min\left\{l \middle| \sum_{i=1}^{l} w_i > j - \frac{3}{2}\right\}$, $r^* = max\left\{r \middle| \sum_{i=1}^{r} w_i < j - \frac{1}{2}\right\}$.

Формалізація випадкової величини S

Для побудови ВВ S , що визначає результат виконання завдання студентом, можливі різні підходи, які залежать від конкретної постановки задачі. Зокрема можливе застосування пуассонівського біноміального (poissonbinomial)або складеного пуассонівського (compoundpoisson) розподілів. Обидва ці розподіли є розподілами сум ВВ у певному сенсі. У даній роботі зупинимось на визначенні ВВ S як зваженої суми пуассонівських ВВ.

Розглянемо наступну постановку задачі. Нехай для перевірки рівня знань іноземної мови студенту запропоновано виконати одне завдання певного виду, наприклад, написати твір, прочитати вголос

або розкрити усну тему, і у цьому завданні студент може допустити, помилки n різних типів. Наприклад, якщо запропонованим завданням є написання твору, то типами помилок можуть бути синтаксичні, морфологічні, орфографічні, лексичні тощо. Для визначення результату оцінювання знань студента експерти фіксуютькількість зроблених ним помилок. Проте очевидно, що помилки різнихтипів мають різний вплив на кінцевий результат, наприклад, граматичні помилки вважаються більш грубими, ніжорфографічні. Для того, щоб це врахувати, кожному типу помилок присвоюється ваговий коефіцієнт$\theta_i > \mathbf{0}$, $i = \overline{1,n}$, чим більшим є значення вагового коефіцієнта, тим грубішим є тип помилки. Виходячи з цього, результат виконання завдання студентом визначається як зважена сума

$$s = \sum_{i=1}^{n} \theta_i x_i$$

де x_i - кількість помилок і-го типу, допущених студентом. До цієї суми експерти застосовують свої критеріїоцінювання і визначають остаточну консолідовануоцінку.

Допущення студентом помилок певного типу доцільно розглядати як найпростішийпотікподій. Події, що полягають у допущенні помилки, наступають на інтервалі, що вимірюєтьсякількістюслів у випадку написання твору або читання тексту та тривалістю мовлення у випадку усного завдання. Умови, накладені на найпростішийпотік [7], можна вважати виконаними.

Справді, кількість помилок, зроблених на певному проміжку не залежить анівід його розміщення, анівід помилок на попередніхпроміжках, також події наступають поодинці: не може бути одночасно допущено кілька помилок одного типу. Тоді дискретна ВВ $X(t)$, що задає кількість помилок певного типу на

інтервалі $[0,t)$, є розподіленою за законом Пуассона з параметром λt ($X(t) \in Poiss(\lambda t)$), де λ –інтенсивність потоку.

Для таких завдань як написання твору або читання тексту вякостіλ виступає середня кількість помилок цього типу на одне слово, вякостіt –кількістьслів у тексті. Для перевірки усного мовлення λ – це середня кількість помилок цього типу за одиницю часу, t – час мовлення.

У ситуаціїперевірки знань t можна вважати фіксованим, адже обсяг твору та тексту для читання або тривалість мовлення задається зазделегідь. Тому будемо розглядати ВВ $X_i \in Poiss(a_i)$, $i = \overline{1,n}$, що задають кількість помилок і-го типу, зроблених студентом протягом повного виконання завдання. При цьому $a_i = \lambda_i t$, де t є відомим параметром, який визначається ситуацією перевірки знань, а λ_i – середня кількість помилок і-го типу на одиницю вимірювання є параметром, специфічним для кожного студента. Параметри λ_i необхіднооцінювати за вибіркою звідомими результатами виконаних студентом завдань.

Таким чином, за наведеної постановки задачі дискретна ВВS, що характеризує якість виконання завдання студентом, є зваженою сумою n ВВ, розподілених за законом Пуассона:

$$S = \sum_{i=1}^{n} \theta_i X_i \qquad (3)$$

Для обчислення імовірностей отримання студентом кожної з можливих оцінокзгідно з формулами (1), (2)необхідно знайти ФР ВВ S для цього студента з урахуванням його особистих параметрів λ_i, $i = \overline{1,n}$.

Попри те, що розподіл Пуассона є стійким, зважена сума пуассонівських ВВ, очевидно, вже не буде пуассонівською ВВ, також

ВВ S в загальному випадку є дійснозначною. Розподіл ВВ S не має відомої параметричної форми, є сильно нерегулярним, в чому полягає складність його вивчення. Тому необхідно виробити підходи до знаходження ФР такої ВВ.

Знаходження функції розподілу

Спочатку розглянемо випадок, коли події, які полягають в допущенні помилки і-го та j-го типу, є незалежними $\forall i,j=\overline{1,n}$, $i\neq j$, тобто ВВ X_i є незалежними у сукупності. Таке припущення може справджуватись на практиці, адже деякі типи помилок можна вважати незалежними, наприклад, орфографічні та пунктуаційні помилки у творі.

Пропонується 3 методи точного обчислення ФР ВВ S.

Перший метод заснований на знаходженні значення ФР ВВ S в кожній точці x напряму за допомогою рекурсивної процедури, яка описується співвідношенням:

$$F_S(x)=\sum_{i_1:\theta_1 i_1\le x} P(X_1=i_1)\cdot P\left(\sum_{k=2}^{n}\theta_k X_k\le x-\theta_1 i_1\right)=\sum_{i_1:\theta_1 i_1\le x} P(X_1=i_1)\times$$

$$\times\sum_{i_2:\theta_2 i_2\le x-\theta_1 i_1} P(X_1=i_1)\cdot P\left(\sum_{k=3}^{n}[\theta_k X_k\le(x-[\theta_1 i_1)-\theta_2 i_2\right)=\cdots$$

Другий метод базується на понятті генератриси ВВ. У традиційному розумінні генератриса визначається для цілочисловихневід`ємнозначних дискретних ВВ [8]. Але у роботі Мануеля [9] наведено узагальнення поняття генератриси для дискретних дійснозначних ВВ. Для дискретної ВВ X, яка набуває дійсних значень $\{\alpha_k\}_{k\in Z}$ (нехай $\alpha_k<0,\quad k<0$; $\alpha_k>0, k>0$; $\alpha_k=0, k=0$), вона визначається як:

$$G_X(t)=\sum_{k=-\infty}^{+\infty}[P(X=[\alpha_k)t^{\alpha_k}$$

Властивість генератриси суми незалежних ВВ при цьому зберігається:

$$G_{X+Y}(t) = G_X(t) \cdot G_Y(t)$$

Також відомо, що послідовність ВВ $X^{M,N}$ визначених, як:

$$\begin{cases} P(X^{M,N} = \alpha_k) = P(X = \alpha_k) \quad \forall k \in \{-M, \dots, -1, 1, \dots, N\} \\ P(X^{M,N} = \mathbf{0}) = P(X = \mathbf{0}) + \sum_{k=-\infty}^{-M-1} P(X = \alpha_k) + \sum_{k=N+1}^{+\infty} P(X = \alpha_k) \end{cases}$$

збігається за розподілом до ВВ X [9].

Виходячи з цього, пропонується обрати мале число $\varepsilon > 0$, яке визначає, якими імовірностями можна знехтувати. Потім необхідно знайтигенератриси ВВ $\theta_i X_i$ (значення імовірностей збігаються з відповідними імовірностями розподілу пуассонівської ВВ X_i), нехтуючи тими доданками, де $P(X_i = \alpha_k) < \varepsilon$. Далі отримані таким чином n виразів необхідно перемножити та звести подібні. Цю процедуру для полегшення рекомендується виконувати засобами для символьних обчислень деякого прикладного пакету, наприклад, WolframMathematica, Matlab або бібліотекиsympy мови програмування Python. За отриманою в результаті генератрисою ВВ S можна знайти її розподіл та за ним обчислити ФР.

Третій метод базується на понятті характеристичної функції та є ідейно схожим з другим. Але якщо другий метод є метод є універсальним і може бути аналогічно застосований для пошуку ФР зваженої суми будь-яких дискретних ВВ, навіть не обов'язково розподілених за одним законом, то цей метод враховує, що ВВ у зваженійсумірозподілені саме за законом Пуассона.

Якщо $X \in Poiss(a)$, то характеристична функція має вигляд: $\varphi_X(t) = exp[a(e^{it} - 1)]$. Справедливі наступні теореми [8]:

Теорема. Нехай $Y = \alpha X + \beta$. Тоді:

$$\varphi_Y(t) = e^{it\beta}\varphi_X(\alpha t) \qquad (4)$$

Теорема. Нехай X та Y є незалежними ВВ. Тоді:

$$\varphi_{X+Y}(t) = \varphi_X(t) \cdot \varphi_Y(t) \quad (5)$$

Отже, за (4), якщо $X_j \in Poiss(a_j)$, то $\varphi_{\theta_j X_j} = exp\left[a_j\left(e^{i\theta_j t} - 1\right)\right]$.

Таким чином, за (5) для зваженої суми (3) маємо:

$$\varphi_S(t) = exp\left[\sum_{j=1}^{n} a_j\left(e^{i\theta_j t} - 1\right)\right] = exp\left[-\sum_{j=1}^{n} a_j\right] \cdot exp\left[a_j e^{i\theta} \quad (6)\right.$$

Пропонується розкласти множник$exp[a_j e^{i\theta_j t}]$ у виразі (6) за формулою Тейлора в околі точки0:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + o(x^n)$$

до певного члену N. Якщо необхідно знайти ФР на відрізку $[x_{min}; x_{\max}]$, то в якості N рекомендується взяти $N = \left\lfloor \frac{x_{max} + 1}{\theta_i} \right\rfloor$. При такій кількості доданків будуть враховані всіможливіваріанти отримання значення, близького до $x_{\max}$, як лінійноїкомбінації вагових коефіцієнтівθ_i, $i = \overline{1,n}$.

Таким чином після перемноження та зведення подібних$\varphi_S(t)$ з певною точністю апроксимується деякою сумою членів виду $\beta e^{i\alpha t}$. Кожен такий член відповідає значенню α, яке ВВ S приймає з імовірністю β. Тобто отримуємо розподіл ВВ S, за яким можна знайти ФР.

Зручність такого підходу в тому, що операції додавання, множення та ділення на число виразів виду $\beta_1 e^{i\alpha_1 t} + \cdots + \beta_m e^{i\alpha_m t}$ виконуються достатньо просто і їх представлення у програмі можно ефективно організувати у вигляді, наприклад, словника, ключами якого є дійсні значення α. Загалом для отримання шуканого виразу для $\varphi_S(t)$ необхідно виконати N ітерацій алгоритму, який на кожній k-ій ітерації додає до поточного виразу вираз $\frac{\left(exp\left[a_j e^{i\theta_j t}\right]\right)^k}{k!}$, причому

чисельник обчислюється на основі збереженого на попередньому кроці виразу для $\left(exp\left[a_j e^{i\theta_j t}\right]\right)^{k-1}$. На початку роботи алгоритму необхідно задати $\varphi_S(t) := 1e^{i0t}$.

Всі три методи точного знаходження ФР зваженої суми пуассонівських ВВ були протестовані на наступному прикладі:

$$S = 0{,}35X_1 + 0{,}2X_2 + 0{,}12X_3 + 0{,}33X_4,$$

$$X_1 \in Poiss(2{,}1), X_2 \in Poiss(2{,}5), X_3 \in Poiss(3{,}0), X_4 \in Poiss(1{,}7)$$

Алгоритми були програмно реалізовані мовою Python (версія 2.7).

На рис. 1 наведено графіки отриманих трьома методами ФР ВВ S на відрізку $[0;\ 7]$ з кроком дискретизації 0.01.

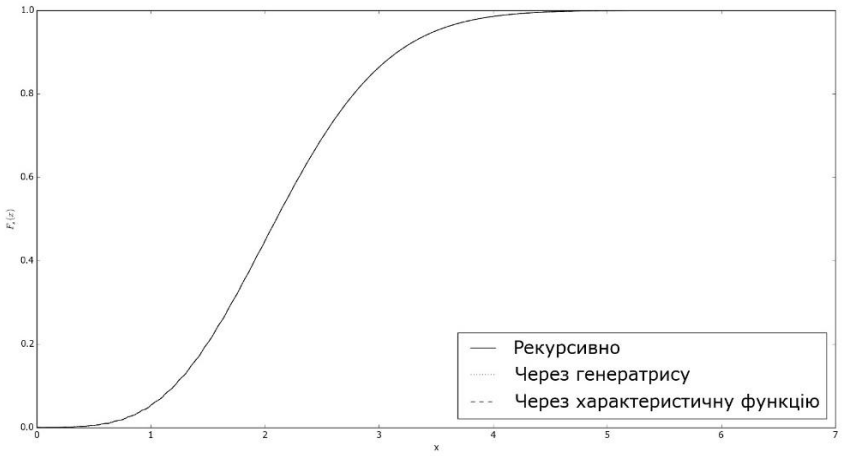


Рис. 1Графіки ФР, отриманих трьома методами

Вони є ідентичними (максимальне з попарних максимальних абсолютних відхилень для всіх трьох методів має порядок 10^{-3}), з чого можна зробити висновок, що усі три запропоновані методи знаходження ФР працюють правильно. Відхилення виникають винятково через представлення числових даних у пам'яті.

Середній час роботи 1-го методу на цьому прикладі становить 10 с, 2-го методу – 70 с, 3-го методу – 15 с. Отже швидкодія другого методу досить істотно нижча, ніж у інших двох. Це пояснюється тим, що у другому методі використовуються засоби символьної

математики, що сповільнюють обчислення. Швидкодія усіх трьох методів суттєво знижується зі зменшенням кроку дискретизації та особливо зі збільшенням кількості доданків n у зваженій сумі.

У загальному випадку найбільш оптимальним є третій метод з використанням характеристичних функцій. При невеликій кількостідоданківn він незначно поступається у швидкодії рекурсивному методу, але при збільшенні n, він стає суттєво швидшим, а при достатньо великому n (вже при $n = 8$) стає безальтернативним, оскільки в першому методі стає неприпустимою глибина рекурсії, а другий працює надмірний час. Також він, як і другий метод дозволяє отримати, крім ФР, побічно також і розподіл ВВ S.

Розподіл ВВ S, що розглядалася у прикладі, отриманий за допомогою третього методу наведений на рис. 2.

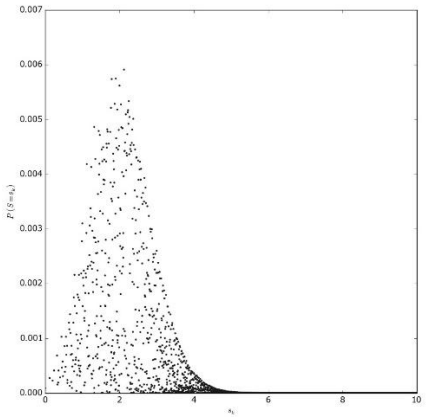


Рис.2 Розподіл ВВ S

Як видно з графіка, розподіл є сильно нерегулярним і його неможливо достатньо точно апроксимувати нормальним розподілом, спираючись на центральну граничну теорему (ЦГТ), оскільки є багато значень, імовірності прийняття яких знаходяться нижче за обвідну. Цьому сприяють також невеликі значення параметрів пуассонівських розподілів у сумі. Проте, оскільки пуассонівські ВВ у представленій моделі використовуються для підрахування кількості помилок, які за

змістом допускаються досить рідко, то параметри, що виникатимуть у моделі будуть саме невеликими.

Враховуючи, що при збільшенні кількості доданків навіть найшвидший з методів точного обчислення ФР ВВ S працює неприйнятний час (730 с. при $n = 8$) доцільно розглянути методи апроксимації ФР, що розглядається.

Найпоширенішим способом апроксимації ФР є апроксимація за допомогою ФР нормального закону, спираючись на ЦГТ.

В даному випадку $E[S] = \sum_{i=1}^{n} \theta_i a_i$, $D[S] = \sum_{i=1}^{n} \theta_i^2 a_i$ і нормальна апроксимація для ФР ВВ S задається наступною формулою [8]:

$$F_S(x) \approx \frac{1}{2}\left(1 + \Phi\left(\frac{x - E[S]}{\sqrt{2D[S]}}\right)\right)$$

де $\Phi(x)$ – функція Лапласа.

Іншим способом апроксимації ФР є використання методу Монте-Карло, тобто апроксимації за допомогою емпіричної ФР, яка за теоремою Глівенко-Кантеллі збігається до теоретичної ФР за імовірністю [9].

У даному випадку для використання цього підходу необхідно згенерувати N наборів реалізацій n пуассонівських ВВ з параметрами a_i, $i = \overline{1,n}$ та за ними знайти реалізації ВВ S, як лінійної комбінації пуассонівських ВВ. За згенерованими реалізаціями знаходиться емпірична ФР ВВ S.

Всі три методи апроксимації ФР зваженої суми пуассонівських ВВ були протестовані на наступному прикладі:

$$S = 0{,}35X_1 + 0{,}2X_2 + 0{,}12X_3 + 0{,}33X_4 + 0{,}24X_5 + 0{,}5X_6 + 0{,}7X_7 + 0{,}17X_8,$$

$$X_1 \in Poiss(2{,}1), X_2 \in Poiss(2{,}5), X_3 \in Poiss(3{,}0), X_4 \in Poiss(1{,}7),$$

$$X_5 \in Poiss(3{,}5), X_6 \in Poiss(1{,}3), X_7 \in Poiss(4{,}5), X_8 \in Poiss(3{,}8)$$

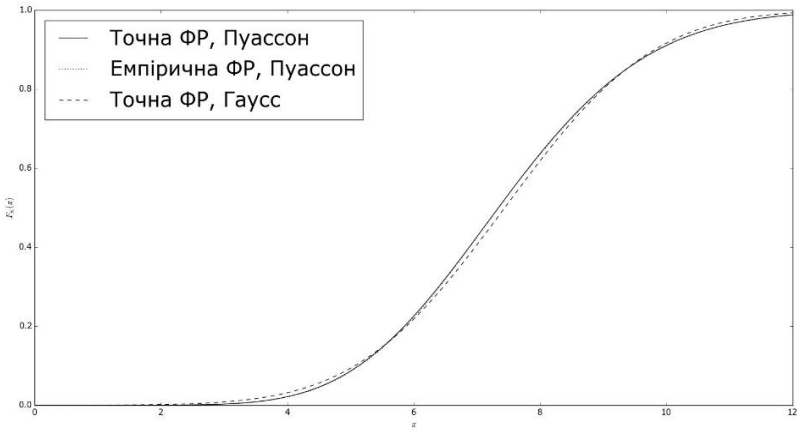


Рис. 3 Графіки точної та апроксимованих ФР ВВ S

На рис. 3 наведено графіки отриманих точної ФР, обчисленої за допомогою характеристичних функцій, та апроксимацій ФР обома методами на відрізку $[0;\ 12]$ з кроком дискретизації 0.01 . Використане значення N в методі Монте-Карло 10000 .

Теоретична ФР ВВ S та емпірична ФР для зваженої суми пуассонівських ВВ є майже ідентичними, максимальне абсолютне відхилення становить 0.0024 , тобто має порядок 10^{-3}. Натомість має місцепомітневідхилення цих графіків від ФР при гауссівськійапроксимації, максимальне абсолютне відхилення становить 0.0216 , тобто має порядок 10^{-2}. Таким чином, точність апроксимації методом Монте-Карло є на порядок вищою, ніж при використанні нормального розподілу, причому відхилення є прийнятним для використання такої апроксимації замість знаходження точної ФР. Отже, у практичних застосуваннях рекомендується використовувати для апроксимації ФР зваженої суми пуассонівських ВВ з малими значеннями параметрів саме метод Монте-Карло.

Необхідно також окремо розглянути випадок, коли ВВ X_i є залежними. Це у багатьох випадках відповідаєпрактичнійзадачі. Допущення деяких типів помилок не можна вважати незалежними

подіями. Наприклад, очевидно, залежними є помилки, що стосуються неправильного вживання граматичного часу та неправильної побудови речення, бо перша тягне за собою другу.

Знаходження точної ФР у цій ситуації є математично надто складним, тому доцільно використовувати апроксимацію. З огляду на те, що у випадку незалежних ВВ X_i апроксимація емпіричною ФР є точнішою за нормальну, у цьому випадку також рекомендується використовувати метод Монте-Карло. Необхідні реалізації пуассонівського випадкового вектора з заданими параметрами можуть бути згенеровані за допомогою алгоритму Сіма, викладеного зокрема в роботі [10]. Але він передбачає деякі обмеження на коефіцієнти коваріаційної матриці.

Застосування моделі на практичному прикладі

Для демонстрації роботи моделі наведемо практичний приклад.Нехай екзамен студента полягає у написанні твору обсягом 200 слів. Студента оцінюють 5 експертів, їхня консолідована оцінка є середнім арифметичним суб'єктивних оцінок кожного експерта. За виконання твору студент може отримати одну з оцінок від A до F згідно з Болонською системою. Необхідно знайти імовірності отримання ним кожної оцінки.

Типи оцінок та середні частоти їх допущення – значення параметрів моделі λ_i були взяті для даного прикладу з роботи [11]. У ній викладені результати написання твору з англійської мови обсягом 200-250 слів 72-ма учнями старшої школи міста Семенійх (Малайзія), для яких англійська мова є іноземною. Було виділено 14 типів помилок та обчислено загальну та середню кількість помилок кожного типу, допущених студентами.

У таблиці 1 наведені назви типів помилок, вагові коефіцієнти, погоджені з експертом з англійської мови, які відповідають грубості

кожного типу помилок, та середні частоти допущення помилки кожного типу в розрахунку на одне слово (за браком додаткової інформації вважалось, що кожний з творів містив 200 слів).

Таблиця 1 Параметри типів помилок

№	Назва типу помилки	Середнячастотадопущення, λ_i	Ваговий коефіцієнт, θ_i
1	Singular/PluralForm	0.0286	0.35
2	VerbTense	0.024	1.0
3	Word Choice	0.0226	0.18
4	Preposition	0.02	0.6
5	Subject/VerbAgreement	0.0151	0.9
6	Word Order	0.015	0.32
7	Article	0.0147	0.75
8	Word Form	0.0118	0.68
9	Spelling	0.0104	0.27
10	VerbForm	0.0101	0.85
11	Capitalization	0.009	0.23
12	Wrong/Misused Word	0.0085	0.55
13	Missingword	0.0071	0.2
14	Redundancy	0.0054	0.1

Вважатимемо середні значення, наведені в таблиці 6.1 параметрами λ_i модельного студента, який розглядається. Відповідно до обсягу твору, що вимагається при складанні екзамену, $t = 200$.

Деякі типи помилок можна вважати залежними, відповідно до цього наступні коефіцієнти кореляції є ненульовими: $\rho_{15} = 0{,}3$; $\rho_{210} = 0{,}45$;

$\rho_{312} = 0{,}2$;$\rho_{59} = 0{,}2$;$\rho_{614} = 0{,}3$; $\rho_{89} = 0{,}45$;$\rho_{912} = 0{,}2$;$\rho_{1214} = 0{,}3$.

Порогові значення T_{ij}, $i,j = \overline{1,5}$, 5 експертівE_i, що відповідають можливим 6 оцінкам (від A до F), наведені у таблиці 2.

Таблиця 2 Порогові значення експертів

	E_1	E_2	E_3	E_4	E_5
T_1	12,1	10,3	13,0	11,75	14,7
T_2	18,6	16,5	19,0	17,8	20,3
T_3	24,5	22,0	26,0	23,2	27,5
T_4	33,0	29,7	35,0	31,5	37,9
T_5	39,2	36,4	42,0	38,5	43,5

У таблиці 3 наведено імовірності отримання кожної оцінки, обчислені за формулою (1)з використанням ФР ВВ S – зваженої суми залежних пуассонівських ВВ, апроксимованої за допомогою емпіричної ФР. Обсяг вибірки становить$N = 10000$, реалізації пуассонівського випадкового вектора генеруються за допомогою алгоритму Сіма [10].Обчислення проводились за допомогою програмного продукту, створеного мовою Python.

Таблиця 3Розподілоцінок для модельного студента

Оцінка	**Імовірність отримання**
A	0,0162
B	0,2721

C	0,5629
D	0,1471
E	0,0017
F	0,0

Обговорення отриманих результатів

Можна побачити, що найбільш імовірною оцінкою, яку отримає модельний студент за даних умов, є оцінка C, причому ця імовірність становить більше 0.5 . При цьому студент отримає одну з оцінок B, C або D з імовірністю$\approx 0{,}98$, тобто майже напевно. Можна зробити висновок, що модельний студент точно складе екзамен успішно (неуспішному складанню екзамену відповідає оцінка F), але при своїх нинішніх показниках та такому складі екзаменаційної комісії імовірність отримання найвищої оцінки A вкрай мала.

Такі результати є достатньо адекватними та свідчать про здатність моделі до ефективного прогнозування розподілу оцінок студента.

Висновки

Побудована математична модель оцінювання знань студента дозволяє прогнозувати імовірності отримання студентом кожної з можливих оцінок, на відміну від регресійних моделей, які у таких випадках видають єдиний результат без оцінки імовірності його досягнення. Також прогнозування у даній моделі спирається на особисті параметри студента, специфічні для завдання, яке розглядається, такі як середня частота допущення помилок кожного типу. При цьому модель може бути застосована до будь-якого навчального предмету, приклад іноземної мови був обраний лише тому, що в ньому найбільш чітко розрізняються різні види помилок та оцінюється їхня грубість.

