

Національна академія наук України
Інститут телекомунікацій і глобального інформаційного простору

Кваліфікаційна наукова праця на
правах рукопису

Прим. № 1
004.67:004.89:[351.84:364.3-64](043.3)

ЗАРУДНИЙ ОЛЕКСІЙ БОРИСОВИЧ

ДИСЕРТАЦІЯ

Моделі та методи прогнозування для підтримки прийняття рішень у системі
державного пенсійного забезпечення
05.13.06 – інформаційні технології

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей,
результатів і текстів інших авторів мають посилання на відповідне джерело

_____ О.Б. Зарудний

Науковий керівник
Трофимчук Олександр Миколайович,
член-кореспондент НАН України,
д.т.н., професор

Київ – 2026

Декларація про використання інструментів ШІ

Під час підготовки рукопису автор використовував інструменти штучного інтелекту ChatGPT (OpenAI, модель GPT-5) на етапах підготовки допоміжних текстових матеріалів. Засоби штучного інтелекту були використані під час стилістичного доопрацювання тексту та структуризації у розділах вступу, аналізу літературних даних, методики дослідження.

Перевірка результатів, отриманих із використанням інструментів штучного інтелекту, здійснювалася шляхом ручного редагування тексту з урахуванням наукового змісту роботи та вимог до написання.

Використання інструментів штучного інтелекту не вплинуло на зміст отриманих наукових результатів і не змінило висновків дослідження. Вплив таких інструментів обмежувався допомогою в мовному та структурному доопрацюванні рукопису.

Олексій ЗАРУДНИЙ

АНОТАЦІЯ

Зарудний О. Б. Моделі та методи прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення. – Кваліфікаційна праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за фахом 05.13.06 - Інформаційні технології – Інститут телекомунікацій і глобального інформаційного простору, Національна академія наук України, Київ, 2026.

Дисертаційна робота присвячена вирішенню актуальної науково-прикладної проблеми - розробленню науково обґрунтованих підходів автоматизації підтримки прийняття рішень у системі державного пенсійного забезпечення, основу яких становлять моделі, методи прогнозування процесів різної природи. Науковими результатами є нові моделі високого ступеня адекватності, ансамблі моделей, методи аналізу даних, алгоритми, методи прогнозування інформаційній технології, що призначені до використання у системі державного пенсійного забезпечення. Представлені в роботі моделі та методи прогнозування відрізняються стійкістю, забезпечують вищу точність прогнозів та є адаптивними та гнучкими. Вони доповнюють існуючі підходи, враховуючи зміни демографічної ситуації, вплив екологічних факторів, наслідки війни, зміни у законодавстві, розвиток інформаційних технологій. Їх застосування підвищить якість обслуговування населення, обґрунтованість рішень, що приймаються на різних рівнях у системі державного пенсійного забезпечення.

Ключові слова: математичні моделі, прогноз, державне пенсійне забезпечення, інформаційні технології, інтелектуальний аналіз даних, підтримка прийняття рішень.

ABSTRACT

Zarudnii, O. B. Models and forecasting methods to support decision-making in the state pension system. Qualification paper submitted as a manuscript.

This thesis is submitted for the degree of Candidate of Technical Sciences in the specialty 05.13.06 'Information Technologies' – Institute of Telecommunications and Global Information Space, National Academy of Sciences of Ukraine, Kyiv, 2026.

This dissertation is devoted to addressing a pressing scientific and applied problem-the development of scientifically sound approaches to automating decision support within the state pension system, based on models and methods for forecasting processes of various kinds. The scientific results include new models with a high degree of accuracy, model ensembles, data analysis methods, algorithms, and forecasting methods in information technology, all designed for use in the state pension system. The models and forecasting methods presented in this work are robust, ensure higher forecast accuracy, and are adaptive and flexible. They complement existing approaches by taking into account changes in the demographic situation, the impact of environmental factors, the consequences of war, changes in legislation, and the development of information technology. Their application will improve the quality of service provided to the population and the soundness of decisions made at various levels within the state pension system.