У даній роботі розглянута задача точного знаходження та апроксимації ФР зваженої суми пуассонівських ВВ. Визначено, що серед запропонованих методів точного знаходження ФР оптимальним є третій розглянутий метод з використанням характеристичних функцій. Усі три методи є адаптованими автором саме до розподілу, що вивчається. Для апроксимації ФР рекомендується використовувати метод Монте-Карло, оскільки він дає точніші результати, ніж нормальна апроксимація.

Дослідження властивостей розподілу зваженої суми пуассонівських ВВ та вироблення підходів до роботи з ним є важливим завданням, оскільки він доволі часто виникає на практиці. Наприклад, він виникає у медичних застосуваннях при моделюванні безпосередньо стандартизованих рівнів смертності [12]. Проте через складну ФР він найчастіше апроксимується за допомогою нормального розподілу, що, як показано в даній роботі, може не забезпечувати прийнятної точності. Доцільно вивчити також методи точкового оцінювання та побудови довірчих інтервалів для параметрів цього розподілу.

Роботу моделі було продемонстровано на практичному прикладі. Отримані результати показують, що модель є достатньо адекватною та її можна рекомендувати до впровадження в навчальному процесі.

На основі результатів, отриманих за моделлю, можна також розв'язувати інші задачі, наприклад, задачу знаходження імовірності того, що середній бал студента за кілька іспитів перевищує деяке задане значення. Для даної задачі те, що модель дозволяє знаходити імовірності отримання всіх оцінок, є ключовим.

Подальші дослідження можуть стосуватися побудови аналогічної математичної моделі на основі інших розподілів та залучення інших методів.

Література:

1. Li K., Uvah J., Amin R.PredictingStudents' PerformanceinElementsofStatistics // US-ChinaEducationReview, 2012,Vol. 10, pp. 875-884
2. Dankoa K., Dukea J.C., Franza D.P. PredictingStudentPerformanceinAccountingClasses// TheJournalofEducationforBusiness, 1992,Vol.5, pp. 270-274
3. Fang N., Lu J. A decisiontreeapproachtopredictivemodelingofstudentperformanceinengineeringdynamics// InternationalJournalofEngineeringEducation, 2010, Vol. 1, pp. 87-95.
4. Huang S. PredictiveModelingandAnalysisofStudentAcademicPerformanceinan Engineering Dynamics Course // AllGraduateThesesandDissertations. Paper 1086, 2011,pp. 1-116
5. Chamillard A.T.UsingStudentPerformancePredictionsina ComputerScienceCurriculum// Proceedingsofthe 11th annual SIGCSE conferenceonInnovationandtechnologyincomputerscienceeducation, 2006,pp. 260-264
6. Felder R.M., Forrest K.D., Baker-Ward L., Dietz E.J., Mohr P.H. A longitudinalstudyofengineeringstudentperformanceandretention: I. Successandfailureintheintroductorycourse// AJournalofEngineeringEducation, 1993, Vol. 82, pp. 15-21

7. Haight F.A. Handbookofthepoissondistribution. –NewYork : Wiley, 1967. – 168 p.
8. Wasserman L. AllofStatistics. –NewYork : Springer, 2004. – 442 p.
9. Manuel S.E. ProbabilityGeneratingFunctionforDiscreteRealValuedRandomVariables– 2004. – 17 p.
10. Masahiko G., Kazushi M.ComparisonofAlgorithmsforGeneratingPoissonRandomVectors: A Review // Interstat, August, 2012, pp. 1-10.
11. Saadiyah D., Kaladevi S.ErrorAnalysisoftheWrittenEnglishEssaysofSecondarySchoolStudentsinMalaysia: A CaseStudy// EuropeanJournalofSocialSciences, 2009,Vol. 8, pp. 483-495
12. Breslow N.E., Day N.E. Statisticalmethodsofcancerresearch / -- Lyon : IARC ScientificPublications, 1987. – 406 p.

УДК 519.246.8
Грицаюк М.О.

Автоматизований вибір прогнозуючих моделей в СППР

Вступ

У даній статті поставлена та вирішена задача автоматизованого вибору прогнозуючих моделей у системі підтримки прийняття рішень.

Сьогодні існує необхідність побудови адекватних математичних і статистичних моделей для оцінювання коротко- і довгострокових прогнозів з метою підвищення якості рішень, які приймаються на їх основі.

Також, оскільки у таких поширених програмних додатках, як Eviews та STATISTICA, відсутній автоматизованих вибір кращої моделі, то тема є надзвичайно актуальною.

Дослідження присвячене розробці комбінованого критерію якості, який поєднав би в собі критерії якості моделей та оцінок прогнозів.

Постановка задачі

– Розробити комбінований критерій для автоматизованого вибору кращої регресійної моделі, побудованої на основі статистичних даних.

– Розробити програмне забезпечення для виконання необхідних обчислювальних експериментів (мова програмування C#).

– Зібрати статистичні дані для виконання обчислювальних експериментів.

– Побудувати регресійні моделі на основі зібраних даних та обчислити статистичні оцінки адекватності моделей та якості прогнозів.

– Вибрати кращі моделі для короткострокового прогнозування та виконати порівняльний аналіз отриманих результатів.

Інтегральний критерій якості моделі

Якість моделі оцінюють за допомогою багатьох статистичних критеріїв якості, зокрема таких: коефіцієнта множинної детермінації (R^2), який характеризує інформативність моделі по відношенню до інформативності даних; статистики Дарбіна-Уотсона (DW), що

визначає ступінь автокорельованості похибок моделі; інформаційного критерію Акайке (AIC) і статистики Байєса-Шварца (BSC); суми квадратів похибок моделі ($\sum e^2(k)$); $F-$статистики Фішера та інших[1]. Для автоматизованого вибору кращої моделі можна скористатись запропонованим в даній статті інтегральним критерієм якості:

$$IK = e^{1-R^2} + \frac{SSE}{N} + \begin{cases} \ln(AIC + BSC), & якщо\ AIC + BSC > 0 \\ e^{AIC+BSC}, & якщо\ AIC + BSC \le 0 \end{cases} + e^{2-DW} +$$

$$+ \ln(СКП) + \ln(САПП) + e^{U}$$

Даний критерій комбінує основні статистичні характеристики, які використовуються для визначення якості моделі та критерії, що використовуються для визначення якості оцінок прогнозів.

Архітектура та функціональна схема розробленої СППР

Розглянемо архітектуру розробленої в рамках даної роботи системи підтримки прийняття рішень, вона представлена на рис.1.

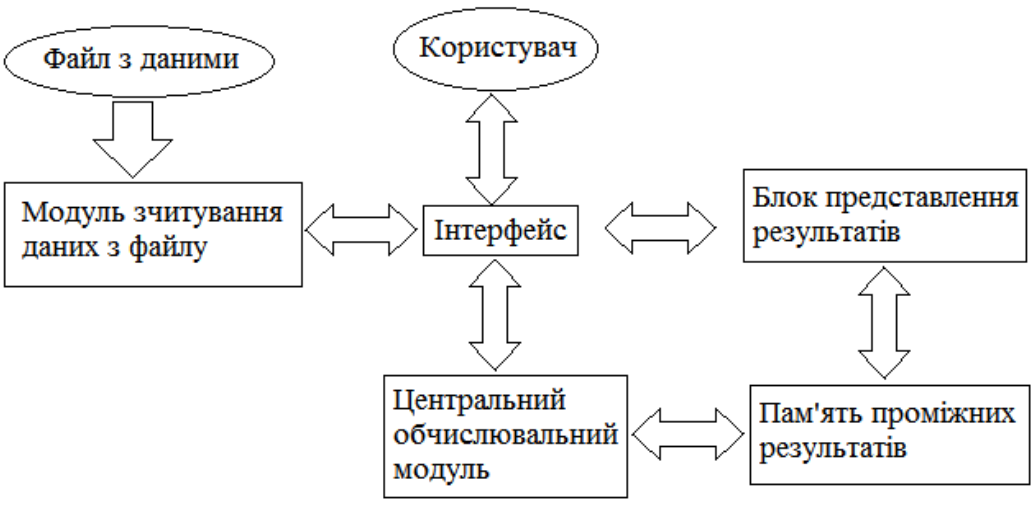


Рисунок 1 – Архітектура розробленої СППР

Як видно з рис. 1, користувач взаємодіє з програмним додатком через інтерфейс програми. Також через інтерфейс відбувається

керування зчитуванням даних з файлу, за яке відповідає окремий модуль. Зчитуванням, звичайно, керує користувач через інтерфейс. Далі дані потрапляють до центрального обчислювального модуля, в якому і відбувається найважливіший етап роботи програми: попередня обробка даних та необхідні розрахунки (див. функціональну схему на рис. 2). Всі обчислення зберігаються в пам’яті проміжних результатів та передаються до блоку представлення даних, а звідти на інтерфейс. Відповідно, на інтерфейсі користувач бачить всі необхідні результати обчислень.

Розглянемо детально функціональну схему розробленої в рамках даної роботи системи підтримки прийняття рішень, вона представлена на рис. 2.

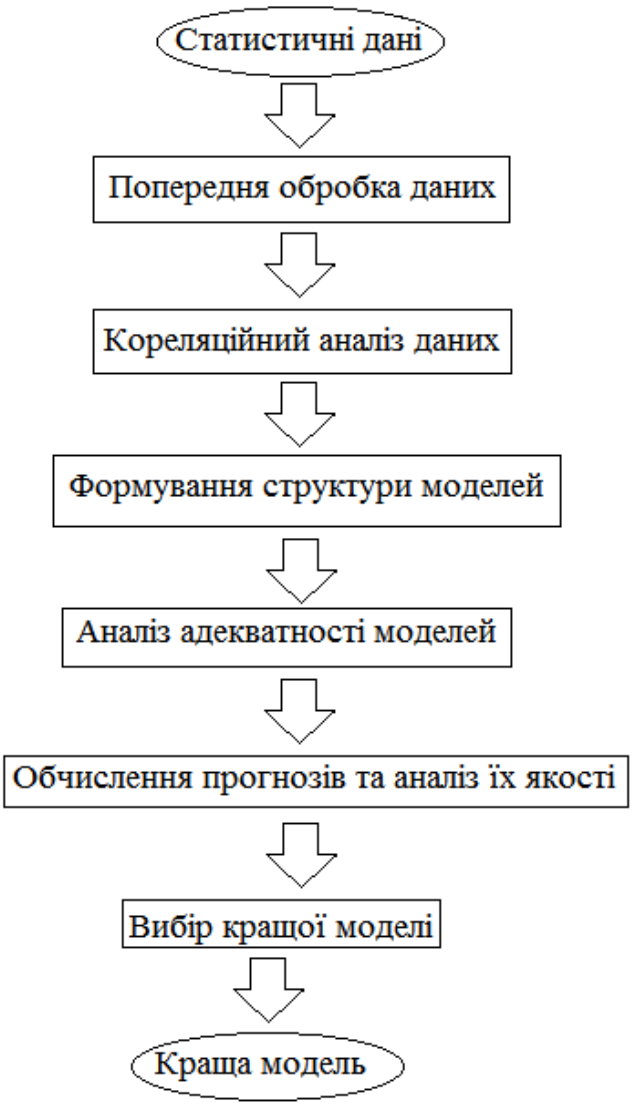


Рисунок 2 – Функціональна схема розробленої СППР

На вхід системи подається вибірка – статистичні дані. Вибірка проходить попередню обробку даних, в розробленій СППР – це нормування. Після чого відбувається кореляційний аналіз даних, в СППР, що представлена в даній роботі, використана автокореляційна функція. Далі відбувається формування структури моделей-кандидатів, для цього в роботі були використані МНК та РМНК. На наступному етапі виконується аналіз адекватності моделей, в розробленій СППР використовувались такі статистики: *SSE,* R^2, *DW, AIC, BSC*. Після чого виконується обчислення прогнозів та аналіз їх якості (в розробленій СППР використані такі критерії якості оцінок

прогнозів: *САПП, СКП, U*). На фінальному етапі відбувається вибір кращої моделі за допомогою запропонованого в рамках даної роботи інтегрального критерію якості моделей. На виході системи ми отримуємо кращу модель за обраними характеристиками та розробленим інтегральним критерієм якості.

Інструментальна платформа та вимоги до інструментарію по експлуатації розробленої СППР

Для даної роботи була вибрана мова програмування C# та середовищерозробки програмного забезпечення Microsoft VisualStudio 2015.

Для експлуатації розробленої СППР необхідний не дуже потужний сучасний комп’ютер з такими характеристиками:

- персональний комп'ютер або ноутбук з процесором частотою 2 Ггц або більше;
- операційні системи: XP, Vista, 7, 8, 8.1, 10;
- встановлений на комп’ютері .NET Framework 4.5.

Приклади використання розробленої СППР

Створена система підтримки прийняття рішень була апробована для моделювання та прогнозування реальних економічних процесів, а саме: валовий внутрішній продукт України, індекс споживчих цін України та грошовий агрегат М3 (загальна грошова маса) в Україні. Для даних процесів були автоматично обрані за допомогою СППР кращі регресійні моделі. Для цих моделей за допомогою СППР були побудовані графіки. Далі детально розглянутий кожний з проведених експериментів. А зараз розглянемо докладно інтерфейс СППР та її функціональні можливості (див. рис. 3).

Інтерфейс розробленої СППР

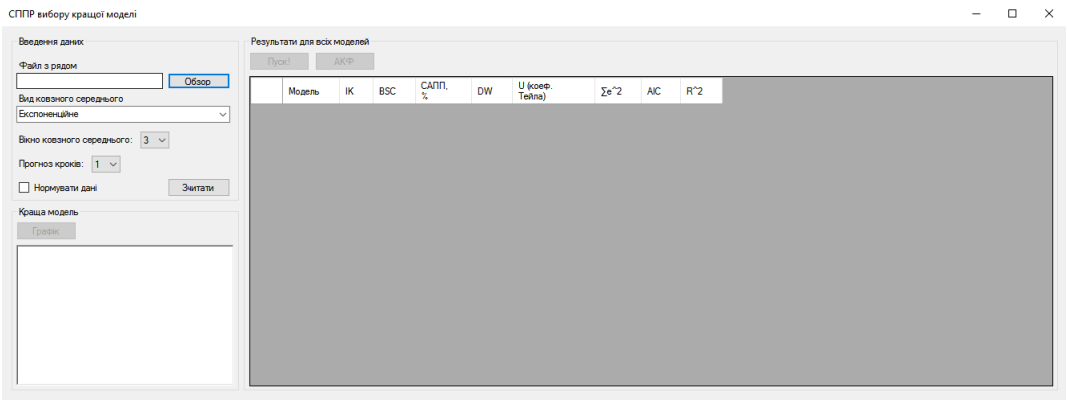


Рисунок 3 – Інтерфейс розробленої СППР

Зліва у верхньому куті, де написано «Файл з рядом», можна обрати файл з вибіркою, файл повинен бути у форматі .txt, виміри вибірки повинні бути розділені між собою новим рядком (кнопка «Enter»).

«Вид ковзного середнього» дозволяє обрати спосіб обрахунку ковзного середнього, існують такі варіанти: експоненційне ковзне середнє зі зменшенням коефіцієнтів від поточного виміру, експоненційне ковзне середнє у формі e^n, просте ковзне середнє, шум накладений на вибірку – рекомендується використовувати тільки для нормованих даних (див. розділ 1 для більш детальної інформації про види ковзного середнього).

Також можна обрати розмір вікна ковзного середнього: 3, 5, 7 або 9.

«Прогноз кроків» дозволяє задати кількість кроків прогнозування: від 1 до 5.

Також можна поставити птичку та обрати нормувати дані чи ні.

Після налаштування всіх параметрів, що були описані вище, треба нажати на кнопку «Зчитати», програма зчитає дані з файлу та можна продовжити роботу.

Для того, щоб переглянути АКФ треба натиснути на відповідну кнопку, яка стане активною після натискання на «Зчитати».

Для того, щоб обрахувати характеристики якості моделей та оцінок прогнозів для всіх моделей-кандидатів та обрання СППР кращої моделі, необхідно натиснути кнопку «Пуск!». У таблиці, що зараз пуста на рис. 3 з'являться відповідні параметри та у лівому нижньому куті буде показана краща модель.

Таблицю можна буде сортувати за параметрами, які вам потрібно, натискаючи на відповідний стовпець.

До речі, перебір моделей-кандидатів відбувається до тих степеней, для який АКФ > 0.3.

Валовий внутрішній продукт України

Дані взято з сайту Державної служби статистики України [2].

Таблиця 1 ВВП України у 1996-2005 рр. (млн.грн)

Місяці	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Січень	5138	5920	6196	7500	9934	13616	14128	15021	19099	24278
Лютий	6109	6107	6398	7887	10887	13720	13412	15562	19231	
березень	7084	6670	8080	9097	11910	15529	14876	17763	23764	
квітень	6628	7524	7173	8912	12435	15719	16128	18903	24551	
травень	5730	6998	6955	8501	12705	16104	17080	18509	23225	
червень	5021	6517	7527	9601	13466	16611	16877	19125	27671	
Липень	7209	7860	7498	10310	16104	19926	20317	20759	31579	
серпень	6740	7503	7727	10526	16564	21388	23474	25481	33282	
вересень	5115	7927	8350	11420	17969	19393	19043	25433	31903	
жовтень	7310	8847	9036	11396	15410	18421	18701	25128	32143	
листопад	6529	8205	8825	10765	16662	18121	18623	24654	31341	
грудень	8306	9876	11058	12545	18906	21500	18536	24314	30287	

Результат роботи програми з файлом, у якому знаходяться дані з таблиці 4 представлений на рис.4. Тут і далі для всіх експериментів буде використовуватись експоненційне ковзне середнє з вікном 3, прогнозом на один крок та нормуванням даних, адже саме такі параметри дають кращі результати.

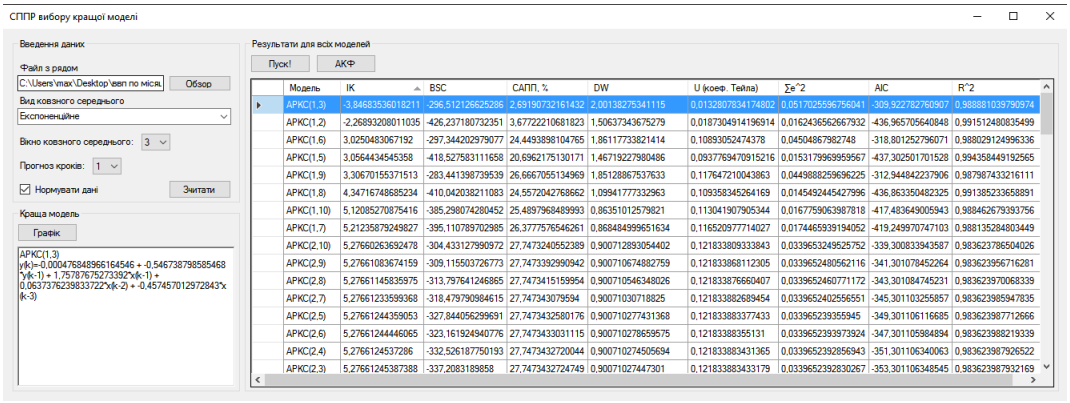


Рисунок 4 – Результат роботи програми на вибірці ВВП України

Як видно з рис. 4 кращою моделлю виявилась АРКС(1,3): $\mathbf{y(k) = -0,000476848966164546 - 0,546738798585468 * y(k-1) + 1,75787675273392 * v(k-1}$. Загальні результати для всіх експериментів можна переглянути в табл.1.

На рис. 5 представлена АКФ для даного процесу.

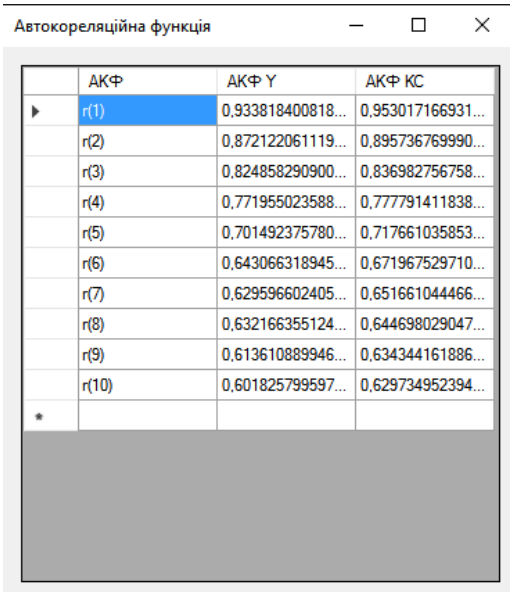


Рисунок 5 – АКФ для процесу ВВП

Виходячи з результатів АКФ програма робила перебір по усім степеням до 10.

На рис. 6 представлений графік кращої моделі та вибірки.

Рисунок 6 – Графік кращої моделі та вибірки для процесу ВВП

Індекс споживчих цін України

Дані взято з сайту Державної служби статистики України [2].

Таблиця 2 Індекс споживчих цін України у 1991 – 2009 рр. (до попереднього місяця, %)

Місяці	1991	1992	1993	1994	1995	1996	1997	1998	1999
Січень	…	385,2	173,2	119,2	121,2	109,4	102,2	101,3	101,5
Лютий	…	115,3	128,8	112,6	118,1	107,4	101,2	100,2	101
Березень	…	112,1	122,1	105,7	111,4	103	100,1	100,2	101
Квітень	…	107,6	123,6	106	105,8	102,4	100,8	101,3	102,3
Травень	…	114,4	127,6	105,2	104,6	100,7	100,8	100	102,4
Червень	…	126,5	171,7	103,9	104,8	100,1	100,1	100	100,1
Липень	…	122,1	137,6	102,1	105,2	100,1	100,1	99,1	99
Серпень	104	108,3	121,7	102,6	104,6	105,7	100	100,2	101
Вересень	104,5	110,6	180,3	107,3	114,2	102	101,2	103,8	101,4
Жовтень	105,9	112,4	166,1	122,6	109,1	101,5	100,9	106,2	101,1
Листопад	116,5	122	145,3	172,3	106,2	101,2	100,9	103	102,9
Грудень	124,6	135,1	190,8	128,4	104,6	100,9	101,4	103,3	104,1

Місяці	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Січень	104,6	101,5	101	101,5	101,4	101,7	101,2	100,5	102,9	102,9
Лютий	103,3	100,6	98,6	101,1	100,4	101	101,8	100,6	102,7	101,5
Березень	102	100,6	99,3	101,1	100,4	101,6	99,7	100,2	103,8	101,4
Квітень	101,7	101,5	101,4	100,7	100,7	100,7	99,6	100	103,1	100,9
Травень	102,1	100,4	99,7	100	100,7	100,6	100,5	100,6	101,3	100,5

Продовження таблиці 2

Місяці	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Червень	103,7	100,6	98,2	100,1	100,7	100,6	100,1	102,2	100,8	101,1

Липень	99,9	98,3	98,5	99,9	100	100,3	100,9	101,4	99,5	99,9
Серпень	100	99,8	99,8	98,3	99,9	100	100	100,6	99,9	99,8
Вересень	102,6	100,4	100,2	100,6	101,3	100,4	102	102,2	101,1	100,8
Жовтень	101,4	100,2	100,7	101,3	102,2	100,9	102,6	102,9	101,7	100,9
Листопад	100,4	100,5	100,7	101,9	101,6	101,2	101,8	102,2	101,5	101,1
Грудень	101,6	101,6	101,4	101,5	102,4	100,9	100,9	102,1	102,1	100,9

Результат роботи програми з файлом, у якому знаходяться дані з таблиці 2 представлений на рис.7.

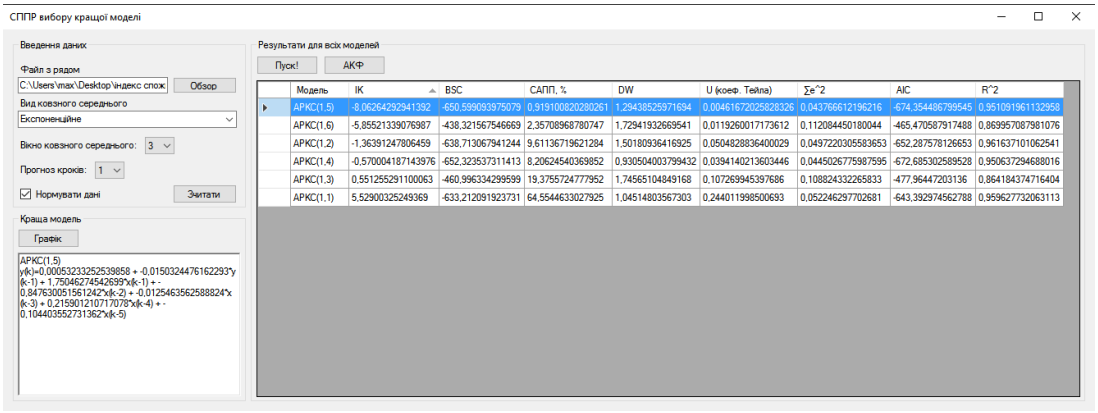


Рисунок 7 - Результат роботи програми на вибірці ІСЦ України

Як видно з рис. 7 кращою моделлю виявилась АРКС(1,5): $\mathbf{y(k) = 0,00053233252539858 - 0,0150324476162293 * y(k-1) + 1,75046274542699 * v(k-1}$. Загальні результати для всіх експериментів можна переглянути в табл.4.

На рис. 8 представлена АКФ для даного процесу.

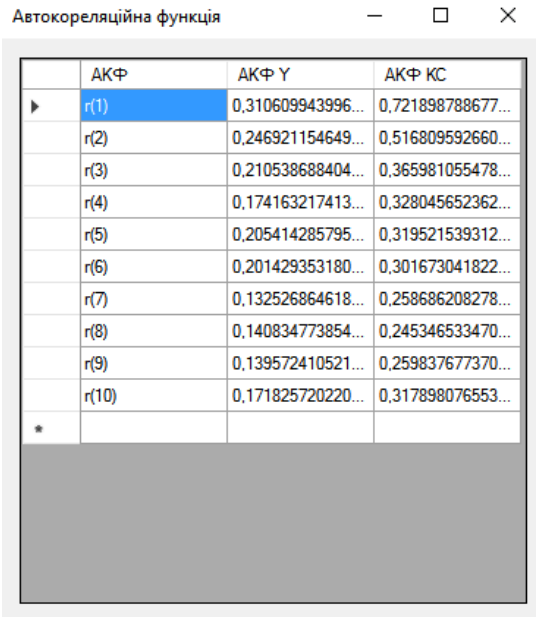


Рисунок 8 – АКФ для процесу ІСЦ

Виходячи з результатів АКФ програма робила перебір для авторегресійної складової лише першої степені, а для складової ковзного середнього до 6 степені.

На рис. 9 представлений графік кращої моделі та вибірки.

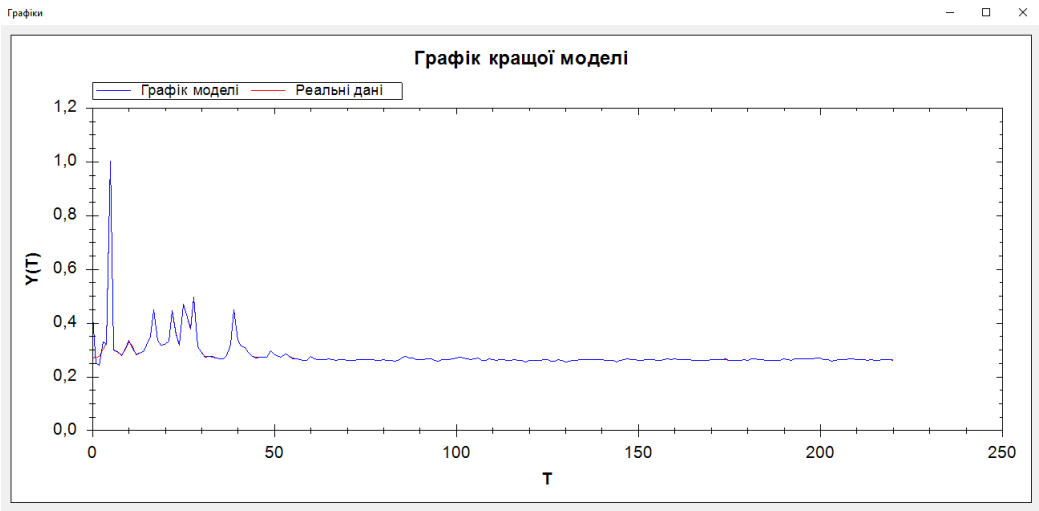


Рисунок 9 – Графік кращої моделі та вибірки для процесу ІСЦ

Грошовий агрегат М3 (загальна грошова маса) в Україні

Дані взято з сайту Державної служби статистики України [2].

Таблиця 3 Грошовий агрегат М3 (загальна грошова маса) в Україні (млн. грн.)

Рік	Місяці					
	1	2	3	4	5	6
1993	32.73	39.51	59.31	67.19	75.73	111.02
1994	477.6	551.7	680.2	780.2	898	1080.3
1995	3264	3581	3931	4242	4700	5306
1996	6594	6891	7184	7331	7484	7672
1997	9359	9335	9795	10215	10410	11102
1998	12074	12238	12960	12971	13305	13458
1999	15185	15366	15923	16663	17496	18579
2000	22048	22970	24211	25097	26032	27098
2001	30816	31638	33026	34092	35157	36953
2002	43619	45032	47345	48389	48813	51195
2003	63200	65304	70114	72509	73977	79034
2004	92643	96049	101151	105104	109435	113971
2005	125726	130942	140107	146495	147896	156339
2006	188844	191283	195304	201214	207370	214136
2007	256226	261348	272464	282359	288185	303016
	7	8	9	10	11	12
1993	156.42	235.88	306.18	356.7	414.84	481.21
1994	1274.8	1575.6	1859.3	2247.8	2827.8	3206.7
1995	5922	6200	6380	6227	6227	6930
1996	7987	7672	7759	7999	8187	9364
1997	11539	12380	12355	11869	11669	12541
1998	13880	13756	14325	14414	14845	15718
1999	18816	19694	20468	20899	21042	22070
2000	28127	29485	28975	28866	29395	32084
2001	37373	38275	39643	40750	41508	45555
2003	80786	83048	86495	86856	88295	95043
2004	117130	121476	130277	126224	125251	125801
2005	159139	164780	171011	174830	180061	194071
2006	221536	226376	234770	238461	244097	261063
2007	316969	328967	348226	354179	365591	396156

Результат роботи програми з файлом, у якому знаходяться дані з таблиці 3 представлений на рис.10.

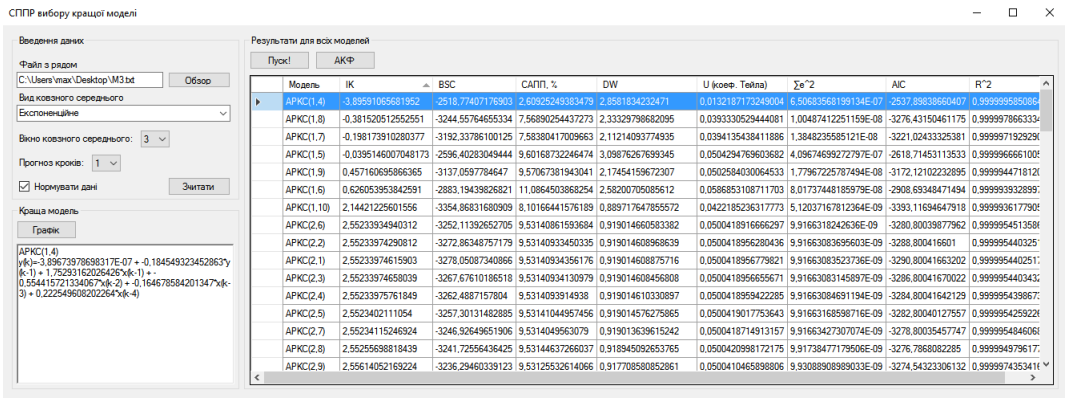


Рисунок 10 - Результат роботи програми на вибірці грошовий агрегат М3 України

Як видно з рис. 10 кращою моделлю виявилась АРКС(1,4): $\mathbf{y(k) = -3,89673978698317E-07 - 0,184549323452863 * y(k-1) + 1,75293162026426 * v(k-}$. Загальні результати для всіх експериментів можна переглянути в табл.4.

На рис. 11 представлена АКФ для даного процесу.

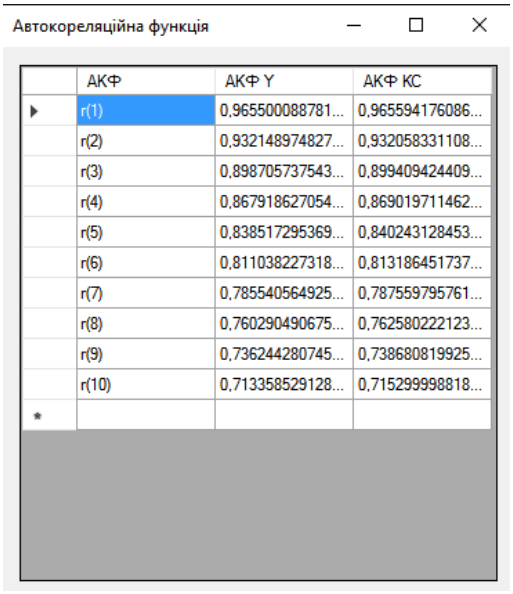


Рисунок 11 - АКФ для процесу грошовий агрегат М3 України

Виходячи з результатів АКФ програма робила перебір по усім степеням до 10.

На рис. 12 представлений графік кращої моделі та вибірки.

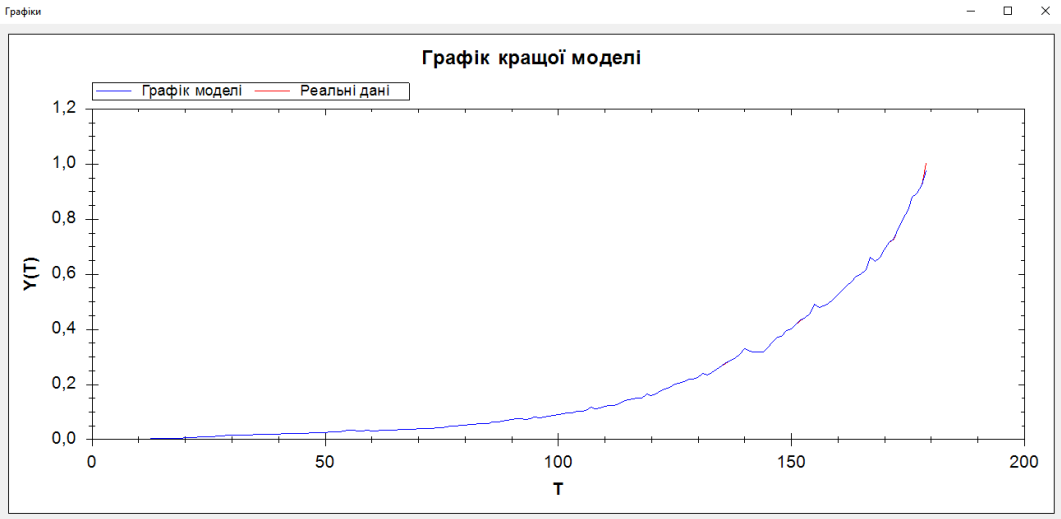


Рисунок 12 – Графік кращої моделі та вибірки для процесу грошовий агрегат М3 України

Загальні результати для всіх експериментів

В табл. 4 представлені по 3 кращі моделі для кожного з розглянутих вище фінансово-економічних процесів та їх оцінки якості моделей та оцінок прогнозів.

Таблиця 3.4 Загальні результати для всіх експериментів

Процес	Модель	ІК	Адекватність моделі			Якість прогнозу	
			R^2	$\sum e^2$	DW	САПП, %	U
ІСЦ	АРКС(1,5)	-8,062	0,951	0,043	1,294	0,919	0,004
	АРКС(1,6)	-5,855	0,869	0,112	1,729	2,357	0,011
	АРКС(1,2)	-1,363	0,961	0,049	1,501	9,611	0,05
ВВП	АРКС(1,3)	-3,846	0,988	0,051	2,001	2,691	0,013
	АРКС(1,2)	-2,268	0,991	0,016	1,506	3,677	0,018
	АРКС(1,6)	3,025	0,988	0,045	1,861	24,449	0,108
Грошовий агрегат М3	АРКС(1,4)	-3,895	0,999	6,506E-07	2,858	2,609	0,013
	АРКС(1,8)	-0,381	0,999	1,004E-08	2,333	7,568	0,039
	АРКС(1,7)	-0,198	0,999	1,384E-08	2,112	7,583	0,039

Висновки

Розроблений інтегральний критерій якості моделі, який поєднав у собі критерії якості моделей та оцінок прогнозів.

Були проведені експерименти з використанням розробленої СППР, змодельовані та спрогнозовані (короткостроковий прогноз) такі реальні економічні процеси, як ВВП України, ІСЦ України та грошовий агрегат М3 (загальна грошова маса) в Україні. За допомогою СППР було автоматично обрано такі кращі моделі для вищезгаданих процесів (всюди використовувалось експоненційне ковзне середнє, що зменшується з віддаленістю від поточного виміру, з вікном 3, прогнозом на один крок та нормуванням даних):

– для ВВП України – модель АРКС(1,3), що має вигляд:

$$y(k) = -0,000476848966164546 - 0,546738798585468 * y(k-1) + 1,75787675273392 * v(k-1$$

, де v – ковзне середнє.

– для ІСЦ України – модель АРКС(1,5), що має вигляд:

$$y(k) = 0,00053233252539858 - 0,0150324476162293 * y(k-1) + 1,75046274542699 * v(k-1$$

– для грошового агрегату М3 (загальнагрошовамаса) в Україні – модель АРКС(1, 4), щомаєвигляд:

$$y(k) = -3,89673978698317E-07 - 0,184549323452863 * y(k-1) + 1,75293$$

Література

1. *Бідюк П.І.* Аналіз часових рядів. / *П.І. Бідюк, В.Д. Романенко, О.Л. Тимощук* – Київ: «Політехніка»,НТУУ “КПІ”, 2013. – 600 с.
2. Державна служба статистики [Електронний ресурс] – Режим доступу: http://www.ukrstat.gov.ua/– Назва з екрана
3. *Бідюк П.І., Половцев О.В.* Аналіз та моделювання економічних процесів перехідного періоду. — Київ: НТУУ «КПІ», 1999. — 230 с.
4. *Лукашин Ю.П.*Адаптивные методы краткосрочного прогнозирования. – М.: Финансы и статистика, 2003. – 414 с.

5. *Chatfield C.* Time series forecasting. – London: Chapman & Hall, 2000. – 267 p.

УДК 519.248
Зубарева П.К

Оцінка впливу депутатських груп у Верховній Раді Українина прийняття колективного рішення за допомогою індексу Банзафа та порядкового α - індексу

Науковий керівник:к.ф-м.н., доцент Каніовська І.Ю.

Вступ

В законодавчих органах депутати, як правило, об’єднуються у фракції (частіше всього на основі партійної приналежності). Структуризація парламенту дозволяє зробити його роботу більш організованою і ефективною, прискорити та полегшити процес прийняття рішень. Той факт, що фракції, утворені з депутатів, обраних на загальнодержавних виборах, є головними акторами в парламентах, робить представницьку демократію принципово можливою. Прийняття рішення залежить від результату голосування депутатів: рішення буде ухвалено тільки в тому випадку, коли кількість голосів на його користь перевищить деяку заздалегідь встановлену квоту. У ситуації, коли в законодавчому органі представлені три і більше фракції, можлива ситуація, при якій жодна з них не володіє достатньою кількістю голосів, щоб подолати квоту. У таких випадках фракції змушені об'єднуватися один з одним - утворювати коаліції. Академічна політологія значну увагу приділяє питанням, пов'язаним з механізмами створення коаліцій та закономірностями їх існування.

Спостереження за роботою законодавчих органів робить очевидним те, що фракції в парламентах мають різний вплив в процесі прийняття законів. Визначення того, якою мірою та чи інша фракція впливає на прийняття рішень, дозволить спрогнозувати їх характер. Знання про розподіл впливу між фракціями в парламенті дозволить зробити певні висновки про особливості політичної системи в цілому. Складання загальної картини розподілу впливу в парламенті дозволить обчислювати найбільш ймовірні стратегії фракцій, що, наприклад, надзвичайно важливо в передвиборчий період. Вивчення розподілу впливу фракцій в парламентах є актуальним не тільки з точки зору академічної, а й з точки зору прикладної політології.

Обрання Верховної Ради України як об'єкту дослідження пов'язано з особливостями сучасної політичної історії України. Активна діяльність

українського парламенту протягом останніх двадцяти років, його змінне положення як гілки влади, зміни політичних еліт, поляризація українського суспільства роблять політичне життя України насиченим, а українськийпарламент привабливим для вивчення.

Питання вимірювання впливу акторів в законодавчих органах є одним з найважливіших і вивчається протягом тривалого часу (див., наприклад, [1, 2, 3, 4, 5]). На перший погляд, вплив фракції в парламенті безпосередньо залежить від того, яку частку місць вона займає. Однак, як буде показано в даній статті, ця думка є хибною: впливовість фракції не завжди залежить від кількості її місць в парламенті. Для вимірювання впливу гравців в органах прийняття колективного рішення були розроблені спеціальні індекси (напр., Індекси Банзафа, Шеплі-Шубіка, Дігена-Пакела, [5]).

Дослідження присвячене аналізу вимірювання впливу груп (фракцій) на прийняття колективного рішення та застосуванню отриманих результатів для прогнозування характеру подальших політичних процесів у країні.

Постановка задачі

Нехай на вхід подається чисельний склад кожної голосуючої групи (фракції). Ми хочемо визначити, яка з груп (фракцій) – «сильніша», тобто має найбільше впливу серед інших груп на прийняття колективного рішення. За основу голосування береться квота, побудована за правилом більшості. Ми також знаємо, що для подолання квоти, групам вкрай необхідно об’єднатись у коаліції. Проте такий процес утворення коаліцій не є випадковим, а найчастіше грунтується на взаємних бажаннях груп до об’єднання.

Визначення міри вливу фракцій (груп) на прийняття рішення дозволить спрогнозувати найбільш імовірні політичні, економічні, соціальні та інші процеси в державі і на міжнародному рівні; визначити характер політичної системи в цілому; обчислити найкращі стратегії фракцій у передвиборчий період.

Визначення індексу Банзафа

Коаліція називається виграшною, якщо кількості голосів в ній досить для подолання заданої квоти. Якщо після виключення деякої фракції коаліція перестає бути виграшною, то така фракція називається ключовою. Фракція тим більш впливовіша, чим в більшій кількості виграшних коаліцій вона є ключовою. Саме на підставі такої логіки побудований індекс впливу Банзафа[6]- величина, що

дозволяє оцінити, наскільки впливовою є та чи інша фракція в парламенті. Індекс Банзафа $\beta(i)$ для фракціїi обчислюється як наступне відношення:

$$\beta(i) = {}^{b_i}\!/_{\sum_{j=1}^{k} b_i},$$

де b_i - число коаліцій, в яких дана фракція i є ключовою, k - кількість всіх фракцій в парламенті.

Індекс Банзафа, поряд з іншими індексами впливу (Шеплі - Шубіка, Дігена - Пакела, [6; 3]) використовується для оцінки впливовості гравців в органах прийняття колективних рішень, в тому числі й фракцій в парламентах.Надалі, на практичному прикладі, буде побудовано індекс Банзафа для фракцій Верховної Ради України 8-госкликання. Передбачається, що фракції і групи не мають переваг у створенні коаліцій, тобто всі коаліції розглядаються як рівноймовірні, а депутати однієї фракції або депутатської групи голосують одноголосно. Індекси розраховані для простої та конституційної більшості (квота рівна$0{,}5n + 1$ та$\frac{2}{3n}$ відповідно, де n - число депутатів в парламенті).

Однак припущення про те, що всі коаліції рівноймовірні, часто не відповідає дійсності, і індекс Банзафа виявляється неспроможним відображати реальний розподіл впливу в виборчих органах. В [7] було запропоновано узагальнення індексу Банзафа -α -індекс, який дозволяє врахувати різного роду переваги фракцій в утворенні коаліцій. Побудова узагальнення - порядкового (ординального) індексу - грунтується на тому, що переваги фракцій в утворенні коаліцій розглядаються як відносини порядку.

Визначення порядкового α -індексу

Для оцінки впливу фракцій та груп в українському парламенті з урахуванням їх переваг у формуванні коаліцій пропонується використовувати модифікацію індексу Банзафа - порядковий α -індекс. При побудові порядкового α -індексу передбачається, що для кожної фракції в парламенті можна побудувати впорядкування інших фракцій відповідно до її бажанням вступати з ними в коаліцію. Тобто для фракції i на декартовому добутку$(P\{i\} \times P\{i\})$ можна визначити відношення виду $R_i = (j, k)$, яке показує, що фракція i віддасть перевагу вступити в коаліцію скоріше з фракцією j, ніж з фракцією k.

Відношення R_i є відношенням лінійного порядку, отже можна визначити ранг p_{ij} фракції j в R_i, що дорівнює кількості фракцій, з якими фракція $\mathbf{i}$ буде вступати в коаліцію менш охоче, ніж з фракцією j. Найбільш краща для вступу в коаліцію фракція матиме ранг (ціле невід'ємне число) $p_{ij} = |P| - 1$, де $|P|$ - кількість фракцій у парламенті. Таким чином, p_{ij} буде відображувати бажання фракції $\mathbf{i}$ об'єднатися з фракцією j в коаліцію, причому в загальному випадку $p_{ij} \neq p_{ji}$, що дозволяє врахувати несиметричність переваг фракцій.

Після того як вподобання фракцій представлені у вигляді чисел, ці вподобання можна врахувати при побудові індексу впливу. Для цього в [7] пропонується використовувати функції інтенсивності зв'язку (intensity functions).

Нехай ω - виграшна коаліція. Будемо позначати функцію інтенсивності зв'язку фракції i з коаліцією ω як $f(i, \omega)$. В [7] запропоновано кілька способів її побудови. Розглянемо три з них.

1. Функція інтенсивності прямих вподобань фракції^i :

$$f^+(i, \omega) = {\Sigma_{j \in \omega} p_{ij}} \Big/ {|\omega|}$$

При побудові даного типу функції інтенсивності зв'язку враховується виключно інтенсивність переваг фракції i в формуванні коаліції з іншими учасниками виграшної коаліції. Вона розраховується як сума рангів p_{ij} кожної фракції j з виграшної коаліції, нормована на кількість фракцій у виграшній коаліції.

2. Функція інтенсивності зворотних переваг по відношенню до фракції i :

$$f^-(i, \omega) = {\Sigma_{j \in \omega} p_{ji}} \Big/ {|\omega|}$$

При побудові функції зворотних переваг враховується тільки інтенсивність переваг фракцій з виграшної коаліцій по відношенню до фракції i .

3. Усереднена функція інтенсивності зв'язку:

$$f(i, \omega) = \frac{f^+(i, \omega) + f^-(i, \omega)}{2}$$

Усереднена функція інтенсивності зв'язку фракції i з виграшною коаліцією ω розраховується як середнє арифметичне функцій прямих і зворотних переваг.

Нехай фракція i є ключовою в виграшній коаліції ω. Визначимо θ_i як абсолютний вплив фракції i, який дорівнює загальній кількості значень функції інтенсивності зв'язку фракції i з усіма виграшними коаліціями ω, в яких вона є ключовою:

$$\theta_i = \sum_{\omega,\ i-\text{ключова в } \omega} f(i, \omega)$$

Нормуючи θ_i на $\sum_j \theta_i$ (суму значень абсолютного впливу всіх фракцій в виграшній коаліції), отримаємо відносний вплив фракції i. Це і буде порядковий α-індекс, індекс впливу фракції i, в якому враховані переваги фракцій в утворенні коаліцій:

$$\alpha_{(i)} = \frac{\theta_i}{\Sigma_j \theta_i}$$

Використовуючи в якості опції інтенсивності зв'язку різні її варіації, можна побудувати відповідні різновиди α-індексу.

Як уже згадувалося, α-індекс є узагальненням індексу Банзафа. І якщо побудова індексу Банзафа заснована на знаходженні кількості виграшних коаліцій, в яких фракція є ключовою, тобто за кожну коаліцію, яку дана фракція робить виграшною, вона «отримує» одиницю, то в α-індексі замість одиниці за кожну таку коаліцію фракція «отримує» деяке значення, відповідне силі зв'язку даної фракції з коаліцією.

Оцінка впливу депутатських груп у Верховній Раді України 8-го скликання (2016 рік) за допомогою індексу Банзафа

Для побудови розподілу влади Верховної Ради України 8-го скликання було взято зріз її чисельного складу на момент травня 2016року[8].

Наразі в парламенті існують 9 депутатських фракцій і груп загальною кількістю в 418 депутатів. Найбільшою фракцією, що за часткою місць займає майже 35%, є Блок Петра Порошенка з чисельним складом у 142 депутати. Другою за чисельністю є Народний Фронт, очолюваний Арсенієм Яценюком.

Частка місць цієї фракції вже на 20% відсотків меньше частки місць попередньої і становить 19.5% від загального розподілу місць. Опозиційний блок та Позафракційні групи майже однакові за складом, проте поділяють між собою третю та четверту сходинки у кількості місць, - 10.5% та 10% відповідно. Надалі детальніше про кількісний склад у Таблиці 1.Було здійснено перейменування фракцій для простоти подальшого використання.

Індекс Банзафа був розрахований в припущенні, що депутати, які входять до фракцій і груп, голосують одноголосно. Для 9-ти фракцій з математичної точки зору існує 512 коаліцій. У роботі було розглянуто 502 коаліції, виключивши порожню коаліцію та коаліцію з одним гравцем. Оскільки ми рахуємо індекс для простої та конституційної більшості, кількість виграшних коаліцій буде відрізнятись. У першому випадку (випадок простої більшості, де квота становить 226 голосів) кількість виграшних коаліцій становить 231, у другому (випадок конституційної більшості, де квота вже становить 300 голосів) – лише 96. На щастя, парламент України не має у своєму складі фракцій – фіктивних гравців, тих, що жодного разу не виступали критичними у виграшній коаліції.

Таблиця 1 – Чисельний склад та частки місць депутатських груп у Верховній Раді України 8-го скликання (травень 2016 року)

	Назва	Перейменування	Кількість	Частка
1	«Блок Петра Порошенка»	A	144	34.45%
2	«Народний фронт»	B	81	19.38%
3	«Опозиційний блок»	C	44	10.53%
4	«Позафракційні»	D	41	9.81%
5	«Самопоміч»	E	26	6.22%
6	«Партія Відродження»	F	23	5.50%
7	«Радикальна партія Ляшко»	G	21	5.02%
8	«Батьківщина»	H	19	4.55%
9	«Воля народу»	I	19	4.55%
	Всього		418	

Для побудови абсолютного індексу Банзафа для кожної фракції ми знаходимо кількість разів, коли вона виступає критичним гравцем у всіх виграшних коаліціях. Відносну міру Банзафа отримуємо шляхом нормування абсолютної. Нижче - Таблиця 2, 3 містять значення індексу для простої та контитуційної більшості, а Рисунок 1, 2 відображують різницю між індексом

Банзафа та часткою місць у парламенті по всіх фракціях для Верховної Ради України 8-го скликання. Обговорення наведених результатів представлено у подальших пунктах статті.