Keywords: mathematical models, forecasting, public pension system, information technology, data analytics, decision support

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові праці, в яких опубліковані основні наукові результати дисертації

Статті у наукових виданнях, включених до категорії «А» Переліку наукових фахових видань України

1. Зарудний О. Б., Коваль Р. Г. Методика прогнозування часових рядів для визначення потреби у видатках на соціальний захист та соціальне забезпечення», *Міжнародний науково-технічний журнал «Проблеми керування та інформатики»*. 2024. Vol. 69, №о. 5. С. 64–77. <https://doi.org/10.34229/1028-0979-2024-5-5>

Статті у наукових виданнях, включених до Переліку наукових фахових видань України

2. Zarudnyi, O., Koval, R. Application of probabilistic and stochastic models and data mining for forecasting the contingent of old age pension recipients in the context of systemic uncertainty. *Technology Audit and Production Reserves*, 2024. Vol. 5. №. 2(79), P. 56–62. <https://doi.org/10.15587/2706-5448.2024.313960> (Scopus)

3. Зарудний О. Застосування інтелектуального аналізу еколого-економічних даних у системі пенсійного страхування та для прогнозування витрат на соціальний захист і соціальне забезпечення. *Екологічна безпека та природокористування*, 2024. Т. 51(3), С. 161–176. UR: <https://es-journal.in.ua/article/view/314099>. <https://doi.org/10.32347/2411-4049.2024.3.161-176>

4. Трофимчук О., Коваль Р., Зарудний О. Методика прогнозування кількості інвалідів з числа санітарних втрат. *Екологічна безпека та природокористування*, 2025. Т. 54 (2). С. 121–135. <https://doi.org/10.32347/2411-4049.2025.2.121-135>

Статті у виданнях, індексованих у базі даних Scopus

5. Zarudnyi, O., Koval, R. Methodology for Constructing an Analytical Subsystem of the Unified Information System of the Social Sphere of Ukraine. *Cybernetics and Systems Analysis*. 2025. Vol. 61. №. 4. P. 487–494. ISSN 1573-8337. <https://doi.org/10.1007/s10559-025-00785-9> (Scopus Q3)

6. Korbicz J., Sholokhov O., Koval R., Zarudnyi O. Application of SAS Text Miner for the analysis of citizens' appeals in the system of social protection and social

security. 8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society. (AISTDS 2024). (Kyiv, 1 Oct. 2024). *Proceedings of the 8th International Scientific and Practical Conference "Applied Information Systems and Technologies in the Digital Society*. 2024. Vol. 3942, P. 113 – 124. ISSN 1613-0073 URL: https://ceur-ws.org/Vol-3942/S_09_Korbicz.pdf, (Scopus)

Статті у наукових виданнях України

7. Корбич Ю., Шолохов О., Зарудний О., Коваль Р. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб населення у соціальному захисті та соціальному забезпеченні. *Прикладні інформаційні системи та технології в цифровому суспільстві*. 2024. Т. 1 № 1. С. 172-184. <https://doi.org/10.17721/3041-2323.2024.172-184>

Наукові праці, які засвідчують апробацію матеріалів дисертації

8. Зарудний О.Б., Коваль Р. Г. Проблеми прогнозування факторів, що впливають на обсяг видатків Пенсійного фонду України Інформаційно-комунікаційні технології для перемоги та відновлення // Колективна монографія за матеріалами XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток» (Київ, 14-15 листопада 2023 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2023. С. 76-79. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf

9. Зарудний О.Б., Коваль Р. Г. Застосування методу сингулярного розкладання у задачі аналізу електронних звернень громадян до Пенсійного фонду України. Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення // Колективна монографія за матеріалами XXIII Міжнародної науково-практичної конференції «Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення» (Київ, 12-13 листопада 2024 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2024. С. 156-159. URL: https://itgip.org/wp-content/uploads/2024/11/2024-11-24_zbirka_all_07_11_2024_148x210.pdf