Таблиця 2 – Індекс Банзафа депутатських груп у Верховній Раді України 8-го скликання (травень 2016 року) за простою більшістю

	Назва	Перейменування	Індекс Банзафа
1	«Блок Петра Порошенка»	A	44.27%
2	«Народний фронт»	B	12.81%
3	«Опозиційний блок»	C	10.11%
4	«Позафракційні»	D	8.76%
5	«Самопоміч»	E	5.62%
6	«Партія Відродження»	F	5.62%
7	«Радикальна партія Ляшко»	G	4.27%
8	«Батьківщина»	H	4.27%
9	«Воля народу»	I	4.27%

Таблиця 3 – Індекс Банзафа депутатських груп у Верховній Раді України 8-го скликання (травень 2016 року) за конституційною більшістю

	Назва	Перейменування	Індекс Банзафа
1	«Блок Петра Порошенка»	A	44.27%
2	«Народний фронт»	B	12.81%
3	«Опозиційний блок»	C	10.11%
4	«Позафракційні»	D	8.76%
5	«Самопоміч»	E	5.62%
6	«Партія Відродження»	F	5.62%
7	«Радикальна партія Ляшко»	G	4.27%
8	«Батьківщина»	H	4.27%
9	«Воля народу»	I	4.27%

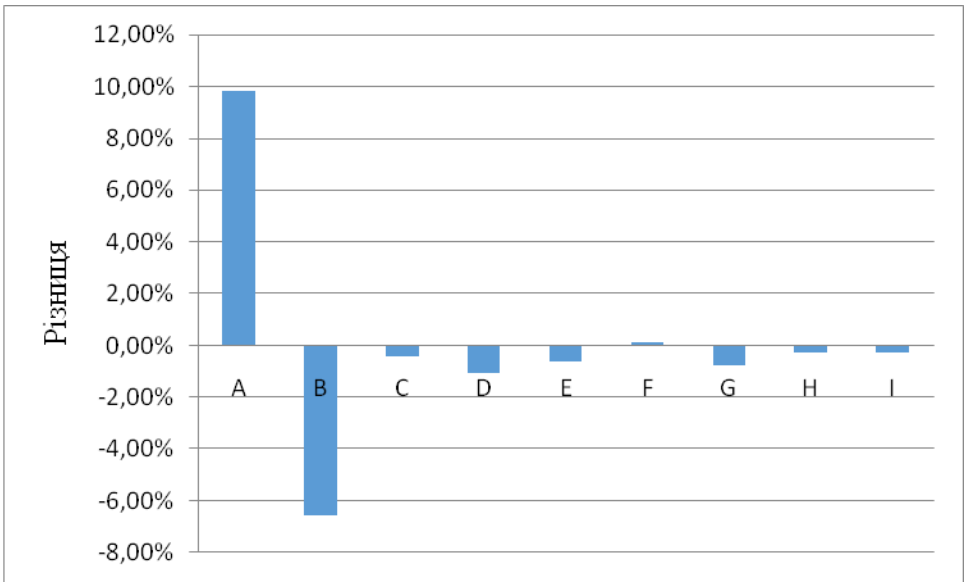


Рисунок 1 – Різниця між індексом Банзафа та часткою місць (проста більшість)

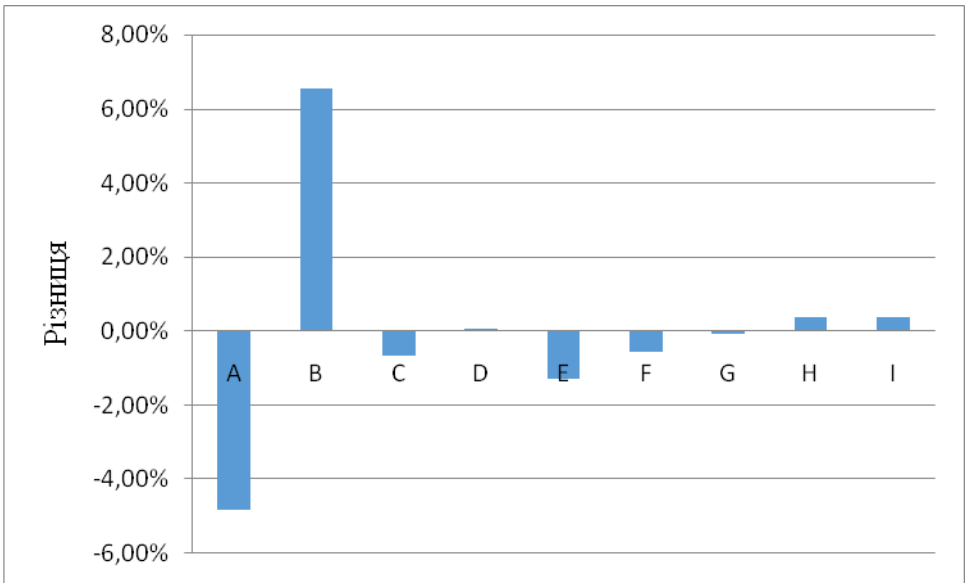


Рисунок 2 – Різниця між індексом Банзафа та часткою місць (конституційна більшість)

Оцінка впливу депутатських груп у Верховній Раді України 8-го скликання (2016 рік) за допомогою порядкового α -індексу

У ВРУ 8-го скликання існує поляризація партій на основі їх ідеологій. Також спостерігається чітке структурування за принципом «пропрезидентські - антипрезидентські сили». У Таблиці 4 наведено ідеологічний спектр. Насправді, на сьогоднійшній день однозначно достовірно визначити ідеологічні принципи партій досить складно, оскільки їх трансформація відбувається швидко. Більш

достовірні результати можна було б отримати, лише дивлячись на ВРУ 8-го скликання в ретроспективі, тобто – у майбутньому.

За основу такої структуризації (Таблиця 4) бралися первинні офіційні ідеології партій, а також результати спостережень за характером прийняття рішень партіями під час пленарних засідань у ВРУ.

Таблиця 4 - Положення фракцій і груп у ВРУ 8-го скликання на ідеологічному спектрі:

Пропрезидентські сили			Антипрезидентські сили	
Націоналісти, радикали	Консервативні - лібералісти	Ліберали	Консерватори	Радикали
«Батьківщина», «Самопоміч», РПЛ	БПП, «Народний фронт»	Позафракційні	«Воля народу», «Відродження»	«Опозиційний блок»

Визначення положення фракцій і груп на ідеологічному спектрі дозволяє кожній з них присвоїти ранг r_i. Фракціям і групам, які займають однакове становище на спектрі, будемо привласнювати однаковий ранг. Нехай значення r_i у націоналістів-радикалів пропрезидентських сил дорівнює 1, а у радикалів антипрезидентських сил - 6. Ліберали матимуть ранг, рівний 3, а консерватори антипрезидентських сил – вже 5 (Таблиця 5).

Таблиця 5- Позиції фракцій і груп у ВРУ 8-го скликання

Назва	Перейменування	Ранг
Батьківщина	H	1
Самопоміч	E	1
Радикальна партія Ляшко	G	1
Блок Петра Порошенка	A	2
Народний фронт	B	2
Позафракційні	D	3
Воля народу	I	5
Партія Відродження	F	5
Опозиційний блок	C	6

На підставі наданих кожній фракції (групі) рангів можна побудувати матрицю вподобань $\|p_{ij}\|$.. Введемо відстань між фракціями - міру віддаленості позицій двох фракцій μ_{ij}, $\mu_{ij} = |r_i - r_j|$. Чим більше відстань між двома

фракціями (групами) на осі, тим менш імовірно, що вони будуть прагнути утворювати коаліцію, значить значення p_{ij}, що відображує бажання фракції утворити коаліцію, буде тим менше, чим міра віддаленості буде більше. Будемо вважати, що фракції і групи не утворюють коаліцію, якщо відстань між ними перевищує значення 3, тобто при $\mu_{ij} > 3$ значення p_{ij} дорівнює 0. Якщо міра віддаленості позицій фракцій приймає значення 2, відповідне значення p_{ij} дорівнюватиме 1, якщо $\mu_{ij} = 1 p_{ij}$ буде приймати значення 2 і найбільш близьким фракціям $\mu_{ij} = 0$ з відповідатиме $p_{ij} = 3$.

Саме на основі такої технології побудови матриці вподобань було прийнято рішення пропустити ранг = 4, щоб уникнути ситуації, коли пропрезидентські та антипрезидентські партії (крім Позафракційні) матимуть імовірність утворити коаліції.

Таблиця 6 - Матриця вподобань ВРУ 8-го скликання

	H	E	G	A	B	D	I	F	C
H		3	3	2	2	1	0	0	0
E	3		3	2	2	1	0	0	0
G	3	3		2	2	1	0	0	0
A	2	2	2		3	2	0	0	0
B	2	2	2	3		2	0	0	0
D	1	1	1	2	2		1	1	0
I	0	0	0	0	0	1		3	2
F	0	0	0	0	0	1	3		2
C	0	0	0	0	0	0	2	2	

У Таблиці 7, 8 представлені результати обчислень порядкового α -індексу на основі функції інтенсивності прямих переваг для прийняття законів на основі простої та конституційної більшості відповідно.

Таблиця 7–Порядковий α -індекс депутатських груп у Верховній Раді України 8-го скликання (травень 2016 року) за простою більшістю

Назва	Перейменування	Alpha	Delta(Alpha-Banzhaf)
Батьківщина	H	4.76%	0.50%
Самопоміч	E	6.48%	0.86%
Радикальна партія Ляшко	G	4.76%	0.50%
Блок Петра Порошенка	A	50.07%	5.80%
Народний фронт	B	15.99%	3.18%

Позафракційні	D	8.10%	-0.67%
Воля народу	I	2.57%	-1.70%
Партія Відродження	F	3.36%	-2.26%
Опозиційний блок	C	3.90%	-6.21%

Таблиця 8–Порядковий α -індекс депутатських груп у Верховній Раді України 8-го скликання (травень 2016 року) за конституційною більшістю

Назва	Перейменування	Alpha	Delta(Alpha-Banzhaf)
Батьківщина	H	5.66%	0.72%
Самопоміч	E	5.66%	0.72%
Радикальна партія Ляшко	G	5.66%	0.72%
Блок Петра Порошенка	A	34.47%	4.84%
Народний фронт	B	30.65%	4.72%
Позафракційні	D	10.08%	0.21%
Воля народу	I	2.35%	-2.59%
Партія Відродження	F	2.35%	-2.59%
Опозиційний блок	C	3.12%	-6.75%

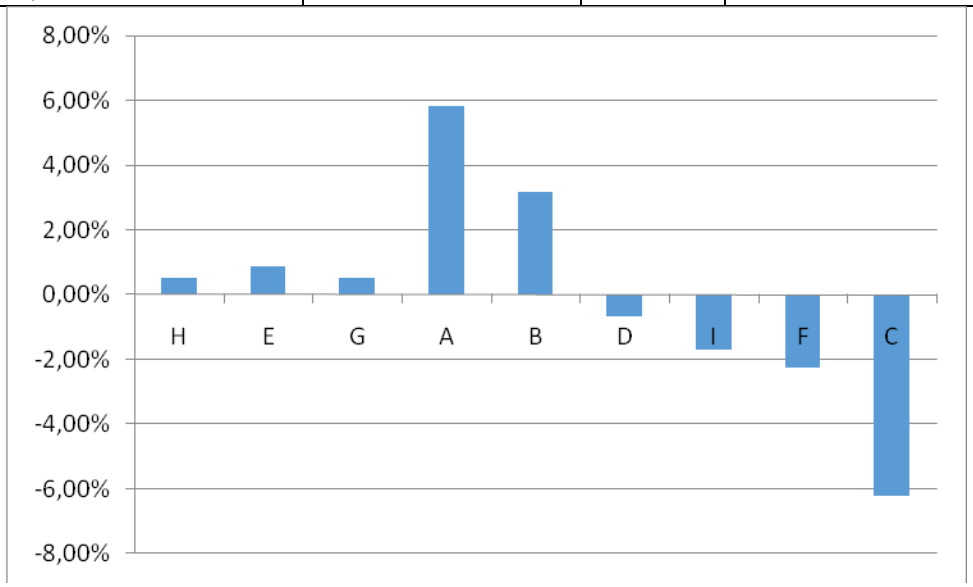


Рисунок 3 – Різниця між порядковим α -індексом та індексом Банзафа (проста більшість)

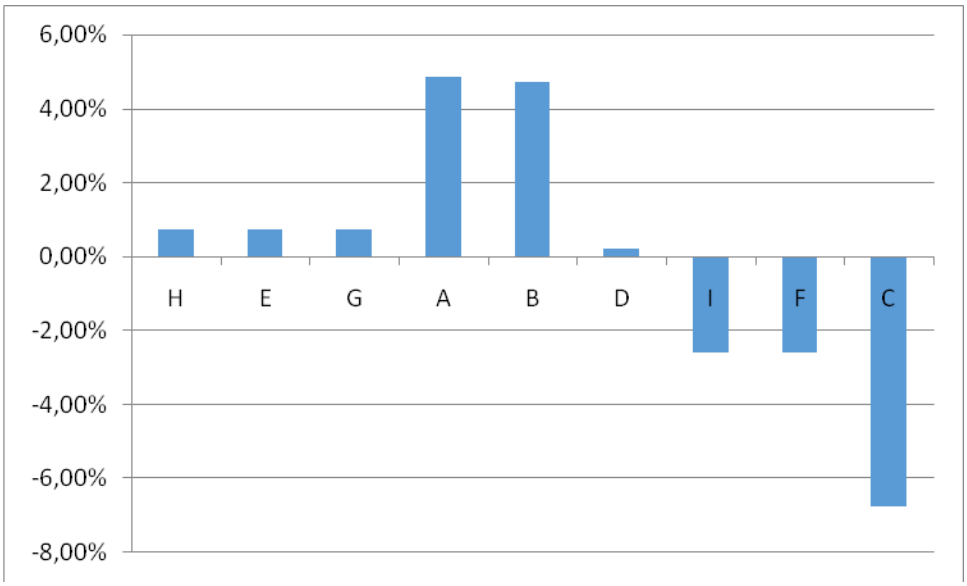


Рисунок 4 – Різниця між порядковим α -індексом та індексом Банзафа (конституційна більшість)

Аналіз отриманих результатів

Картина отриманих результатів відрізняється від «офіційної» ситуації в парламенті. Виявляється, що деякі фракції, займаючи певний відсоток місць, насправді мають дещо інший відсоток впливу на прийняття коллективного рішення.

Для початку розглянемо результати, отримані за допомогою індексу Банзафа. Наприклад, «Блок Петра Порошенка», що займає всього третину місць у парламенті, насправді впливає на прийняття рішення за простою більшістю з силою у майже 45%, в той час як «Народний фронт» – навпаки втрачає свої позиції, що дорівнюють 7% впливу. Цікава ситуація склалась з останнім, рахуючи вплив партії Арсенія Яценюка на прийняття рішень за конституційною більшістю, виявилось, що вона одна з найбільших, яка значно посилює свій вплив.

Результати, отримані за допомогою порядкового α -індексу, як відомо з теорії, більш точніші. Тут вплив «Блоку Петра Порошенка» виявився значнішим для рішень за простою більшістю, та становить трохи більше 50%. А от позиції впливу на прийняття рішення за конституційною більшістю значно покращила лише партія «Народний фронт». Різниця з їх чисельним складом складає 11%. Варто відмітити, що «Блок Петра Порошенка» у обрахунках порядкового α -індексу за конституційною більшістю у всіх виграшних коаліціях виступає

критичним гравцем. Це означає, що жодне конституційне рішення не може бути прийняте без одноголосної підтримки депутатів з «Блоку Петра Порошенка».

Висновки

За період з 1990 по 2016 р Україна була проголошена незалежною державою, була прийнята нова Конституція, кілька разів змінювалася виборча система і двічі - форма правління, змінювалася політична еліта. Верховна Рада України за час свого існування пройшла шлях від парламенту з безліччю дрібних груп, утворених за непартійним ознаками, до парламенту з невеликою кількістю великих фракцій, здатних формувати відносно стійкі коаліції. Однак під час усіх скликань ситуація в парламенті була нестабільною, взаємодія більшості та опозиції, які формувалися в деяких скликаннях, складна, що часто ускладнювало прийняття рішень.

В статті показано, що вплив фракцій і груп ВРУ 8-го скликання, оцінений за допомогою індексу Банзафа, не завжди залежить від їх частки місць у парламенті. Яскравий тому приклад – «Блок Петра Порошенка». Оцінки розподілу впливу фракцій та груп, проведені за допомогою α-індексу, – більш точніші, оскільки враховують перевагифракцій щодо утворення коаліцій. Воничасто значно відрізняються від побудованих індексів Банзафа. Показано, що фракції і групи, які займають центричне положення, збільшують свій вплив, в той час як радикальні фракції і групи зменшують його. Такий результат було отримано також і під час досліджень кількісного складу та коаліційних переваг ВРУ 2-го та 3-го скликань. На основі цього, можна стверджувати, що більшість конституційних та загальних рішень у парламенті будуть відбуватися під впливом консервативно-ліберальних пропрезидентських партій – «Блок Петра Порошенка» та «Народний фронт», допоки ситуація з розподілом місць у парламенті не зміниться.

Література:

1. *Aleskerov F., Holler M., Kamalova R.* Power Distribution in the Weimar Reichstag in 1919 – 1933. The University – Higher School of Economics. M.: Publishing House of the University Higher School of Economics, 2010. 54 p.

2. *Costello R., Thomson R.* The distribution of power among EU institutions: who wins under codecision and why? // *Journal of European Public Policy*. 2013. Vol. 20. P. 1025-1039.

3. *Eligius. H*Coalition Formation: The Role of Procedure and Policy Flexibility // *Group Decision &Negotiation*. 2013. Vol. 22. P. 407-427.

4. *Leech D.* Power indices in large voting bodies // *Public Choice*. 2013. Vol. 155. P. 61-79.

5. *Алескеров Ф. Т., Благовещенский Н., Сатаров Г., Соколова А. В., Якуба В.И.* Влияние и структурная устойчивость в российском парламенте (1905-1917 и 1993-2005 гг.). М.: Физматлит, 2007

6. *Banzhaf John F./ Weighted voting doesn't work: A mathematical analysis*– N.:, Nertiper, 1965. – 317–343pp

7. *Acemoglu D., Egorov G., Sonin K.* Coalition Formation in Non-Democracies // *Review of Economic Studies*. 2008. Vol. 75. P. 987-1009

8. Схема розміщення депутатських фракцій і груп у сесійній залі[Электронный ресурс]:http://w1.c1.rada.gov.ua/pls/radan_gs09/ns_zal_frack

УДК 004.6

РЕАЛІЗАЦІЯ МЕТОДИКИ ПОБУДОВИ МОДЕЛЕЙ ЧАСОВИХ РЯДІВ

Б.Р. Дудка, П.І. Бідюк

На сьогодні існує широке коло методів прогнозування, які ґрунтуються на різноманітних аналітичних процедурах, чітких та нечітких логічних правилах і раціональному експертному мисленні. Однак, у багатьох випадках вони не дають можливості досягти бажаного результату стосовно якості оцінок прогнозів. В зв'язку з цим виникає проблема значного і прискореного підвищення якості оцінок коротко- та середньострокових прогнозів. Розв'язання задач ефективного прогнозування на новому якісному рівні вимагає застосування до досліджуваних процесів сучасних методів системного аналізу, коректного використання методів математичного моделювання процесів довільної природи на основі досягнень теорії оцінювання і статистичного аналізу даних. Деякі можливості стосовно розв'язання задачі адаптивного прогнозування розглядаються у відомих роботах [1, 2, 3]. Однак, методи, представлені в цих роботах, не грунтуються на системному підході до розв'язання задач прогнозування і не дають відповіді на основне запитання: як організувати процес обробки даних таким чином, щоб отримати кращі оцінки прогнозів в умовах наявності невизначеностей структурного, параметричного і статистичного характеру. Подібні невизначеності можуть бути зумовлені прихованою нестаціонарністю процесу, розвиток якого прогнозується, пропусками даних, неякісними зашумленими даними та екстремальними значеннями. Ефективні методи адаптивного прогнозування за допомогою фільтра Калмана представлені в роботі [4]. Для адаптації алгоритму оцінювання та прогнозування стану процесу використовують обчислені в реальному часі оцінки статистичних характеристик збурень стану і шумів вимірів.

Постановка задачі. Ставиться задача розробки та реалізації методики аналізу нестаціонарних процесів довільної природи на основі підходів та методів системного аналізу, які передбачають ієрархічний аналіз процесів моделювання та прогнозування, врахування невизначеностей структурного параметричного і статистичного

характеру, адаптування моделей до змін у процесах та застосування альтернативних методів оцінювання з метою пошуку кращих структур моделей за допомогою множини статистичних критеріїв якості моделей. Запропонувати нові обчислювальні схеми побудови прогнозуючих систем із зворотним зв'язком на основі використання множини статистичних параметрів якості оцінок прогнозів і застосувати створену концепцію адаптивного прогнозування до реальних процесів.

Концепція побудови адаптивної прогнозуючої системи

Спрощена концептуальна схема процесу моделювання, прогнозування та керування (керування – логічне завершення двох попередніх етапів) наведена на рис. 1. Розглянемо докладніше кожний з етапів аналізу даних. Створення системи адаптивного прогнозування починається з вибору процесу, аналізу його поточного стану, існуючих моделей та підходів до прогнозування його розвитку. Аналіз спеціальних літературних джерел може суттєво допомогти у встановленні факту існування моделей для опису поведінки (статики чи динаміки) вибраного процесу. Це можуть бути математичні моделі у вигляді систем рівнянь, закони розподілу вхідних та вихідних величин (статистичні моделі) або логічні моделі у вигляді множин правил, які характеризують взаємодію входів і виходів процесу керування. В останні три десятиліття набувають великої популярності ймовірнісні методи і моделі різноманітних структур і моделі у вигляді правил нечіткої логіки, які мають відносно хороше наближення до характеру мислення експерта. Ймовірнісні моделі представляють собою хороше (логічне) доповнення до моделей регресійного типу. Вибір типу та структури моделі відіграє суттєву роль для реалізації подальших етапів створення прогнозуючої системи.

Так, модель, створена на основі теоретичних уявлень та закономірностей стосовно конкретного процесу, може потребувати лише деякого уточнення її параметрів за допомогою статистичних даних.

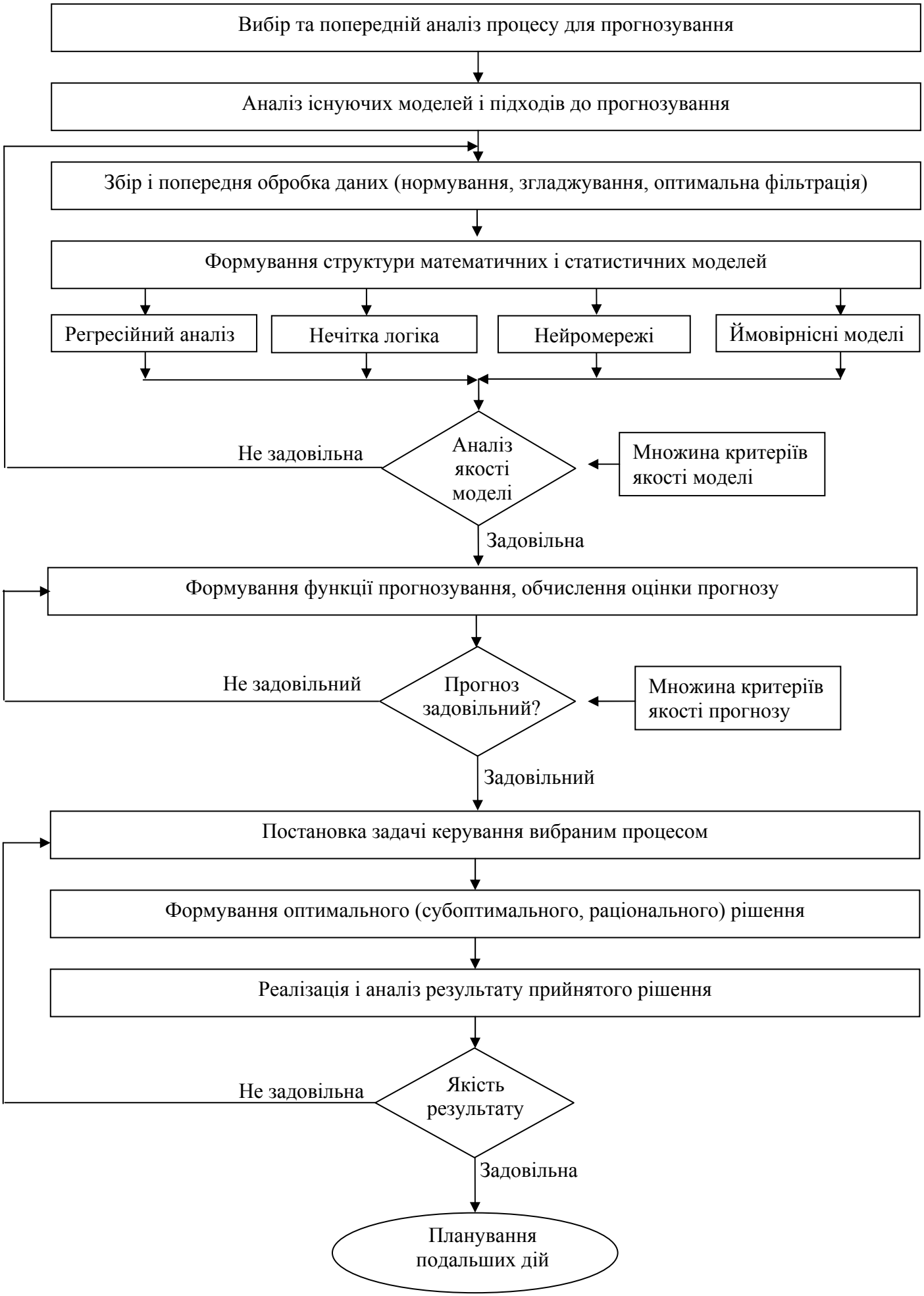


Рисунок 1 – Структура адаптивної прогнозуючої системи

А модель, яка повністю грунтується на статистичних дослідженнях, може потребувати значно більших об’ємів інформації та часу для її побудови. Огляд літературних джерел може бути також корисним з точки зору вибору методів адаптивного оцінювання параметрів моделі. Оскільки кожний метод має свої особливості та межі застосування, то бажано знати і врахувати ці особливості до його застосування на практиці. Наприклад, звичайним методом найменших квадратів (МНК) можна скористатись для оцінювання параметрів лінійних і псевдолінійних моделей за умови нормального розподілу збурення процесу, а узагальнений МНК не має такого обмеження, але потребує іншої додаткової інформації.

Практика створення прогнозуючих систем для процесів довільної природи свідчить про те, що готові для практичного використання моделі зустрічаються дуже рідко. Навіть існуючі апробовані моделі потребують корегування їх структури та/або параметрів з метою адаптування до конкретних умов. Тому у більшості випадків необхідно будувати нову модель високого ступеня адекватності на основі наявних статистичних даних. Якість даних відіграє надзвичайно важливу роль при побудові математичної моделі, а тому при зборі даних необхідно керуватись відомими вимогами стосовно їх інформативності, синхронності та коректності [4].

Попередня обробка даних необхідна для приведення їх до форми, яка забезпечить можливості коректного застосування методів оцінювання параметрів моделі та отримання їх статистично значущих оцінок. Так, досить часто необхідно заповнювати пропуски даних, корегувати значні імпульсні (екстремальні) значення, нормувати значення у заданих межах та логарифмувати великі значення. Ключовим моментом обробки зашумлених даних є фільтрація шумових складових за допомогою цифрових або оптимальних фільтрів. Наприклад, ефективне моделювання і прогнозування процесів ціноутворення на біржі неможливе без згладжування високочастотних складових вимірів.

На основі даних, коректно підготовлених для подальших обчислювальних процедур, оцінюють структури та параметри

можливих математичних моделей процесів, вибраних для прогнозування і керування. Вибір (оцінювання) структури моделі – ключовий момент процесу побудови математичної моделі. Нагадаємо, що структура моделі включає такі елементи: (1) *вимірність* (кількість рівнянь, що утворюють модель); (2) *порядок моделі* – максимальний порядок диференціальних або різницевих рівнянь, які входять в модель; (3) *нелінійність* та її тип (нелінійності стосовно змінних або параметрів); (4) *час затримки* (лаг) виходу процесу відносно входу та його оцінка; (5) *зовнішнє збурення* процесу та його тип (детерміноване або випадкове); (6) *початкові умови та обмеження* на змінні і параметри моделі. Як правило, для одного процесу оцінюють декілька моделей-кандидатів, а потім вибирають з них кращу модель за допомогою множини відомих статистичних параметрів якості моделі.

Більшість вимірів процесів у техніці, економіці та фінансах мають, як правило, детерміновану та випадкову складові. Тому в якості статистичних моделей будемо розглядати моделі процесів у вигляді розподілів випадкових величин [5]. Обґрунтований вибір типу розподілу та отримання коректних оцінок його параметрів представляє собою процес побудови статистичної моделі процесу.

Після побудови моделі її необхідно перевірити на можливість застосування до розв’язання задачі прогнозування. На сьогодні існує широкий спектр методів моделювання та обчислення оцінок прогнозів на їх основі, які застосовують в технічних системах, економіці та фінансах. Однак, далеко не всі методи забезпечують високоякісні оцінки прогнозів у конкретних випадках їх застосування. Тому вибір методу моделювання і прогнозування – це досить непроста задача, яка потребує одночасного застосування декількох альтернативних методів і вибору кращого з них на основі статистичного аналізу отриманого результату.

Самими популярними на сьогодні методами моделювання і прогнозування розвитку процесів довільної природи є такі: методи на основі регресійного аналізу та нечіткої логіки, ймовірнісно-статистичні методи (ЙСМ), метод групового врахування аргументів (МГВА), метод (машина) опорних векторів (МОВ), нейронні мережі, методи на основі „м’яких” обчислень (генетичні та імунні алгоритми) і деякі інші.

Кожний із згаданих методів в тій чи іншій мірі може враховувати невизначеності статистичного або структурного характеру. Кращі результати прогнозування процесів з невизначеностями довільного типу можна отримати за допомогою ймовірнісних методів та нечіткої логіки [6-8]. За своєю природою ці методи близькі до способів моделювання ситуацій та прийняття рішень людиною, а тому їх застосування в системах управління та підтримки прийняття рішень (СППР) може дати значний позитивний ефект.

Одним із сучасних напрямів розвитку ймовірнісних методів моделювання і прогнозування є статичні і динамічні мережі Байєса. Вони призначені для розв'язування задач розпізнавання образів (ситуацій), прогнозування розвитку процесів та їх діагностики, а також для керування процесами і об'єктами різної природи. Перевагами моделей у формі мереж Байєса є їх висока наглядність (ациклічний спрямований граф), можливості створення моделей високої розмірності (в смислі великого числа вузлів-змінних), використання неперервних та дискретних змінних в одній моделі, врахування невизначеностей структурного і стохастичного типів, можливості використання методів точного і наближеного формування ймовірнісного висновку. Загалом мережі Байєса можна характеризувати як досить складний високоресурсний, але високоефективний ймовірнісний метод моделювання та прогнозування розвитку процесів довільної природи. Очевидно, що при створенні адаптивної прогнозуючої системи необхідно використовувати методи, які мають взаємодоповнюючі характеристики щодо можливостей описання процесів та можливостей оцінювання високоякісних прогнозів.

Оцінювання якості моделі і прогнозу

В подальшому викладенні ми будемо розрізняти якість моделі та якість прогнозу, оскільки саме такий підхід дає можливість коректної організації відповідного обчислювального процесу. Якість моделі можна оцінити за допомогою декількох статистичних критеріїв якості, зокрема таких: коефіцієнт множинної детермінації (R^2), який визначає інформативність моделі по відношенню до інформативності даних; статистика Дарбіна-Уотсона (DW) – визначає ступінь

автокорельованості похибок моделі; інформаційний критерій Акайке (AIC) і статистика Байєса-Шварца (BSC) – критерії загального типу, які враховують суму квадратів похибок моделі, кількість вимірів та параметрів моделі; сума квадратів похибок моделі ($\sum e^2(k)$) та інші. Для автоматизованого вибору кращої моделі краще скористатись інтегральним критерієм якості [9]:

$$IK = e^{1-R^2} + \frac{SSE}{N} + \begin{cases} \ln(AIC + BSC), & якщо \;\; AIC + BSC > 0 \\ e^{AIC+BSC}, & якщо \;\; AIC + BSC \le 0 \end{cases} + e^{2-DW} + ,$$
$$+ \ln(СКП) + \ln(САПП) + e^{U}$$

де $СКП$ – середньоквадратична похибка історичного однокрокового прогнозу; $САПП$ – середня абсолютна похибка прогнозу в процентах; U – коефіцієнт Тейла (наближається до нуля, якщо модель придатна для прогнозування).

Важливим моментом процесу прогнозування є об'єктивне коректне визначення якості отриманого прогнозу. Оскільки прогнозовані значення – випадкові величини, то для оцінювання їх якості також необхідно використовувати декілька статистичних критеріїв. Рис. 2 ілюструє часову вісь та відрізки часу, на яких виконується оцінювання моделі і перевірка якості прогнозу.

Рисунок 2 – Види прогнозування за часовим рядом

Наявну вибірку даних доцільно розділити на навчальну та перевірочну. На навчальній вибірці виконується оцінювання параметрів моделі процесу і

реалізується так званий „історичний” прогноз, який дає змогу встановити якість однокрокового прогнозу на цьому участку ряду даних. Прогноз на перевірочній частині вибірки даних в науковій літературі називають ще прогнозом *ex post*. В різних емпіричних дослідженнях для перевірки рекомендують залишати (10 – 50) % значень ряду даних. Загальною рекомендацією є використання для навчання (оцінювання параметрів моделі) не менше 50-60 значень ряду. Прогнозування значень поза вибіркою даних називають прогнозом *ex ante* (рис. 2).

Досить часто якість оцінок прогнозів визначають лише за допомогою середньоквадратичної похибки (СКП), але значення СКП залежить від масштабу даних, а тому цієї характеристики недостатньо для поглибленого аналізу якості прогнозу. Поглиблене оцінювання якості прогнозів досягається за рахунок використання критеріїв, які дають відносні оцінки (наприклад, коефіцієнт Тейла) та відносні оцінки в процентах (наприклад, $САПП$). Перевагою їх використання є те, що вони не залежать від масштабу даних. $САПП$ і коефіцієнт Тейла обчислюють за виразами:

$$САПП = \frac{1}{s}\sum_{i=1}^{s}\frac{|y(k+i)-\hat{y}(k+i,k)|}{|y(k+i)|}\times 100\% = \frac{1}{s}\sum_{i=1}^{s}\frac{|e(k+i)|}{|y(k+i)|}\times 100\%,$$

$$U = \frac{\sqrt{\frac{1}{s}\sum_{k=1}^{s}[y(k+i)-\hat{y}(k+i)]^2}}{\sqrt{\frac{1}{s}\sum_{i=1}^{s}y^2(k+i)}+\sqrt{\frac{1}{s}\sum_{i=1}^{s}\hat{y}^2(k+i)}},$$

де s - число кроків прогнозування; $y(k+i)-$ фактичні значення даних; $\hat{y}(k+i)-$ оцінки прогнозів, обчислені як умовне математичне сподівання відповідної функції [5, 6]. Коефіцієнт Тейла U – це важливий індикатор якості моделі і прогнозу одночасно; за означенням, $0 \le U \le 1$. При $U \to 0$ прогнозовані значення наближаються до фактичних значень ряду – модель ідеальна. Тобто U дає можливість встановити придатність моделі (методу) в принципі для обчислення оцінок прогнозів.

У багатьох випадках кращих результатів прогнозування можна досягти за рахунок усереднення (або комбінування за допомогою вагових коефіцієнтів) оцінок прогнозів, отриманих за допомогою різних методів. Для двох методів прогнозування середня оцінка визначається просто:

$$\hat{y}_c(k) = \frac{\hat{y}_1(k) + \hat{y}_2(k)}{2},$$

де $\hat{y}_c(k)$ – оцінка комбінованого прогнозу; $\hat{y}_1(k)$, $\hat{y}_2(k)$ – прогнози, отримані за різними методами. Якщо окремі прогнози не зміщені (це повинен забезпечувати метод прогнозування), то комбінованний прогноз також буде незміщеним. Похибка комбінованого прогнозу:

$$e_c(k) = y(k) - \hat{y}_c(k) = y(k) - \frac{\hat{y}_1(k) + \hat{y}_2(k)}{2} = \frac{e_1(k) + e_2(k)}{2},$$

Дисперсія похибки комбінованого прогнозу:

$$\text{var}\left[\frac{e_1(k) + e_2(k)}{2}\right] = E\left[\frac{e_1(k) + e_2(k)}{2}\right]^2 = \frac{1}{4}E\left[e_1^2(k) + 2e_1(k)e_2(k) + e_2^2(k)\right] =$$

$$= \frac{1}{4}\left\{E\left[e_1^2(k)\right] + 2E\left[e_1(k)e_2(k)\right] + E\left[e_2^2(k)\right]\right\} =$$

$$= \frac{1}{4}\left[\sigma_1^2 + 2\frac{E[e_1(k)e_2(k)]}{\sigma_1\sigma_2}\sigma_1\sigma_2 + \sigma_2^2\right] = \frac{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2}{4}.$$

Таким чином, дисперсія комбінованого прогнозу обчислюється за виразом:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}{4},$$

де ρ – коефіцієнт кореляції між похибками прогнозу. Якщо похибки прогнозування за двома моделями незалежні, тобто, $\rho = 0$, то остання формула спрощується:

$$\sigma_c^2 = \frac{\sigma_1^2 + \sigma_2^2}{4}.$$

Таким чином, якщо дисперсії близькі за значеннями і похибки оцінок прогнозів незалежні, то дисперсія комбінованої похибки буде значно меншою будь-якої з двох дисперсій. Але навіть при існуванні досить високої кореляції між похибками прогнозування дисперсія похибки комбінованого

прогнозу буде меншою ніж дисперсія кожного методу окремо. Наприклад, при $\sigma_1^2 = \sigma_2^2 = 100$ і $\rho = 0.8$ дисперсія комбінованого прогнозу складе 90.

Якщо інформація стосовно статистичних характеристик індивідуальних прогнозів невідома, то можна присвоїти різні вагові коефіцієнти окремим прогнозам на основі суб'єктивних або експертних суджень:

$$\hat{y}_c(k) = w_1 \hat{y}_1(k) + w_2 \hat{y}_2(k), \qquad (5)$$

де w_1, w_2 – вагові коефіцієнти. Очевидно, що більші значення вагових коефіцієнтів необхідно присвоювати тим індивідуальним прогнозам, які мають меншу дисперсію похибок. При цьому для коректності обчислень необхідно, щоб виконувалась умова: $w_1 + w_2 = 1$.

Як правило, можна обчислити похибки прогнозів для конкретних моделей і процесів. Це дає можливість об'єктивно підійти до розв'язання задачі вибору вагових коефіцієнтів. Оскільки моделі, які дають меші суми квадратів похибок прогнозів, генерують якісніші прогнози, то логічно прийняти цю міру за основу для визначення вагових коефіцієнтів. Позначимо суму квадратів похибок прогнозування (для історичного прогнозу) через

$$sse = \sum_{k=1}^{N} e^2(k).$$

Тепер можна записати вирази для вагових коефіцієнтів окремих прогнозов:

$$w_1 = \frac{1/sse_1}{1/sse_1 + 1/sse_2}, \qquad (6)$$

$$w_2 = \frac{1/sse_2}{1/sse_1 + 1/sse_2}, \qquad (7)$$

де sse_1, sse_2 – суми квадратів похибок прогнозів для кожного з методів, що використовуються. Таким чином, ми об'єктивно присвоїли більший ваговий коефіцієнт точнішому методу прогнозування.

Адаптивне обчислення оцінок прогнозів

Для збереження якості прогнозів в умовах нестаціонарності досліджуваного процесу, а також для підвищення якості оцінок прогнозів процесів з довільними статистичними характеристиками необхідно застосовувати адаптивні схеми оцінювання прогнозів. Вихідними величинами для аналізу якості прогнозів та формування адаптивних схем їх оцінювання є значення похибок прогнозів та їх статистичні характеристики (дисперсія, стандартне відхилення, середні значення похибок). Для розв'язання задачі адаптації прогнозуючої моделі до процесу (іншими словами, до вимог стосовно якості прогнозу) можна скористатись такими обчислювальними можливостями:

– рекурсивне оцінювання параметрів математичних і статистичних моделей з надходженням нових вимірів, що сприяє структроно-параметричному уточненню моделі та підвищенню якості прогнозу по мірі надходження нових даних;

– автоматизований аналіз часткової автокореляційної функції (ЧАКФ) залежної (основної) змінної з подальшим корегуванням структури моделі шляхом введення/вилучення додаткових лагових значень;

– почергове введення в модель можливих регресорів (особливо провідних індикаторів) та аналіз їх впливу на якість прогнозу;

– автоматизований аналіз функції часткової взаємної кореляції основної змінної з регресорами з метою корегування лагових значень регресора у правій частині рівняння;

– оцінювання параметрів збурюючих впливів в реальному часі по мірі надходження вимірів та використання отриманих оцінок для покращення оцінок прогнозу (наприклад, при використанні фільтра Калмана як інструмента згладжування та/або прогнозування) [4];

– адаптування структур нейронних і байєсових мереж на основі нових вимірів та експертних оцінок;

– використання похибок прогнозів для обчислення вагових коефіцієнтів при формуванні комбінованої оцінки прогнозу з декількох індивідуальних оцінок;

– автоматизований (адаптивний) вибір оптимальних вагових коефіцієнтів в процедурах експоненційного згладжування, пошуку подібних траєкторій, регресії на опорних векторах та деяких інших методах;

– автоматизований аналіз залишків регресійних моделей з метою встановлення їх стаціонарності та інформативності і можливе корегування структури моделі процесу на основі результатів аналізу;

– адаптивне формування масивів вимірів змінних стану процесу за допомогою методів ієрархічного комплексування (інтегрування) даних, що забезпечує підвищення їх інформативності.

Очевидно, що існує досить широкий спектр методів адаптування схем оцінювання прогнозів до статистичних характеристик даних і характеристик оцінок прогнозів. Застосування тієї чи іншої схеми залежить від конкретної постановки задачі, якості та об’єму експериментальних (статистичних) даних, сформульованих вимог до якості оцінок прогнозів та часу, наявного для виконання обчислень. Кожний метод адаптивного формування оцінки прогнозу має свої особливості обчислювального характеру, які мають бути враховані при створенні системи адаптивного прогнозування.

Наявність зворотного зв’язку в системі адаптивного прогнозування за похибкою моделі сприяє наближенню якості моделі до рівня, необхідного для досягнення високоякісного прогнозу. Зворотний зв’язок за похибкою прогнозу також спрямовується на підвищення точності оцінок прогнозів, яка досягається за рахунок покращення якості (інформативності) даних, уточнення структури моделі, а також завдяки застосуванню схем усереднення оцінок прогнозів. Цей контур також дає можливість уникнути „перенавчання”, яке приводить до підвищення точності апроксимації даних побудованою моделлю, але, як правило, знижує якість оцінок прогнозів.

Приклад застосування концепції адаптивного прогнозування: побудова моделі для індексу споживчих цін

Формування індексу споживчих цін – один із визначальних макроекономічних процесів, який відображає фактичний стан споживання в Україні. Побудуємо математичні моделі динаміки для наступних макроекономічних процесів України: формування внутрішнього валового продукту (ВВП), індекс споживчих цін (ІСЦ) і грошовий агрегат М3. Для побудови використано фактичні місячні дані з січня 1996 по січень 2005 року, всього 109 значень. Кореляційна матриця для цих змінних:

ІСЦ	ВВП	М3
1	-0.3043610	-0.2491123
-0.3043610	1	0.9317529
-0.2491123	0.9317529	1

Корельованість ІСЦ з ВВП та агрегатом М3 незначна; в подальшому ця інформація буде використана при побудові альтернативних варіантів математичних моделей процесів, що розглядаються.

Модель індексу споживчих цін

Спочатку розглянемо можливість описання індексу споживчих цін за допомогою моделі авторегресії з ковзним середнім. Авторегресійні моделі самі прості за своєю структурою, але досить часто вони мають високу ступінь адекватності досліджуваному процесу, прийнятну для подальшого використання. Автокореляційна функція процесу наведена в таблиці 1. При побудові моделі індекс споживчих цін позначимо як *isc* .

Таблиця 1

Автокореляційна функція процесу формування оптових цін (ІСЦ)

Часові дані: 1996:01 2005:01
Всього спостережень: 109

АКФ	Часткова АКФ		AC	PAC	Q-стат.	Ймов
. |**** |	. |**** |	1	0.570	0.570	36.373	0.000
. |** |	.*|. |	2	0.230	-0.140	42.378	0.000
. |* |	. |* |	3	0.129	0.089	44.284	0.000
. |* |	. |. |	4	0.076	-0.022	44.957	0.000
. |. |	. |. |	5	0.064	0.043	45.433	0.000
. |* |	. |* |	6	0.120	0.101	47.137	0.000
. |* |	. |* |	7	0.180	0.088	50.960	0.000
. |* |	.*|. |	8	0.068	-0.136	51.515	0.000
. |. |	. |* |	9	0.052	0.105	51.840	0.000
. |* |	. |. |	10	0.072	-0.001	52.468	0.000
. |* |	. |* |	11	0.117	0.107	54.166	0.000

В таблиці 1 скорочення AC відповідає АКФ, а PAC – часткова АКФ.

Як видно з АКФ, при побудові моделей необхідно починати з моделей нижчих порядків, які часто мають прийнятну адекватність процесу і забезпечують високу якість прогнозу. Результати оцінювання моделі авторегресії першого порядку за допомогою методу найменших квадратів наведені в таблиці 2.

Таблиця 2. Результати оцінювання моделі AP(1) для ІСЦ.

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1996:02 2005:01
Використано спостережень: 108 після корегування
Модель AP(1): I=C(1)+C(2)*I(-1)

	Оцінки коеф-в	Станд. похибка	t-статистика	Ймов.
C(1)	43.29386	6.624742	6.535178	0.0000
C(2)	0.569826	0.065703	8.672790	0.0000
R-квадрат	0.415067	Середнє залежної змінної		100.7407
Скорегований R-квадрат	0.409549	Станд. відхил. зал. Змінної		1.506189
Станд. похибка регресії	1.157368	Інформ. критерій Акайке		3.148519
Сума квадратів похибок	141.9871	Критерій Шварца		3.198188
		Стат. Дарбіна-Уотсона		1.931805

Отримана модель AP(1):

$$isc(k) = a_0 + a_1\, isc(k-1) + \varepsilon(k) = \\ = 43{,}294 + 0{,}57\, isc(k-1) + e(k),$$

де $e(k)$ – залишки (похибки) моделі, значення яких можна знайти у відповідному файлі пакету програм, що застосовується для побудови моделі. Статистичні характеристики моделі:

$$R^2 = 0{,}415; \quad J = СКП = 141{,}99; \quad DW = 1{,}931.$$

Коефіцієнт детермінації має низьке значення (0,415), сума квадратів похибок досить високе (141,99), а статистика Дарбіна-Уотсона (1,931) наближається до найкращого значення. Таким чином, загалом адекватність моделі AP(1) досить низька, а тому структура моделі потребує уточнення. Характеристики якості (історичного – за навчальною вибіркою) однокрокового прогнозу:

$$СеКП = 1{,}36; \quad САП = 1{,}02; \quad САПП = 1{,}008; \quad U = 0{,}0067,$$

тобто, середньоквадратична похибка (*СеКП*), середня абсолютна похибка (*САП*), середня абсолютна похибка в процентах (*САПП*) і коефіцієнт Тейла,

який свідчить про загальну придатність моделі для прогнозування (ідеальне значення – нуль).

В таблиці 3 наведені результати оцінювання авторегресійної моделі АР(3). Всі три коефіцієнти моделі значимі.

Таблиця 3. Результати оцінювання моделі АР(3).

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1996:04 2005:01
Всього спостережень після корегування крайніх значень: 106
Модель: I=C(1)+C(2)*I(-1)+C(3)*I(-2)+C(4)*I(-3)

	Оцінки коеф-в	Станд. похиб.	t-статистика	Ймов.
C(1)	46.96556	9.330171	5.033729	0.0000
C(2)	0.613458	0.098563	6.223995	0.0000
C(3)	-0.157670	0.113613	-1.387787	0.1682
C(4)	0.077517	0.086493	0.896229	0.3722
R-квадрат	0.316655	Середнє залежної змінної		100.6604
Скорегований R-квадрат	0.296556	Станд. відхил. зал. змінної		1.372428
Станд. похибка регресії	1.151076	Інформ. критерій Акайке		3.156277
Сума квадратів похибок	135.1476	Критерій Шварца		3.256785
		Стат. Дарбіна-Уотсона		1.992389

Таким чином, можна записати наступну модель:

$$isc(k) = a_0 + a_1\, isc(k-1) + a_2\, isc(k-2) + a_3\, isc(k-3) + \varepsilon(k) =$$
$$= 46{,}96 + 0{,}61\, isc(k-1) - 0{,}15\, isc(k-2) + 0{,}08\, isc(k-3) + e(k),$$

Для цієї моделі спостерігається зменшення коефіцієнта детермінації від 0,415 до 0,317 (деяке погіршення); зменшення суми квадратів похибок від 141,99 до 135,148 і деяке покращення статистики Дарбіна-Уотсона: від 1,931 до 1,992 (похибки моделі можна вважати практично некорельованими). Тобто отримані значення характеристик моделі:

$$R^2 = 0{,}317; \quad J = СКП = 135{,}148; \quad DW = 1{,}992.$$

Характеристики однокрокового прогнозу для даної моделі:

$$СеКП = 1{,}36; \quad САП = 1{,}02; \quad САПП = 1{,}01; \quad U = 0{,}0068,$$

тобто модель загалом придатна для прогнозування (за коефіцієнтом Тейла, який наближається до ідеального значення), а три інших показники свідчать про високу точність прогнозу. Необхідно зазначити, що показники якості прогнозу для моделей АР(1) і АР(3) є практично однаковими.

Результати оцінювання моделі 12-го порядку наведені в таблиці 4.

Таблиця 4. Результати оцінювання моделі АР(12) для ІСЦ.

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1997:01 2005:01
Всього використано спостережень після корегування крайніх значень: 97
Модель: I=C(1)+C(2)*I(-1)++C(3)*I(-2)+C(4)*I(-3)+C(5)*I(-6)+C(6)*I(-7)+C(7)*I(

-8)+C(8)*I(-9)+C(9)*I(-11)+C(10)*I(-12)

	Оцінки коеф-в	Станд. похибка	t-статистика	Ймов.
C(1)	26.23923	14.28763	1.836500	0.0697
C(2)	0.666410	0.102198	6.520741	0.0000
C(3)	-0.247996	0.122488	-2.024654	0.0460
C(4)	0.179378	0.103248	1.737345	0.0859
C(5)	-0.038240	0.096715	-0.395392	0.6935
C(6)	0.016650	0.112406	0.148120	0.8826
C(7)	-0.037042	0.111025	-0.333639	0.7395
C(8)	0.030055	0.095744	0.313915	0.7543
C(9)	0.020445	0.093392	0.218911	0.8272
C(10)	0.149194	0.081977	1.819964	0.0722

R-квадрат	0.435002	Середнє залежної змінної	100.6082
Скорегований R-квадрат	0.376553	Станд. відхил. зал. змінної	1.342857
Станд. похибка регресії	1.060301	Інформ. критерій Акайке	3.052366
Сума квадратів похибок	97.80880	Критерій Шварца	3.317800
		Статистика Дарбіна-Уотсона	1.940130

Модель АР(12) має кращі характеристики ніж попередні моделі:

$$R^2 = 0{,}435; \quad J = СКП = 97{,}80; \quad DW = 1{,}94.$$

Значно зменшилась сума квадратів похибок, підвищилось значення R^2, а значення статистики Дарбіна-Уотсона майже таке, як для моделі АР(3).

Характеристики однокрокового прогнозу для цієї моделі:

$$СеКП = 1{,}337; \quad САП = 1{,}02; \quad САПП = 1{,}013; \quad U = 0{,}0066.$$

Таким чином, характеристики однокрокового прогнозу також найкращі для моделі АР(7). Можна зробити висновок, що процес формування індексу оптових цін може бути описаний моделлю авторегресії АР(7) з високим ступенем адекватності. Ця модель забезпечує також отримання кращого однокрокового прогнозу.

Авторегресія з ковзним середнім для ІСЦ

Розглянемо можливість описання ІСЦ за допомогою моделі АРКС. Характеристики моделі АРКС(1,1) наведені в таблиці 9. Вибрані статистичні параметри, які характеризують якість моделі, мають наступні значення:

Таблиця 5. Результати оцінювання моделі АРКС(1, 1) для ІСЦ.

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після корегування крайніх значень: 108

Змінна	Оцінки коеф-в	Станд. похибка	t-статистика	Ймов.

C	100.6529	0.263645	381.7742	0.0000
AR(1)	0.546826	0.093191	5.867793	0.0000
КС(1)	0.060696	0.133153	0.455838	0.6494

R-квадрат	0.415932	Середнє залеж. змінної	100.7407
Скорегований R-квадрат	0.404807	Станд. відхил. залеж. змінної	1.506189
Станд. похибка регресії	1.162006	Інформ. критерій Акайке	3.165558
Сума квадратів похибок	141.7772	Критерій Шварца	3.240062
Логарифм правдоподібн.	-167.9401	F-статистика	37.38676
Статист. Дарбіна-Уотсона	1.996202	Ймовірність (F-стат.)	0.000000

Інвертовані АР корені	.55
Інвертовані КС корені	-.06

Вибрані статистичні характеристики адекватності цієї моделі:

$$R^2 = 0{,}416; \quad J = СКП = 141{,}78; \quad DW = 1{,}996.$$

Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}362; \quad САП = 1{,}016; \quad САПП = 1{,}005; \quad U = 0{,}0067.$$

Моделі АР(1) і АРКС(1,1) мають практично однакові характеристики адекватності та якості однокрокового прогнозу, а тому перевагу (при виборі з цих двох моделей) можна надати моделі АР(1), яка є простішою. Нижче наведена порівняльна таблиця для всіх побудованих моделей.

Врахування впливу на ІСЦ агрегату М3

Коефіцієнт кореляції між ІСЦ та агрегатом М3 від'ємний: –0,249, тобто формальний взаємозв'язок між цими змінними незначний, але цікаво розглянути вплив М3 на ІСЦ за допомогою моделі. Врахування регресора може покращити деякі характеристики моделі, а також врахувати причинний зв'язок між вибраними змінними. Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені в таблиці 6.

Таблиця 6

Результати оцінювання моделі ІСЦ: АР(1) + регресор М3

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після корегування крайніх значень: 108
Модель: I=C(1)+C(2)*I(-1)+C(3)*M

	Оцінки коеф-в	Станд. похибка	t-статистика	Ймов.
C(1)	44.97869	6.920057	6.499757	0.0000
C(2)	0.554264	0.068261	8.119764	0.0000
C(3)	-2.82E-06	3.30E-06	-0.854463	0.3948

R-квадрат	0.419106	Середнє залеж. змінної	100.7407
Скорегований R-квадрат	0.408041	Станд. відхил. залеж. змінної	1.506189
Станд. похибка регресії	1.158844	Інформ. критерій Акайке	3.160108
Сума квадратів похибок	141.0066	Критерій Шварца	3.234612
Логарифм правдоподібн.	-167.6458	Стат. Дарбіна-Уотсона	1.919823

Спостерігається незначне покращення характеристик моделі та якості прогнозу у порівнянні з АР(1), але коефіцієнт при М3 дуже малий (-2,82Е-06). Однак, формально він є значимим. Таким чином, на розглянутому часовому інтервалі вплив агрегату М3 на індекс споживчих цін незначний і ним можна знехтувати.

Характеристики прогнозу:

$$СеКП = 1{,}340; \quad САП = 1{,}004; \quad САПП = 0{,}994; \quad U = 0{,}0066.$$

Можна припустити, що в обороті було недостатньо грошової маси для того, щоб її вплив на ІСЦ був істотним. З іншого боку, недостатній об'єм грошової маси в національній валюті компенсувався (і продовжує компенсуватись) "твердою" іноземною валютою, зокрема доларами США і, в деякій мірі, євро. Таким чином, слабка українська економіка позитивно впливає на курс долара США завдяки фактичному введенню його в частковий оборот на відносно великій території. Встановити фактичне співвідношення між об'ємами національної та іноземної валют в обороті можливо, але для цього необхідно отримати додаткові статистичні дані і виконати спеціальне дослідження. Зокрема, необхідно мати дані щодо об'ємів реалізації торговельних операцій у валюті підприємствами всіх форм власності. Очевидно, що отримані такі дані для тіньового обороту непросто.

Визначення впливу ВВП на ІСЦ

Як було показано на початку цього параграфа, коефіцієнт кореляції між ІСЦ та ВВП складає –0,304, тобто формально це невелике значення. Також логічно припустити, що зростання ВВП має приводити до зменшення ІСЦ (про це свідчить також знак коефіцієнта кореляції між цими змінними).

Характеристики змішаної моделі: авторегресія АР(1) + парна регресія наведені в таблиці 7.

Таблиця 8. Результати оцінювання змішаної регресії для ІСЦ і ВВП.

Метод оцінювання: метод найменших квадратів
Скорегована часова вибірка даних: 1996:02 2005:01
Всього використано спостережень після корегування крайніх значень: 108
Модель: I=C(1)+C(2)*I(-1)+C(3)*V

	Оцінки коеф-в	Станд. похибка	t-статистика	Ймов.
C(1)	45.53125	7.156626	6.362112	0.0000
C(2)	0.549611	0.070124	7.837741	0.0000
C(3)	-1.37E-05	1.64E-05	-0.833627	0.4064
R-квадрат	0.418913	Середнє залеж. змінної		100.7407
Скорегований R-квадрат	0.407844	Станд. відхил. залеж. змінної		1.506189
Станд. похибка регресії	1.159037	Інформ. критерій Акайке		3.160441
Сума квадратів похибок	141.0536	Критерій Шварца		3.234945
Логарифм правдоподібн.	-167.6638	Стат. Дарбіна-Уотсона		1.916289

Отримано рівняння:

$$isc(k) = 45{,}53 + 0{,}55\ isc(k-1) - (1{,}37E-05)\ vvp(k) + e(k)$$

з характеристиками: $R^2 = 0{,}419$; $J = СКП = 141{,}05$; $DW = 1{,}92$.
Характеристики якості однокрокового прогнозу:

$$СеКП = 1{,}335;\quad САП = 1{,}004;\quad САПП = 0{,}993;\quad U = 0{,}0066.$$

Коефіцієнт при змінній *ВВП* невеликий і знаходиться на межі статистичної значимості, тобто вплив *ВВП* на *ІСЦ* незначний. Причиною такого незначного впливу може бути некоректний розподіл *ВВП*, який призводить до того, що більшість населення проживає на межі або навіть нижче межі бідності. Одночасно інша (менша) частина привласнює більшість благ і користується ними, але, очевидно, що це користування не приводить до позитивного впливу на *ІСЦ*. Тому і маємо, що *ВВП* зростає начебто значними темпами, а добробут населення сильно від нього відстає. Таким чином, можна зробити висновок, що на часовому періоді, який розглядається в даному прикладі, відбувався несправедливий розподіл суспільних благ, що проявилось, в даному випадку, в незначному впливі *ВВП* на споживчі ціни для більшості населення України. Побудована математична модель є формальним (об'єктивним) підтвердженням відомого факту нерівномірного розподілу матеріальних благ, вироблених в Україні.

В таблиці 8 зведені характеристики математичних моделей, побудованих для індексу споживчих цін, і характеристики однокрокових прогнозів, обчислених на основі цих моделей. Ця таблиця дає можливість оперативно порівняти результати моделювання та прогнозування, отримані за допомогою методики Бокса-Дженкінса, а також встановити можливість практичного використання результатів.

Таблиця 8. Результати моделювання і однокрокового прогнозування індексу оптових цін.

Тип моделі	Характеристики моделі			Характеристики прогнозу			
	R^2	$\sum e^2(k)$	DW	$СеКП$	$САП$	$САПП$	Коеф-т Тейла
АР(1)	0,415	141,99	1,931	1,360	1,020	1,008	0,0067
АР(3)	0,317	135,148	1,992	1,360	1,020	1,011	0,0068
АР(7)	0,346	127,244	1,811	1,360	1,012	1,002	0,0067
АР(12)	0,435	97,80	1,941	1,337	1,020	1,013	0,0066
АРКС(1,1)	0,416	141,78	1,996	1,362	1,016	1,005	0,0067
АР(1)+М3	0,419	141,007	1,919	1,340	1,004	0,994	0,0066
АР(1)+ВВП	**0,419**	**141,054**	**1,916**	**1,335**	**1,004**	**0,993**	**0,0066**
Моделі для відхилень ІСЦ від середнього							
АР(1)	0,415	141,99	1,931	1,360	1,020	100,40	0,662
АР(13)	0,440	95,69	1,965	1,346	1,020	109,09	0,848

Результати моделювання, наведені в таблиці 12, свідчать про те, що практично всі моделі, побудовані для ІСЦ (перші сім моделей), є придатними для прогнозування, оскільки коефіцієнт Тейла вимірюється тисячними долями. Характеристики однокрокового прогнозу для моделей різного порядку відрізняються несуттєво. Найкращі характеристики щодо прогнозування має модель АР(1)+ВВП. Для моделі АР(12) отримано найменше значення суми квадратів похибок моделі, але характеристики прогнозів, отриманих за цією моделлю, не кращі від інших.

Дві моделі, які побудовані для відхилень ІСЦ від середнього, мало придатні для прогнозування. Про це свідчать високі значення коефіцієнта Тейла та середньої абсолютної похибки в процентах. Очевидно, що для прогнозування відхилень від середнього необхідно знайти більш придатну структуру математичної моделі.

Реалізація методів заповнення пропусків даних

Для заповнення пропусків даних було обрано наступні методи: заповнення пропусків середнім значенням, заповнення пропусків оцінками прогнозів, заповнення пропусків випадковими величинами з нормального розподілу, заповнення пропусків випадковими величинами з рівномірного розподілу.

Для проведення експериментів обрано часовий ряд: курс валют долар/гривня 2014 року, потужністю 249 значень.

Результати експериментів наведено у табл. 9.

Тип моделі	Адекватність				Якість прогнозу		
		R^2	Σe^2	DW	RMSE	MAPE	V
Модель АР(1)(без пропусків)		0,9931	7,7861	1,6387	0,79633	5,5794	0,03275
Модель АР(1)(5 пропусків)		0,5827	801,64	2,6234	2,27192	16,135	0,11506
Модель АР(1)(10 пропусків)		0,4740	1404,5	2,6735	3,24695	19,565	0,13838
Модель АР(1)(15 пропусків)		0,3780	2107,8	2,5680	3,67795	22,820	0,15775
Заповнення пропусків середнім значенням							
Модель АР(1)(5		0,9666	35,7187	1,9063	1,27412	9,93203	0,05208

пропусків)							
Модель AP(1)(10 пропусків)		0,9641	35,6761	1,9051	1,36985	10,5608	0,05577
Модель AP(1)(15 пропусків)		0,9611	35,6248	1,9038	1,43505	10,9474	0,05820
Заповнення пропусків оцінками прогнозів							
Модель AP(1)(5 пропусків)		0,9931	7,8350	1,6413	0,79676	5,5638	0,03276
Модель AP(1)(10 пропусків)		0,9929	7,9625	1,6469	0,79721	5,52365	0,03271
Модель AP(1)(15 пропусків)		0,9926	8,1649	1,6554	0,79676	5,45940	0,03271
Заповнення пропусків нормально розподіленими випадковими величинами							
Модель AP(1)(5 пропусків)		0,9923	7,8340	1,7132	0,8092	5,57099	0,03323
Модель AP(1)(10 пропусків)		0,9921	8,7618	1,7083	0,8125	5,67968	0,03334
Модель AP(1)(15 пропусків)		0,9921	8,6837	1,6762	0,8113	5,62726	0,03328
Заповнення пропусків рівномірно розподіленими випадковими величинами							
Модель AP(1)(5 пропусків)		0,9898	11,5444	2,2222	0,8545	6,23299	0,03503
Модель AP(1)(10 пропусків)		0,9792	22,3559	1,6652	1,0633	8,53448	0,04337
Модель AP(1)(1 пропусків)		0,96951	31,4562	1,7379	1,1997	9,40108	0,04895
Експоненційне згладжування							
Модель AP(1)		0,96643	4,05299	0,82982	0,75214	5,14824	0,03116
Згладжування за допомогою ковзного середнього							
Модель AP(1)		0,99586	4,59200	0,47552	0,77069	5,25861	0,03189

Таблиця 9 – Результати експериментів по заповнення пропусків для часового ряду : курс валют долар/гривня 2014 року

Висновки

Запропонована концепція формулювання та розв’язання задач адаптивного моделювання і прогнозування на основі методології системного аналізу, яка ґрунтується на комплексному використанні

методів попередньої обробки і аналізу даних, математичного і статистичного моделювання, прогнозування та оптимального оцінювання станів процесів довільної природи. Концепція суттєво відрізняється від відомих системним підходом до розв'язання задач прогнозування і можливістю врахування в процесі формування оцінок прогнозів невизначеностей статистичного, параметричного та структурного характеру. Розв'язана задача заповнення пропусків даних альтернативними методами. Ефективність заповнення пропусків проілюстрована на прикладі фінансового часового ряду. Врахування особливостей та обробка невизначеностей забезпечується інструментарієм, який грунтується на методах фільтрації даних, формування структури моделі та оцінювання параметрів моделей-кандидатів.

Використання запропонованої концепції забезпечує отримання високоякісних (за точністю) коротко- та середньострокових прогнозів за умови наявності інформативних даних, а також формування на їх основі альтернативних оптимальних і раціональних рішень та передбачає універсальність застосування до широкого класу процесів. Запропонована структурна схема оцінювання прогнозів містить два контури адаптації і якість функціонування кожного з них аналізується за допомогою окремих множин статистичних критеріїв якості. Ефективність застосування концепції проілюстрована на прикладі адаптування структури мережі Байєса до нових масивів даних, як інструменту класифікації та прогнозування.

В подальших дослідженнях передбачається застосування запропонованої концепції до розв'язання задач прогнозування конкретних процесів з використанням альтернативних адаптивних методів формування оцінок прогнозів.

Література

1. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования. – Москва: Финансы и статистика, 2003. – 414 с.

2. Зельнер А. Байесовские методы в эконометрии. – Москва: Статистика, 1980. – 438 с.

3. Chatfield C. Time series forecasting. – London: Chapman & Hall, 2000. – 267 p.

4. Згуровский М.З., Подладчиков В.Н. Аналитические методы калмановской фильтрации. – Київ: Наукова думка, 1995. – 285 с.

5. Бідюк П.І. Системний підхід до прогнозування на основі моделей часових рядів // Системні дослідження та інформаційні технології, 2003, № 3, с. 88-110.

6. Бідюк П.І., Савенков О.І., Баклан І.В. Часові ряди: моделювання і прогнозування. – Київ: ЕКМО, 2003. – 144 с.

7. Зайченко Ю.П. Основи проектування інтелектуальних систем. – Київ: Слово, 2003. – 351 с.

8. Згуровский М.З., Бидюк П.И., Терентьев А.Н. Методы построения байесовских сетей на основе оценочных функций // Кибернетика и системный анализ, 2008, № 2, с. 81-88.

9. Демківський Є.О., Бідюк П.І. Система підтримки прийняття рішень при прогнозуванні нестаціонарних процесів // Наукові праці Миколаївського державного гуманітарного університету ім. Петра Могили, 2008, Вип. 77, с. 137-159.

АНОТАЦІЇ

Михайлюк О.О.

СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ МОДЕЛЮВАННЯ ТА ПРОГНОЗУВАННЯ ПРОЦЕСІВ НА БІРЖІ

Проблематика. В умовах динамічної мінливості процесів на біржі дуже важливо мати відповідне програмне забезпечення, яке б допомагало своєчасно приймати відповідну стратегію дій. Вже існуючі системи підтримки прийняття рішень (наприклад, Eviews та SPSS) надають широкі можливості для моделювання та прогнозування процесів ціноутворення акцій, але для їх застосування потрібне знання внутрішньої системи команд і займає досить тривалий проміжок часу. Також важливо мати змогу спрогнозувати не тільки майбутню ціну акцій, але й можливе відхилення від цього прогнозу. Це підвищить надійність обраної стратегії поведінки на біржі.

Мета дослідження. Головна мета дослідження: побудова математичних моделей для вибраних біржових процесів, оцінювання прогнозів на основі цих моделей, розробка програмного забезпечення для виконання обчислювальних експериментів.

Методика реалізації. Для розв’язання поставлених задач використано такі методи: методи аналізу процесів на біржі, математичні моделі авторегресії ковзного середнього і авторегресії умовної гетероскедастичності, метод найменших квадратів (і його рекурентна модифікація) для визначення знаходження параметрів моделей.

Результати дослідження. Основними результатами дослідження є такі: створено інформаційну аналітичну систему для моделювання та прогнозування процесів на біржі на базі авторегресійних моделей з ковзним середнім. Отримані результати прогнозування вибраних цін акцій та можливого відхилення від цього прогнозу.

Висновки. Розглянуто основні типи процесів на біржі. Виявлено особливості моделювання процесів з трендом і гетероскедастичних процесів. Виконано огляд існуючих методів моделювання і прогнозування процесів на біржі.На

основі моделювання і прогнозування визначені кращі моделі для вибраних процесів на біржі.

Ключові слова: стаціонарні процеси ,інтегровані процеси, гетероскедастичні процеси, модель авторегресії ковзного середнього, авторегресійна умовно гетероскедастична модель, система підтримки прийняття рішень.

К.В. Богун

МОДЕЛЮВАННЯ І ПРОГНОЗУВАННЯ ВВП УКРАЇНИ ЗА ДОПОМОГОЮ РЕГРЕСІЙНИХ МОДЕЛЕЙ

Проблематика. Одним з найважливіших показників системи національних рахунків є валовий внутрішній продукт (ВВП) країни. Він є узагальнювальним показником економічного розвитку держави, який відображає кінцеві результати діяльності резидентів країни, що виробляють товари та послуги.За допомогою моделі ВВП можливо визначити подальший економічний розвиток та чинники, що на нього впливають..

Мета дослідження. Головна мета дослідити динаміку ВВП України за 1991-2015 роки. Визначення особливості ВВП, чинники, що впливають на його динаміку, а також знайти найкращу модель для прогнозування.

Методика реалізації. Для розв’язання поставленої задачі використовувалися регресійні рівняння. Було досліджено багато моделей серед яких авторегресійні рівняння різних ступенів, авто регресійні рівняння з ковзним середнім, авторегресіні рівняння з ковзним середнім з екзогенним фактором. Оскільки,процес оцінювання параметрів рівняння – це випадковий процес, то для підвищення надійності було використано статистичні параметри: коефіцієнт множинної детермінації, сума квадратів похибок, статистика Дарбіна-Уотсона, середня абсолютна похибка.

Результати дослідження. Були представленні моделі для прогнозування на три кроки, найкращою виявилась ARIMA(6,1), оскільки середня абсолютна похибка в процентах дорівнює 1,308944, а коефіцієнт Тейла 0,004171

Висновки. Прогноз показника ВВП характеризує кількісні параметри економічного розвитку країни та його складові за етапами економічного обігу, оскільки відображає вартість кінцевих результатів економічної діяльності за певний проміжок часу.

Ключові слова: ВВП, моделювання, прогнозування, регресійні рівняння, статистичні параметри.

О.М.Бойко, П.І. Бідюк

ПРОЦЕСИ КРЕДИТУВАННЯ ФІЗИЧНИХ ОСІБ З ВИКОРИСТАННЯМ НЕЛІНІЙНИХ МОДЕЛЕЙ І БАЙЄСІВСЬКИХ МЕРЕЖ

Проблематика. Належним чином спроектовані системи підтримки прийняття рішень (СППР) для моделювання і прогнозування кредитоспроможності фізичних осіб надають можливість врахування невизначеностей ймовірнісного, статистичного і структурного типів. Це приводить до підвищення якості розроблюваних моделей та оцінок прогнозів.

Мета дослідження. Головна мета дослідження: розробити концепцію дослідження процесів кредитування; проаналізувати і вибрати методи дослідження кредитоспроможності; побудувати математичні моделі за статистичними даними.

Методика реалізації. Для розв’язання поставлених задач використано такі методи: лінійна регресія, дерева рішень, нелінійні математичні моделі і методи байєсівського аналізу даних.

Результати дослідження. Основними результатами дослідження є такі: побудовано лінійну регресію, дерева рішень, мережі Байєса та нелінійні логістичні регресії за статистичними даними; знайдено показники якості моделі та якості прогнозу; проаналізовано статистичні критерії аналізу якості класифікації клієнтів, отримано прийнятні за якістю результати оцінювання.

Висновки. Таким чином виконано короткий огляд сучасних систем для оцінювання фінансових ризиків. Проаналізовано їх переваги та недоліки. Докладно описано вибрані методи прогнозування кредитоспроможності фізичних осіб: логістична регресія, лінійна регресія, дерева рішень та

Байєсівські мережі. За допомогою програм GeNIe 2.0 і EViews 7.0 побудовано мережі Байєса, лінійні та нелінійні логістичні регресії, а також за допомогою SPSS побудовані дерева рішень. Сформульовано рекомендації для їх практичного застосування.

Ключові слова: кредитний ризик, банк, логістична регресія, байєсівська мережа, позичальник, лінійна регресія, дерева рішень.

О.В. Фомін, Н.В. Кузнєцова

СКОРИНГОВІ МОДЕЛІ ПОВЕДІНКИ КЛІЄНТІВ-ВЛАСНИКІВ КРЕДИТНИХ КАРТОК ДЛЯ ОЦІНКИ ЇХ ПЛАТОСПРОМОЖНОСТІ

Проблематика. Скорингові моделі, що основані на теорії аналізу виживання, дозволяють застосовувати регресори, які змінюються із плином часу. Такі коваріанти характеризують певні особливості поведінки клієнтів-власників кредитних карток. Тому даний підхід забезпечує високу якість побудованих моделей, що відображається в достовірності класифікації клієнтів та прогнозі їх дій у майбутньому.

Мета дослідження. Головна ціль статті полягає у наступному: проаналізувати існуючі методи оцінки платоспроможності клієнтів-власників кредитних карток, розробити альтернативні моделі для поведінкового скорингу, що основані на методах аналізу виживання, забезпечити покрокову методологію для побудови таких моделей, а також порівняти ефективність розроблених моделей із загальноприйнятими, які основуються на логістичній регресії.

Методика реалізації. Для вирішення поставлених задач були застосовані наступні методи: аналіз, синтез, абстрагування, моделювання, порівняння та узагальнення. Більш вузькоспеціалізовані методи теорії аналізу виживання включають в себе моделі пропорційних ризиків та статистику Каплан-Мейер.

Результати дослідження. Головний результат роботи – побудовані моделі, що дозволяють брати до уваги залежні від часу змінні та забезпечують таким чином динамічне оцінювання власників кредитних карток. Також важливим є універсальність розробленого підходу у вигляді набору команд мови

програмування R, що дозволяє повторювати подібні дослідження як на таких самих даних, так і на зовсім інших.

Висновки. Отже, був запропонований покроковий алгоритм побудови альтернативних поведінкових скорингова моделей, що основуються на моделі пропорційних ризиків (PH) та статистиці Каплан-Мейер (KM). Такі моделі були побудовані, а їх показники порівняні із загальноприйнятими підходами (логістична регресія). В результаті побудови та аналізу моделі PH виявились значно кращими.

Ключові слова: Поведінкова скорингова модель, аналіз виживання, пропорційні ризики, функція ризику, ризик-менеджмент, логістична регресія.

А.В. Слюсар, О.А. Гусєв, Дульський Д.С.

ВЕЙВЛЕТ АНАЛІЗ ХВИЛЬ ЕЛЛІОТТА

Проблематика. Належним чином спроектовані алгоритми для виявлення хвиль Елліотта і подальшого прогнозування поведінки фінансових ринків надають можливість передбачити стан економіки. Це приводить до підвищення якості розроблюваних моделей та оцінок прогнозів.

Мета дослідження. Головна мета дослідження: розробити індикатор на основі вейвлет перетворення, який дозволить виділяти та досліджувати хвилі Елліотта. Дослідити хвилі Елліотта вже існуючими індикаторами; проаналізувати результати композицій індикаторів та вейвлет-аналізу задля дослідження хвиль Елліотта.

Методика реалізації. Для розв’язання поставлених задач використано такі методи: MACD та показник Херста для аналізу проміжків сталого тренду для хвиль Елліотта; дискретне вейвлет перетворення для фільтрації та згладжування хвилі; неперервне вейвлет перетворення для виділення хвиль Елліотта на часовому ряді; кепстр для аналізу структури хвиль Еллотта та виявлення циклічних частот.

Результати дослідження. Основними результатами дослідження є такі: створення індикатора, який дозволяє виділяти хвилі Елліотта на часовому ряді;

розроблені методі дослідження структурі хвиль та виділення основних компонентів; наведено алгоритм поетапного дослідження .

Висновки. У роботі було розглянуто класичний підхід до виявлення хвиль Елліотта та створено новий індикатор чіткого виявлення хвиль Елліотта(у порівнянні з аналізом вихідного графіку котирувань). За допомогою вейвлет перетворення та кепстр можна досліджувати структуру хвиль та порівнюючи вже з відомими характеристиками хвиль передвіщати кінець та поведінку циклу.

Ключові слова: вейвлет-аналіз, хвилі Елліотта, показник Херста, ковзне середнє, показник MACD, кепстр.

Логін В. В.

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ НАДАННЯ РЕКЛАМНОЇ ІНФОРМАЦІЇ ШЛЯХОМ МАСОВОЇ РОЗСИЛКИ

У даній роботі розглянуто задачу підвищення ефективності масової рекламної розсилки, що є дуже актуальною, оскільки рекламна розсилка є досить поширеною в наш час, причому її проведення без попереднього аналізу може спричинити збитки для компанії замість отримання прибутку.

Дослідження, що описується у даній статті, присвячене підвищенню ефективності масової рекламної розсилки. Зокрема, здійснюється опис побудови моделей відгуку клієнтів на рекламу та вибору кращої моделі для подальшого створення ефективної стратегії розсилки рекламної інформації.

Для розв’язання поставлених задач було використано методологію DataMining, зокрема розв’язувалась задача бінарної класифікації з використанням декількох підходів. Під час дослідження застосовано такі методи: логістична регресія, дерева рішень, нейронні мережі. Після проведення моделювання відгуку клієнтів було здійснено порівняння побудованих моделей за ефективністю.

Основним результатом, що був отриманий в результаті дослідження є таким: після порівняння побудованих моделей за передбачуваним прибутком, виявилось, що найбільш ефективною є модель на основі дерева рішень. При

використанні даної моделі спостерігається підвищення ефективності рекламної розсилки, у порівнянні з моделлю «Розіслати всім», на 8879 од.

Отже, в результаті дослідження було виявлено, що при проведенні рекламної розсилки найкраще використовувати не стратегію «Розіслати всім» чи першочергової розсилки клієнтам, від яких був найбільший дохід, а стратегію, що заснована на моделі відгуку клієнтів на основі дерева рішень.

Також варто зазначити, що за даною темою існує перспектива для подальших досліджень, оскільки рекламна розсилка є досить поширеною у наш час. Зокрема, можна побудувати нові класифікаційні моделі, наприклад, алгоритми «k-means», SVM та ін. Також є потреба у розробці автоматичного підбору оптимального розбиття вихідної вибірки на навчальну та тестову. Ще можливими покращеннями могли б стати вдосконалення побудованих моделей та виконання перевірки розроблених моделей на реальних даних.

Є.І.Морильов, М.В.Дідковська

СИСТЕМА РЕКОНСТРУКЦІЇ ТРИВИМІРНОЇ МОДЕЛІ ПО ДВОВИМІРНИМ ЗОБРАЖЕННЯМ

Проблематика. З кожним днем кількість камер помітно зростає і постає питання обробки отриманої інформації, але для реконструювання тривимірних об’єктів використовуються сканери, ціна яких дуже висока. Крім того, в деяких областях неможливо застосування сканерів чи декількох камер одночасно. Цю проблему можна вирішити за допомогою системи тривимірної реконструкції по двовимірним зображенням. Таким чином, гостро стоїть питання реконструювання моделей, використовуючи звичайну камеру

Мета дослідження. Головна мета дослідження: створити інтелектуальну систему реконструкції тривимірної моделі по двовимірним зображенням; запропонувати модифікацію методу реконструкції тривимірних моделей;розглянути базовий метод реконструювання тривимірних моделей.

Методика реалізації. Для розв’язання поставлених задач використано такі методи: підхід до оцінки відностих положень та поворотів камер;

процедури мінімізації похибки реконструювання; знаходження фундаментальної та істотної матриць.

Результати дослідження. Основними результатами дослідження є такі: запропоновано алгоритм, який дозволяє реконструювати тривимірну модель по двовимірним зображенням використовуючи звичайну камеру та базується на попередній оцінці всіх позицій та поворотів камер, триангуляції та використанні алгоритму bundleadjustment; проаналізовано та обгрунтовано преваги та недоліки представленої системи; проведено аналіз результатів реконструкції, порівняння швидкості роботи та якості реконструкції.

Висновки. Таким чином, запропоновано модифікацію алгоритму реконструювання тривимірних моделей по двовимірним зображенням, яка полягає у попередній оцінці позицій та поворотів камер на всій сцені, що реконструюється. Результати реконструювання показали, що якість реконструйованої моделі не змінилась, але швидкість реконструкції збільшилася.

Ключові слова: тривимірна реконструкція, дескриптори особливих точок, співставлення зображення, точкова тривимірна модель, стереоскопія.

І.В. Пудло, П.І. Бідюк

ПРОГНОЗУВАННЯ ДОХІДНОСТІ БАНКІВСЬКИХ ПРОДУКТІВ З ВИКОРИСТАННЯМ СКОРИНГОВОГО ПІДХОДУ

Проблематика. Стратегія управління ризиком займає важливе місце в забезпеченні ефективного функціонування банківських установ в умовах ринкової економіки та в умовах кризових явищ. Скорингові моделі є базовим індикатором кредитоспроможності потенційного позичальника. На їх основі експерт ухвалює остаточне рішення про видачу кредиту.

Мета дослідження. Розглянути загальні підходи до оцінки кредитного ризику. Проаналізувати супутні ризики в процесі кредитування. Розробити аплікаційну модель для револьверних карток із грейсовим (пільговим) періодом, яка дозволить контролювати дохідність даного продукту.

Методика реалізації. Для розв'язання поставлених задач використано метод оцінки кредитного ризику за допомогою побудови скорингових моделей. Для побудови моделі використовується логістична регресія, що пов'язано із її простотою інтерпретації та стабільністю.

Результати дослідження. Основними результатами дослідження є: розглянуті загальні підходи до оцінки кредитного ризику; розглянуті супутні ризики, що виникають в процесі кредитування;побудована аплікаційна скорингова модель для револьверних карток із грейсовим (пільговим) періодом. Крім того, показано стабільність моделі в майбутньому.

Висновки. В даній роботі було розглянуто проблему кредитного ризику, визначені причини його виникнення та підходи до оцінки. Проаналізовано супутні ризики в процесі кредитування на прикладі револьверних карток з грейсовим (пільговим) періодом. Було розглянуто основні етапи побудови скорингової моделі: формування вибірки, сегментації характеристик тощо. Результатом роботи стала побудована аплікаційна скорингова модель, яка, на основі визначених параметрів, дозволяє провести сегментацію клієнтів таким чином, щоб банк не зазнав збитків від недоотримання прибутку.

Для побудованої моделі індекс GINI мав значення 24%. Динаміка індексу GINI коливається в межах допустимого рівня протягом всього розглянутого періоду, що свідчить про високу стійкість побудованої моделі до змін у даних.

Ключові слова: кредитування; кредитний ризик; скоринг; кредитоспроможність; револьверні картки; грейсовий період; регресія.

В.Г. Вертилецький

АНАЛІЗ ФІНАНСОВО- ЕКОНОМІЧНИХ ПРОЦЕСІВ ЗА ДОПОМОГОЮ СППР

Проблематика. Зараз все актуальнішими стають питання моделювання та прогнозування фінансово-економічних процесів в Україні, за умов перехідної економіки, набувають нових особливостей. Методи, які існують зараз, не завжди дають можливість отримати задовільні результати. Тому вкрай актуальним є пошук нових методів та комбінація існуючих для досягнення

більшої точності та адекватності моделей, що дозволить формувати ефективні управлінські рішення.

Мета дослідження. Головна мета — побудова системи підтримки прийняття рішень (СППР), дослідження, аналіз існуючих та нових моделей прогнозування для макроекономічних та фінансово-економічних процесів та порівняння методів.

Методика реалізації. Для реалізації поставленої задачі була спроектована СППР. Було досліджено та порівняно багато моделей: авторегресійні моделі різних порядків, авторегресії з видаленням незначущих параметрів, також була побудована модель методом групового урахування аргументів. Для порівняння прогнозів, які були розраховані за допомогою побудованих моделей використано статистичні параметри коефіцієнт детермінації, сума квадратів похибок, статистика Дарбіна-Уотсона, середня абсолютна похибка.

Результати дослідження. Були побудовані моделі для прогнозування на чотири кроки, найкращою виявилась AP(13), оскільки середня абсолютна похибка в процентах дорівнює 2,8957, а коефіцієнт Тейла 0,018085.

Висновки. Побудовані авторегресіні моделі виявились достатньо точними при досягненні високих порядків. Розроблена СППР засвідчила високу точність отриманих результатів в порівнянні з комерційними продуктами. В подальшому за допомогою даної СППР можна досліджувати інші процеси та робити прогнози по ним.

Ключові слова: система підтримки прийняття рішень, моделювання, прогнозування, регресійні рівняння, метод групового урахування аргументів, статистичні параметри.

О.О.Крюков, К.В.Козлов, Д.Г. Діденко

ВИКОРИСТАННЯ АЛГОРИТМУ CLOPE ДЛЯ СТРУКТУРУВАННЯ ДОМЕННОЇ МОДЕЛІ АДАПТИВНОЇ СИСТЕМИ НАВЧАННЯ

Проблематика. Сучасні системи адаптивного навчання мають досить широке направлення і використовуються для надання повноцінних онлайн-

курсів з різних областей знань. Це приводить до підвищення складності доменної моделей та вимагає значного залучення експертів для побудови її структури та впорядкування.

Мета дослідження. Головна мета дослідження: запропонувати методику структурування доменної моделі адаптивної системи навчання на основі алгоритмів кластерного аналізу.

Методика реалізації. Для розв'язання поставлених задач використано такі методи: алгоритм кластеризації категорійних даних CLOPE; модифікований алгоритм кластеризації транзакційних даних CLOPE з обмеженням нижньої та верхньої границі кількості елементів у кластері.

Результати дослідження. Основними результатами дослідження є такі: розробка модифікованого алгоритму кластеризації CLOPE; аналіз результатів побудови структури доменної моделі за допомогою базового і модифікованого алгоритмів CLOPE.

Висновки. Таким чином, запропоновано методику структурування доменної моделі на базі модифікованого алгоритму кластеризації категорійних даних CLOPE, що дозволяє задавати параметри, які обмежують верхню та нижню границю кількості елементів у кластері.Виконано розбиття наданої вибірки навчальних матеріалів на кластери за допомогою базового та модифікованого алгоритму CLOPE з різними значеннями вхідних параметрів. Зроблено висновок, що використання модифікованого алгоритму CLOPEдобре підходить для задач, коли необхідно враховувати такі характеристики як розмір кластеру. Це дозволяє автоматично структурувати контент системи навчання на однакові за розміром та змістом модулі.

Ключові слова: адаптивні системи навчання; доменна модель; кластеризація категорійних даних.

Б.С. Журке, Д.Г. Діденко

ВИКОРИСТАННЯ КОМБІНОВАНИХ ОЦІНОК ПРОГНОЗУ ФАРМАЦЕВТИЧНОГО РИНКУ УКРАЇНИ У АГЕНТІЙ МОДЕЛІ

Проблематика. Належним чином спроектована імітаційна модель для прогнозування фармацевтичного рику України, з автоматичним налаштуванням та використанням статистичної моделі в комбінації з нейронною мережею. Надасть можливість в короткий строк, отримати адекватний та більш якісний прогноз порівняно з іншими методами прогнозування.

Мета дослідження. Головна мета дослідження: розглянути імітаційну модель, як інструмент для побудови прогнозу фармацевтичного ринку України; побудувати прогнози за допомогою статистичної моделі, нейронної мережі, імітаційної моделі; оцінити якість прогнозу для кожного із методів.

Методика реалізації. Для розв'язання поставлених задач використано такі методи: агентний підхід до побудови імітаційної моделі з метою побудови прогнозу для фармацевтичного ринку України; методика побудови прогнозу за допомогою інтегрованої моделі авторегресії ковзного середнього; побудова нейронної мережі з оберненим поширенням похибки; методи оцінювання якості прогнозів.

Результати дослідження. Основними результатами дослідження є: отримання прогнозів обсягів продажу фармацевтичних засобів на ринку України; оцінювання якості отриманих прогнозів.

Висновки. Таким чином, запропонована імітаційна модель для побудови прогнозу дає більш якісні результати, порівняно з іншими методами. З отриманих результатів видно, що імітаційна модель дає кращі показники не залежно від кількості кроків на яку будується прогноз. Використання імітаційної моделі спрощує задачу побудови прогнозу, оскільки вказаний підхід, надає системі більшу гнучкість та надає можливість враховувати фактори які раніше не враховувались.

Ключові слова: імітаційна модель; агентний підхід; побудова прогнозу.

В.С. Варава, В.М. Подладчіков

ПЕРЕДБАЧЕННЯ СОНЯЧНОЇ АКТИВНОСТІ НА ОСНОВІ АНАЛІЗУ РАДІОВИПРОМІНЮВАННЯ СОНЦЯ

Проблематика. Прогнозування сонячної активності є дуже важливою задачею в сучасному світі, адже космічна погода дуже сильно впливає як на навколоземні супутники та космічні станції, так і на різні процеси на Землі.

Мета дослідження. Головна мета дослідження: розробити ефективну методику прогнозування сонячної активності на 1, 2 та 3 дні.

Методика реалізації. Для розв'язання поставлених задач використано такі методи: адаптація фільтру Калмана до задачі прогнозування сонячної активності на 1, 2 та 3 дні, метод експоненційного згладжування, операції над матрицями.

Результати дослідження. Основними результатами дослідження є такі: розроблено алгоритм для прогнозування сонячної активності на 1, 2 та 3 дні за допомогою фільтра Калмана, зроблено порівняння з традиційним алгоритмом прогнозування потоків на 1 день, заснованим на припущенні, що приріст потоку на наступний день збігається з поточним приростом.

Висновки. Таким чином, запропоновано методику для прогнозування сонячної активності на 1, 2 та 3 дні. Представлено модель досліджуваного процесу в просторі станів та адаптовано до неї фільтр Калмана. Порівняння цього підходу із традиційним алгоритмом показали, що ця методика є ефективною.

Ключові слова: прогнозування, сонячна активність, потік сонячного радіовипромінювання, індекс F10.7, фільтр Калмана.

О.І. Гур'янов

ПРОГНОЗУВАННЯ РЕЗУЛЬТАТІВ ЕКСПЕРТНОГО ОЦІНЮВАННЯ РІВНЯ ЗНАНЬ ІНОЗЕМНОЇ МОВИ ЗА ДОПОМОГОЮ СУМ ВИПАДКОВИХ ВЕЛИЧИН

Проблематика. Прогнозування ймовірностей отримання студентом кожної можливої оцінки на екзамені на основі моделі, які враховує індивідуальні параметри студента, дозволяє оцінювати рівень підготовки студентів та покращувати методи навчання, а також проводити планування.

Мета дослідження. Головна мета дослідження: створити математичну модель екзамену, що складається з виконання єдиного завдання, наприклад, написання твору або усного мовлення, та оцінюється експертною комісією; знайти ймовірності отримання студентом кожної можливої оцінки з використанням випадкової величини, що визначає якість виконання завдання; побудувати таку випадкову величину, що враховує кількість і грубість допущених помилок; дослідити підходи до знаходження функції розподілу цієї випадкової величини.

Методика реалізації. Для характеризації результатів виконання завдання використовуються зважені суми пуассонівських випадкових величин. Для знаходження точної функції розподілу, знаючи параметри розподілів Пуассона, пропонуються рекурсивний метод, метод з використанням генератрис та метод з використанням характеристичних функцій. Для знаходження апроксимації функції розподілу використовується нормальна апроксимація та метод Монте-Карло. Методи були реалізовані мовою програмування Python.

Результати дослідження. Основним результатом дослідження є створення моделі для визначення розподілу оцінок, які може отримати студент, за певних припущень.Можливості та використання моделі продемонстровані на практичному прикладі за допомогою розробленої програми. Запропоновано та проаналізовано кілька методів знаходження функції розподілу зваженої суми пуассонівських випадкових величин.

Висновки. Висунута модель може бути рекомендована для використання в навчальному процесі, що показує практичний приклад. Вона враховує те, що процес допущення помилок не є детермінованим, тому має аналізуватися за допомогою апарату теорії ймовірностей та математичної статистики, оскільки регресійні моделі є непридатними. Аналіз методів знаходження функції розподілу зваженої суми пуассонівських випадкових величин є важливою задачею, оскільки цей розподіл виникає у деяких застосуваннях, зокрема медичних та страхових, але не є достатньо вивченим. Метод з використанням характеристичних функцій рекомендується для знаходження точних значень та метод Монте-Карло для апроксимації. На основі результатів, отриманих за

моделлю, можна розв’язувати інші задачі, наприклад, знаходження імовірності того, що середній бал студента буде перевищувати задане значення.

Ключові слова: зважена сума пуассонівських випадкових величин; функція розподілу; експертне оцінювання.

М.О. Грицаюк

АВТОМАТИЗОВАНИЙ ВИБІР ПРОГНОЗУЮЧИХ МОДЕЛЕЙ В СППР

Проблематика. Сьогодні існує необхідність побудови адекватних математичних і статистичних моделей для оцінювання коротко- і довгострокових прогнозів з метою підвищення якості рішень, які приймаються на їх основі. У таких поширених програмних додатках, як Eviews та STATISTICA, відсутній автоматизованих вибір кращої моделі, тому розробка такого інструментарію є надзвичайно актуальна.

Мета дослідження. Головною метою дослідження є розробка комбінованого критерію для автоматизованого вибору кращої регресійної моделі, побудованої на основі статистичних даних та розробка програмного забезпечення для виконання необхідних обчислювальних експериментів (мова програмування C#).

Методика реалізації. Для розв’язання поставленої задачі розроблений інтегральний критерій якості моделі, який поєднав в собі критерії якості моделей та оцінок прогнозів. Для реалізації СППР була використана платформа розробки програмних додатків Microsoft VisualStudio 2015.

Результати дослідження. Результатом дослідження є розроблений інтегральний критерій якості моделі та запропоновані моделі для таких фінансово-економічних процесів, як ВВП України, ІСЦ України та грошова маса М3 України, які були обрані за допомогою СППР, в якому реалізований алгоритм автоматичного вибору кращої моделі, який заснований на розробленому інтегральному критерію якості моделі.

Висновки. Таким чином, запропонований інтегральний критерій якості моделі можна використовувати при розробці СППР та інших програм для статистичної обробки даних. Запропоновані регресійні моделі таких фінансово-

економічних процесів, як ВВП України, ІСЦ України та грошова маса М3 України дають високий рівень адекватності оцінок прогнозів, що дає можливість використовувати їх для прийняття рішень у державному управлінні. У розробленій СППР можна розширити функції програми до комерційної системи підтримки прийняття рішень для побудови прогнозуючих математичних моделей на основі статистичних даних.

Ключові слова: інтегральний критерій якості, автокореляційна функція, авторегресія з ковзним середнім, прогнозування, валовий внутрішній продукт, індекс споживчих цін, грошовий агрегат М3.

П. К. Зубарева

ОЦІНКА ВПЛИВУ ДЕПУТАТСЬКИХ ГРУП У ВЕРХОВНІЙ РАДІ УКРАЇНИ НА ПРИЙНЯТТЯ КОЛЕКТИВНОГО РІШЕННЯ ЗА ДОПОМОГОЮ ІНДЕКСУ БАНЗАФА ТА ПОРЯДКОВОГО α -ІНДЕКСУ

Проблематика. Визначення міри вливу фракцій (груп) на прийняття рішення дозволить спрогнозувати найбільш імовірні політичні, економічні, соціальні та інші процеси в державі і на міжнародному рівні; визначити характер політичної системи в цілому; обчислити найкращі стратегії фракцій у передвиборчий період.

Мета дослідження. Проаналізувати методи виміру розподілу влади між гравцями в органах прийняття колективного рішення за умови існування кореляції між їх голосами., виявити параметри впливу на ефективність деяких моделей, що використовуються у виборчому процесі.

Методика реалізації. Аналіз та порівняння існуючих методологій вимірювання впливу фракцій (груп) на прийняття колективного рішення; моделювання підрахунку індексів впливовості на основі чисельного складу фракцій (груп) та їх політичних вподобань щодо утворення коаліцій.

Результати дослідження. Проведено аналіз результатів парламентських виборів у Верховну Раду України VIII скликання за допомогою індексу Банзафа та порядкового α -індексу.

Висновки. Було показано, що вплив фракцій (груп) не завжди залежить від їх частки місць у парламенті. Також було виявлено, що фракції (групи), які займають центричне положення, збільшують свій вплив, в той час як радикальні фракції (групи) зменшують його.

Ключові слова: представницька демократія, колективне рішення, індекс Банзафа, порядковий $\propto$ -індекс.

Дудка Б.Р., Бідюк П.І.

РЕАЛІЗАЦІЯ МЕТОДИКИ ПОБУДОВИ МОДЕЛЕЙ ЧАСОВИХ РЯДІВ

Проблематика. На сьогодні існує широке коло методів прогнозування, які ґрунтуються на різноманітних аналітичних процедурах. Однак, у багатьох випадках вони не дають можливості досягти бажаного результату стосовно якості оцінок прогнозів. В зв'язку з цим виникає проблема організації процесу обробки даних таким чином, щоб отримати кращі оцінки прогнозів в умовах наявності невизначеностей структурного, параметричного і статистичного характеру. Також, дуже поширеною проблемою під час первинної обробки даних є наявність пропусків у даних, яка суттєво погіршує статистичні характеристики моделей та якість оцінок прогнозів.

Мета дослідження. Головна мета дослідження це розробка і реалізація методики аналізу нестаціонарних процесів довільної природи на основі системного підходу, яка передбачає ієрархічний аналіз процесів моделювання та прогнозування, врахування невизначеностей структурного параметричного і статистичного характеру. Запропонувати критерій для автоматизованого вибору кращої моделі та нові обчислювальні схеми побудови прогнозуючих систем включаючи концепцію адаптивного прогнозування та моделювання.

Методика реалізації. Для розв'язання поставлених задач використано такі методи: регресійний аналіз, заповнення пропусків альтернативними методами: середнім значенням, заповнення пропусків оцінками прогнозів з використання прогнозуючих функцій, випадковими величинами з нормального розподілу та комбінування прогнозів за допомогою вагових коефіцієнтів.

Результати дослідження. Розроблена та реалізована методика аналізу нестаціонарних процесів на основі системного підходу, який передбачає ієрархічний аналіз процесів моделювання та прогнозування, врахування невизначеностей структурного параметричного і статистичного характеру. Створений критерій автоматизованого вибору кращої моделі з урахування статистичних характеристик адекватності моделей та якості прогнозу. Сформована та реалізована концепція адаптивного прогнозування та моделювання.

Висновки. Розроблена концепція адаптивного моделювання і прогнозування суттєво відрізняється від відомих системним підходом до розв'язання таких задач і можливістю врахування в процесі формування оцінок прогнозів невизначеностей різного характеру. Розв'язана задача заповнення пропусків даних альтернативними методами. Врахування та обробка невизначеностей забезпечується інструментарієм, який грунтується на методах фільтрації даних, формування структури моделі та оцінювання параметрів моделей-кандидатів. Ефективність застосування концепції проілюстрована на прикладі адаптування структури регресійної моделі. В подальших дослідженнях передбачається застосування запропонованої концепції до розв'язання задач прогнозування конкретних процесів з використанням альтернативних адаптивних методів формування оцінок прогнозів.

Ключові слова: методика адаптивного моделювання і прогнозування, регресійний аналіз, системний підхід до моделювання.

ABSTRACT

Michayluk O.O.

DECISION SUPPORT SYSTEM FOR MODELING AND FORECASTING PROCESSES AT STOCK EXCHANGE

Background. In terms of the dynamic variability of processes at the stock exchange is very important to have the appropriate software, which helped to take the appropriate strategy on time. Existing decision support systems (for instance, Eviews and SPSS) provide opportunities for modeling and forecasting processes of stock pricing, but their use requires knowledge of the internal system of command and takes a long time. It is also important to be able to predict not only future stock price, but also the possible deviation from its forecast. This will improve the reliability of the chosen strategy of behavior on the market.

Objective. The main purpose of the research: building mathematical models of selected stock exchange processes, evaluating forecasts based on these models, the development of software to perform computational experiments.

Methods. To solve the tasks used following methods: analysis methods to stock exchange processes, mathematical auto regression models with moving average and autoregressive conditional heteroscedasticity, the method of least squares (and its recurrent modification) to determine the parameters of the models.

Results. The main results of the research are an information system for analytical modeling and forecasting processes at the stock exchange based on autoregressive models with moving average. In addition, there are results of forecasting selected stock prices and a possible deviation from its forecast.

Conclusions. The main types of processes at the stock exchange are considered. There are founded features of process with trend and heteroscedastic processes modeling. A review of existing methods of modeling and forecasting processes at the stock exchange is completed. Based on modeling and forecasting selected the best models for processes at the stock exchange.

Keywords: stationary processes, integrated processes, heteroscedastic processes, autoregression models with moving average, autoregressive conditional heteroscedastic model, decision support systems.

K. Bohun

MODELING AND FORECASTING GDP UKRAINE THROUGH REGRESSION MODELS.

Background. One of the most important indicators of national accounts is the gross domestic product (GDP). It is a generalized indicator of economic growth, which reflects the outcomes of residents of countries that produce goods and services. With GDP model may determine the future economic development and the factors that affect it.

Objective. The main objective is to investigate the dynamics of Ukraine's GDP for the years 1991-2015. We want to define features of GDP, factors that influence its dynamics, and to find the best model for forecasting.

Methods. To solve this problem used the regression equation. It was researched many models including autoregressive equation of different degrees, auto regression equation of the moving average, auto regression equation with moving average with exogenous factor. Since the evaluation parameters of the equation - a random process, to improve the reliability of statistical parameters were used: the coefficient of multiple determination, the sum of squared errors Durbin-Watson statistic, the average absolute error.

Results. Were representation model for predicting three steps.ARIMA(6,1)is proved best one, the average absolute error as a percentage equal to 1.308944 and 0.004171 coefficient Theil

Conclusions. Forecast GDP characterizing quantitative parameters of economic development and its components by stages of economic circulation, because it reflects the cost of the final results of economic activity over time.

Keywords: GDP, modeling, forecasting, regression equation, statistical parameters.

O.M. Boiko, P.I. Bidyuk

PROCESSES of CREDIT INDIVIDUALS WITH NONLINEAR MODELS AND BAYESIAN NETWORKS

Problems. Properly designed decision support system (DSS) for the simulation and prediction of the creditworthiness of individuals enable consideration of uncertainties probabilistic, statistical and structural types. This leads to better quality of developed models and estimates of forecasts.

The aim of the study. The main purpose of the study: research to develop the concept of lending; analyze and select methods creditworthiness; build mathematical models for the data.

Methods of implementation. To solve the tasks, the following methods: linear regression, decision trees, nonlinear mathematical models and methods for Bayesian analysis.

Research results. The main results of the study are as follows: built linear regression, decision trees, Bayesian networks and nonlinear logistic regression statistics; found quality models and quality predictions for the credit borrowers solvency; analyzed statistical classification criteria for quality analysis customers.

Conclusions. Thus completed a brief overview of modern systems for evaluating financial risks. Analyzed their advantages and disadvantages.

Here are described in detail the methods that can be used for predicting the creditworthiness of individuals: logistic regression, linear regression, decision trees and Bayesian networks. With GeNIe 2.0 and EViews 7.0 program built Bayesian networks, linear and nonlinear logistic regression, and using SPSS constructed decision trees. Recommendations for their practical applicationare given.

Key words: credit risk, commercial bank, logistic regression, Bayesian network, credit borrower, linear regression, decision trees.

O.V. Fomin, N.V. Kuznietsova

BEHAVIOURAL SCORING MODELS OF CREDIT CARDS HOLDERS FOR THEIR SOLVENCY ESTIMATION

Background. Scoring cards based on survival analysis model provide a way for using of time-depending covariates. Such covariates reflect special features of the credit card holders' behavior. This approach provides high quality of developed models as well as trustworthiness of clients' classification and prediction of clients' future actions.

Objective. The main purpose of this article is as follows: to analyze existing methods for estimating credit card holders' solvency, to develop alternative models for behavioral scoring based on survival analysis techniques, to provide stepwise methodology for building such model and to compare performance of proposed models with conventional models based on logistic regression.

Methods. Such general methods as analysis, synthesis, abstraction, modeling, comparing and generalization were employed for solving pointed problems. More specifically methods from survival analysis were used, including proportional hazard models and Kaplan-Meier estimator.

Results. The main results of this study are built models which incorporate time-dependent variables and provide dynamical assessment credit card holders as well as universal methodology in a form of commands from R-language for repeating such results.

Conclusions. Thus, a step-by-step algorithm was proposed for creating behavioral scoring models based on Proportional hazards (PH) and Kaplan-Meier estimator. Such models were built and compared with more orthodox approaches such as logistic regression. Then the superiority of the developed models was shown.

Keywords: Behavioral scoring model, survival analysis, proportional hazards, hazard function, risk management, logistic regression.

A.V. Sliusar, O.A. Husiev, D.S.Dulskyi

WAVELET ANALYSIS OF ELLIOTT WAVES

Background. Properly designed algorithms to detect Elliott Waves and further prediction of the behavior of financial markets make it possible to predict the economy. This leads to better quality of developed models and estimates of forecasts.

Objective. The main purpose of the study is to develop an indicator based on wavelet transform, which will high light and analyze the Elliott wave. To make a research on the Elliott wave with existing indicators; analyze the results and track indicators of wavelet analysis to study Elliott Wave.

Methods. To solve the tasks, the following techniques are used: MACD indicator Hurst and analysis gaps sustainable trend for waves Elliott; discrete wavelet transform filtering and smoothing wave; continuous wavelet transform to highlight the Elliott Wave for sometime; kepstr to analyze the structure and detect Elliott waves and cyclic frequencies.

Results. The main results of the study are: the establishment of the indicator, which allows you to allocate Elliott wave on time series; developed methods for research on structure of waves and the allocation of Elliott wave core components; The algorithm with a phased analysis.

Conclusions. In this work considered the classic approach to identifying Elliott Wave and a new indicator for a clear identification of Elliott waves (compared with the analysis of the original quotes chart). Using the wavelet transformand cepstrum can research the structure of the waves, comparing results with known characteristics of waves in order to detect the end of the cycle and its behavior.

Keywords: wavelet analysis, Elliott Wave, Hurst index, moving average, MACD indicator, cepstrum.

Lohin V. V.

IMPROVING THE EFFICIENCY OF THE ADVERTISING INFORMATION PROVISION BY MASS MAILING

In this paper, considered the problem of improving the efficiency of the mass mailing advertising that is very important because the distribution of advertising is

quite common nowadays, and its implementation without analysis may result in losses for the company instead of profit.

The research described in this article is dedicated to the improving the efficiency of the mass mailing advertising. In particular, described the constructing models of customer response to advertising and choosing the best model for the future creation of an effective strategy for distribution of advertising media.

To solve the tasks, we used the methodology of Data Mining, in particular, was solved binary classification problem using several approaches. During the research were used the following methods: logistic regression, decision trees, neural networks. After modeling of customer responses were made comparing of built models on efficiency.

The main result that was obtained after the research is as follows: after comparing models that were built by predictable profit, it appears that the most effective is model based on the decision tree. When using this model, we can observe the increasing of the efficiency of mailing advertising, compared with the model "Send to all" for 8879 units.

Thus, the research showed that during the promotional mailing is not the best use strategy "Send to all" or priority mailing clients from which the largest revenue was obtained, but strategy, which is based on customer response model based on decision tree.

It is also worth noting that on this theme there is the prospect for further research, because promotional mailing is quite common nowadays. In particular, we can build new classification models, for example, algorithms «k-means», SVM and others. There is also a need to develop automatic selection of optimal partition of the original sample to study and test. Another possible improvements would be improving built models and verification developed models on real data.

E.I.Morylov, M.V. Didkovska

THREE-DIMENSIONAL POINT CLOUD RECONSTRUCTION SYSTEM FROM MULTIPLE IMAGES

Background. Everyday the number of cameras increase ssignificantly and the question of processing the received information for reconstructing three-dimensional objectsusing scanners that costis very high. In addition, some areas can not use scanners or multiple camera ssimultaneously. This problem can be solved by three-dimensional point cloud reconstruction system from multiple images. Therefore, there construction of models using ordinary camera is very important.

Objective. The main purpose: a smart reconstruction of three-dimensional model on a two-dimensional image; propose a modification of the method of reconstruction of three-dimensional models, consider a basic method for reconstructing three-dimensional models.

Methods. Following methods were used to sole this task: vidnostyh approach to assessing the policies and cornering camera; procedures to minimize the reconstruction error; of the fundamental and essential matrices.

Results. The main results of the study are: an algorithm that allows to reconstruct a three-dimensional model on a two-dimensional image using an ordinary camera and based on a preliminary evaluation of all cameras positions and rotations and triangulation algorithm using bundle adjustment; advantage and disadvantages of the presented system were analyzed and justified; reconstruction speed and quality was analyzed.

Conclusions. Thus, theproposedmodificationalgorithmforreconstructingthree-dimensional models on a two-dimensional image, which is the preliminary assessment of cameras positions and rotations through out the reconstructed scene. Results of reconstruction showed that the quality of the reconstructed model has not changed, but the speed of reconstruction has increased.

Keywords: three-dimensional reconstruction, feature point descriptor, image comparision, point cloud model, stereoscopy.

I.V. Pudlo, P.I. Bidyuk

PROFITABILITY PREDICTION OF BANKING PRODUCTS USING SCORING APPROACH

Background. Risk management strategy plays an important role in ensuring the effective functioning of banks in a market economy and under conditions of crisis. Scoring model is a basic indicator of the creditworthiness of potential borrowers. On this basis, the expert makes the final decision about granting of credit.

Objective. Consider common approaches to the evaluation of credit risk. Analyze the related risks in the lending process. Develop application model for revolving cards with grace period, which will monitor the profitability of the product.

Methods. For solving tasks I used method of assessing credit risk by building scoring models. For building a model I used logistic regression, which is connected with its simple interpreting and stability.

Results. The main results of the study are: consideration of general approaches to credit risk evaluation; consideration of related risks involved in granting of credit; development of application scoring model for revolving cards with grace period. In addition, it was shown stability of the model in the future.

Conclusions. This paper examined the problem of credit risk, determined causes and approaches to assessment. Analyzed related risks in the granting of credit process by the example of revolving cards grace with period. It was examined the main stages of scoring model construction: sampling, segmentation characteristics and so on. The result of work is developed application scoring model, which on base of specified parameters, allows to make segmentation of customers in order to prevent bank from losses of revenue.

For developed model GINI index value was 24%. Dynamics of GINI index varies in the acceptable level during the period under review, which testifies the high stability of the constructed model to changes in the data.

Keywords: credit; credit risk; scoring; creditworthiness; revolving card; grace period; regression.

V.Vertiletskij

ANALYSIS FINANCIAL AND ECONOMIC PROCESS BY DSS

Background. Now the question becomes more urgent modeling and forecasting of financial and economic processes in Ukraine, a transition economy, acquire new

features. Methods that exist now, do not always make it possible to obtain satisfactory results. Search for new methods and existing combination to achieve greater accuracy and adequacy of models and that will form the effective management decisions, is very urgent now.

Objective. The main objective - to build a decision support system (DSS), research, analysis of existing and new forecasting models for macroeconomic and financial-economic processes and comparison methods.

Methods. To achieve this task was designed DSS. It was explored and compared many models: autoregressive models of different orders, autoregressive removal of insignificant parameters, was also a model group method of data handling. For comparison, forecasts have been calculated using the constructed model parameters used statistical coefficient of determination, the sum of squared errors Durbin-Watson statistic, the average absolute error.

Results. Were built models for forecasting four steps, the best proved AR (13) as the average absolute percentage error is 2.8957 and the coefficient Theil 0.018085.

Conclusions. Built autoregressive models proved accurate enough when you reach a high order. Developed DSS showed high accuracy of the results compared to commercial products. Later with the help of DSS can explore other processes and forecasts them.

Keywords: Decision support system, modeling, forecasting, regression equation, group method of data handling, statistical parameters.

O. O. Kryukov, K.V. Kozlov, D.G. Didenko

Applying CLOPE algorithm to structuring the domain model of adaptive learning system

Background. Modern adaptive learning systems are extensively used while providing full online courses in various disciplines. This fact leads to increased complexity of domain models and requires significant involvement of experts to build its structure and organization.

Objective. The main purpose of the study is as follows: to offer a method of structuring the domain model of adaptive learning system based on cluster analysis algorithms.

Methods. To solve the problems mentioned the following methods have been used: clustering algorithm of categorical data CLOPE; modified transaction data clustering algorithm CLOPE with lower and upper elements number limit restriction.

Results. The main results of the study are in developing a modified clustering algorithm CLOPE; analysis of building domain model of structure using basic and modified CLOPE algorithms.

Conclusions. Thus, the method of structuring the domain model based on modified clustering algorithm of categorical data CLOPE, which allows to set parameters that limit the upper and lower limits of number of elements in the cluster, was offered.

Partitioning of provided sample of training materials to clusters using basic and modified algorithm CLOPE with different input parameters was performed. It is concluded that the use of the modified algorithm CLOPE is a good fit for tasks where it is necessary to take into account such characteristic as the size of the cluster. It allows to partition the content of the training system into modules of same size and semantic content.

Keywords: adaptive learning system; domain model; categorical data clustering.

B.S. Zhurke, D. G. Didenko

USE OF COMBINED RATINGS IN AGENT MODEL FOR PREDICTIONPHARMACEUTICAL MARKET IN UKRAINE

Background. Properly designed simulation model for predicting the pharmaceutical market of Ukraine, with automatic setup and use of statistical models in combination with neural network. Will provide an opportunity in the short term, to obtain adequate and better prognosis compared with other methods of forecasting.

Objective. The main purpose of the study is as follows: consider the simulation as a tool for forecasting the pharmaceutical market of Ukraine; build

forecasts using statistical models, neural network simulation model; assess the quality of the forecast for each of the methods.

Methods. To solve the problems mentioned the following methods have been used: agent approach to building a simulation model to build a forecast for the pharmaceutical market of Ukraine; forecasting method using autoregressive integrated moving average model; building a neural network inverse spread error; methods of evaluating the quality of forecasts.

Results. The main results of the study are: to obtain forecasts sales of pharmaceuticals in Ukraine; Quality Assessment derived predictions.

Conclusions. Thus, the proposed simulation model for forecasting gives better results than other methods. These results shows that the simulation model gives better performance regardless of the number of steps based on that forecast. Using a simulation model simplifies the task of forecasting, as specified approach gives the system more flexibility and makes it possible to consider factors not previously considered.

Keywords: simulation model; agent approach; construction forecast.

V.S. Varava, V.M. Podladchikov

FORECASTING OF SOLAR ACTIVITY BASED ON THE ANALYSIS OF SOLAR RADIO

Background. Forecasting of solar activity is a very important issue in a modern world, because space weather has a great impact on Earth observation satellites and space stations as well as on different processes on Earth.

Objective. The main objective of the current research is to develop an effective methodology of solar activity forecasting for one, two and three days.

Methods. For addressing respective objections, such methods are applied: the adaptation of Kalman filter to solar activity forecasting for one, two and three days; exponential smoothing method; operations on matrices.

Results. The main results of the research include the development of new algorithm of solar activity forecasting for one, two and three days using Kalman filter; the comparison with a traditional algorithm of forecasting flux for 1 day is

conducted, which is based on the assumption that flux increase for the next day coincides with the current increase.

Conclusions. Thereby, a new methodology is suggested for solar activity forecasting for one, two and three days. The state-space model of examined process is introduced and Kalman filter is adapted to it. The comparison of suggested approach with the traditional one shows that this methodology is effective.

Keywords: forecasting, solar activity, solar radio flux, F10.7 index, Kalman Filter.

O.I. Gurianov

PREDICTING THE RESULTS OF EXPERT ASSESSMENT OF FOREIGN LANGUAGE SKILLS USING SUMS OF RANDOM VARIABLES

Background. Predicting the probabilities of the student getting every possible grade on the exam with a model that takes into account individual parameters of the student provides the possibility to assess the education level and improve the teaching methodsas well as do planning.

Objective. The main purpose of the study is as follows: to create a mathematical model of the exam that consists of accomplishing a single task, i.e. writing an essay or oral speech, and is assessed by an expert panel; to find the probabilities of a student getting each possible consolidated grade given the random variable that determines the quality of task accomplishment; to construct such a random variable taking into account the number and severity of mistakes made; to study the approaches to calculating the cumulative distribution function of this random variable.

Methods. Weighted sums of Poisson random variables are used to characterize the results of task accomplishment. To find the exact cumulative distribution function of this distribution given the Poisson summands parameters, a recursive method, a method based on probability generating functions and a method based on characteristic functions are proposed. To find the approximation to the cumulative distribution function normal approximation and the MonteCarlo method are used.The methods were implemented with Python programming language.

Results. The main result of this study is the development of a model todetermine the distribution of the grades a student can get under certain conditions. The capabilities and use of the model are demonstrated on a practical example using the developed program. Several methods for finding the cumulative distribution function of the weighted sum of Poisson random variables are proposed and discussed.

Conclusions. The proposed model can be recommended for use in the education process, as demonstrated by the practical example. It takes into account that the process of making mistakes is not deterministic, therefore it should be analyzed by means of probability theory and statistics, because regression methods are inapplicable. Analysis of methods of finding the cumulative distribution function of the weighted sum of Poisson random variables is an important task because this distribution occurs in a range of practical applications, such as medicine and insurance, but is relatively little studied. A method using characteristic functions is recommended as the most effective for finding the exact values and the MonteCarlo method as the best approximation. Based on the results provided by the model, other problems may be solved, such as finding the probability of the student's average grades exceeding a given value.

Keywords: weighted sum of Poisson random variables; cumulative distribution function; expert assessment.

M.O. Hrytsaiuk

AUTOMATIZED SELECTION OF FORECASTING MODELS IN DSS

Background. Today there is a need to build adequate mathematical and statistical models to evaluate the short- and long-term forecasts in order to improve the quality of decisions taken on their basis. In such common software applications as Eviews and STATISTICA is no automatic selection of the best model. In this case, the development of such equipment is extremely relevant.

Objective. The main purpose of the researchis to develop combined criteria forautomaticallyselection of best regression model, based on statistical data and

development of software for the necessary computational experiments (programming language C#).

Methods. Integrated quality criterion of the model was developed to solve problem. This criterion combines the quality criteria of the model and assessment of the predictions. To implement DSS was used software applications development platform Microsoft Visual Studio 2015.

Results. The result of the research is designed integrated quality criterion of the model and proposed models for such financial-economic processes, as Ukraine's GDP, CPI Ukraine and M3 money supply of Ukraine. Those models were selected by DSS, which implements an algorithm of automatically selection of best model that is based on developed integrated quality criterion of the model.

Conclusions. The proposed integrated quality criterion can be used in the development of DSS and other software for statistical data processing. The proposed regression models of financial and economic processes, as GDP of Ukraine, Ukraine CPI and M3 money supply of Ukraine provide a high level of adequacy assessments of the predictions, which makes it possible to use for decision making in government management. In the developed DSS can expand the functions of the program to a commercial decision support system to build forecasting models based on statistics.

Keywords: integrated quality criterion, autocorrelation function, autoregressive–moving-average model, forecasting, gross domestic product, consumer price index, M3 money supply.

P. K. Zubareva

IMPACT EVALUATION OF PARLIAMENTARY GROUPS IN THE PARLIAMENT OF UKRAINE TO MAKE A COLLECTIVE DECISION USING THE INDEX BANZAFA AND ORDINAL A-INDEX

Background. Determination of factions' (groups') measure of voting power to make a decision will predict the most likely political, economic, social and other processes in the country and internationally; determine the nature of the political system as a whole; calculate the best faction's strategies in the pre-election period.

Objective.To examine the methods of measuring the distribution of power between players in collective decision-making bodies on the condition of occurring correlation between the voters, to identify the parameters influence the effectiveness of some models used in the election process.

Methods. Analysis and comparison of existing methodologies for measuring the factions' (groups') impact on the adoption of collective decisions; simulation of calculation voting power measures based on factions' (groups') size and their political views on the coalition formation.

Results. The results of parliamentary elections to the Verkhovna Rada of Ukraine VIII convocation were analysed by the Banzhaf power index and ordinal $\propto$ -index.

Conclusions. It is shown that the power of faction (group) might significantly differ from its share of seats. It is found as well that if α -index is used centrist factions (groups) increase their power while factions (groups) with radical positions decrease their power.

Keywords: representative democracy, collective decision, the Banzhaf power index, the ordinal $\propto$ -index.

Dudka B., Bidyuk P.

IMPLEMENTATION OF THE METHODOLOGY FOR CONSTRUCTING TIME SERIES MODELS

Background. Today there exists a wide range of forecasting techniques that are based on alternative analytical procedures. However, in many cases they do not provide a possibility for reaching desirable result regarding quality of forecasting estimates. Thus, the problem arises for proper organization of the data processing procedures so that to improve quality of the forecasts in conditions of availability of structural, parametric and statistical uncertainties. Also widely spread is the problem of availability of the data gaps that leads to substantial degrading of the forecasts estimates.

Objective. The basic purpose of the research is in development and software implementation of properly organized methodology for analysis of the various nature

processes using systemic approach that supposes application of hierarchical procedure to modeling and forecasting as well as taking into consideration possible uncertainties of structural, parametric and statistical nature. It is also necessary to provide a criterion for automatic selection of the best model and new computational schemes for constructing forecasting systems including the concept of adaptive modeling and forecasting.

Methods. The following methods were hired to fulfill the problem statement: regression analysis, filling the data gaps in time series with alternative techniques such as the mean value; the use of forecasting functions; using the pseudorandom numbers from normal distribution and integration of estimates generated with alternative techniques.

Results. The methodology for analysis of nonstationary processes was developed that is based on the system analysis principles. The methodology supposes an application of hierarchical approach to the processes of modeling and forecasting, taking into consideration of possible uncertainties of structural, parametric and statistical nature. A combined criterion was proposed for selection of the best model from generated candidates that takes into consideration model adequacy and forecast quality. A new concept for adaptive modeling and forecasting was proposed.

Conclusions. Thus, a new concept for adaptive modeling and forecasting of the processes of arbitrary nature was proposed that is distinguished from known with the systemic approach to solving such problems and the possibility for taking into consideration of various nature uncertainties. Also the problem of filling the data gaps in time series was solved with alternative techniques. The solution is based on the instrumentation in the form optimal data filtering techniques and appropriate estimation of a model structure and parameters. Effectiveness of the approach was illustrated with adaptive constructing of a regression model. In the future research it is supposed to apply the concept proposed for solving practical problems of quality forecasting with the use of alternative adaptive methods for combining the forecasts estimates.

Keywords: methodology for adaptive modeling and forecasting, regression analysis, systemic approach to modeling.

ВІДОМОСТІ ПРО АВТОРІВ

Бідюк Петро Іванович – д.т.н., професор Інституту прикладного системного аналізу НТУУ «КПІ».

Богун Ксенія Володимирівна – магістр Інституту прикладного системного аналізу НТУУ «КПІ».

Бойко Оксана Михайлівна – магістр Інституту прикладного системного аналізу НТУУ «КПІ».

Варава Владислав Сергійович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Вертелецький Володимир Геннадійович – магістр Інституту прикладного системного аналізу НТУУ «КПІ».

Грицаюк Максим Олегович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Гур'янов Олександр Ігорович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Гусєв Олександр Андрійович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Діденко Дмитро Георгійович – к.т.н., доцент кафедри ММСА Інституту прикладного системного аналізу НТУУ «КПІ».

Дідковська Марина Віталіївна – к.т.н., доцент Інституту прикладного системного аналізу НТУУ «КПІ».

Дудка Богдан Романович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Дульський Дмитро Сергійович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Журке Богдан Сергійович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Зубарева Поліна Костянтинівна – студентка Інституту прикладного системного аналізу НТУУ «КПІ».

Козлов Кирило Валерійович – аспірант Інституту прикладного системного аналізу НТУУ «КПІ».

Крюков Олександр Олександрович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Кузнєцова Наталія Володимирівна – к.т.н., старший викладач Інституту прикладного системного аналізу НТУУ «КПІ».

Логін Вадим Вікторович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Михайлюк Олександр Олегович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Морильов Єгор Ігорович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Подладчіков Володимир Миколайович – д.т.н., професор Інституту прикладного системного аналізу НТУУ «КПІ», м. Київ.

Пудло Ілля Віталійович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Слюсар Андрій Вячеславович – студент Інституту прикладного системного аналізу НТУУ «КПІ».

Фомін Олександр Володимирович – студент Інституту прикладного системного аналізу НТУУ «КПІ».