10. Зарудний О. Б., Коваль Р. Г. Шолохов О. В. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб

населення у соціальному захисті та соціальному забезпеченні Прикладні інформаційні системи та технології в цифровому суспільстві: зб. Тез доповідей і наук. повідомлень учасників VIII Міжнародної науково-практичної конференції «Прикладні інформаційні системи та технології в цифровому суспільстві» (Київ, 01 жовтня 2024 р.) // за заг. ред. В. Плєскач, О. Фєндьо. К.: Київський нац. ун-т ім. Тараса Шевченка, 2024. С. 167-178. URL: https://aistis.knu.ua/wp-content/uploads/2024/10/Text_Book_conference_01.10.2024.pdf

Авторське свідоцтво на твір

11. Зарудний О. Б., Колбун В. А., Малецький О. М., Машкін В. Г., Рябцева Т. Б., Матвєєв С. В., Хандрос Ю. О., Коваль Р. Г., Бондарчук О. М., Педан Ю. А., Павлюков І. О. Комп'ютерна програма «Інтегрована комплексна система Пенсійного фонду України «ІКІС ПФУ», версія 2» №11056; заявл.31.10.2011; опубл. 11.11.2011.

Підручники, посібники, навчальні матеріали

12. Закон України «Про загальнообов'язкове державне пенсійне страхування»: Науково-практичний коментар / Б.О. Зайчук, О.Б. Зарудний, С.Б. Березіна та ін. Київ.: НВП «АВТ», 2005. 1024 с. (Інтегрований навчально-атестаційний комплекс).URL: <https://essuir.sumdu.edu.ua/items/75829b1c-c646-4d79-8e96-5171c0e86c28>

13. Зайчук Б. О., Зарудний О.Б., Березіна С. Б. Пенсійний фонд України. Організаційно-правові та соціально-економічні засади функціонування. . Київ: НВП "АВП", 2006. 298 с.

14. Зайчук Б. О., Зарудний О.Б., Березіна С. Б. Загальнообов'язкове державне пенсійне страхування : навч. посіб. (Інтегрований навчально-атестаційний комплекс). Київ : НВП "АВП", 2005. 256 с.

ПЕРЕЛІК СКОРОЧЕНЬ

- AIC - Akaike Information Criterion – інформаційний критерій Акайке
- ARIMA – autoregressive integrated moving average – авторегресійне інтегроване ковзне середнє)
- BI - Business Intelligence – бізнес-аналітика
- BIC - Bayesian Information Criterion – байєсівський інформаційний критерій
- BP - Breusch–Pagan test – тест Бройша—Пагана
- DW - Durbin–Watson statistic – статистика Дарбіна—Ватсона
- ECE - Expected Calibration Error – очікувана помилка калібрування
- FF- F-statistic - статистика Фішера
- HMM - Hidden Markov Model – прихована марковська модель
- FP – False Positive) — хибно позитивний результат
- FN – False Negative) — хибно негативний результат
- IQR – nterquartile Range міжквартильний розмах
- LB - Ljung–Box Ljung–Box test – тест Льюнга—Бокса
- LightGBM - Light Gradient Boosting Machine
- MAE - Mean Absolute Error – середня абсолютна похибка
- MAPE - Mean Absolute Percentage Error – середня абсолютна процентна похибка
- MPIW - Mean Prediction Interval Width – середня ширина прогнозного інтервалу
- MSE – Mean Squared Error середньоквадратична помилка
- NLP - Natural Language Processing – оброблення природної мови
- PCA - Principal Component Analysis – метод головних компонент
- PICP - Prediction Interval Coverage Probability – ймовірність покриття прогнозного інтервалу
- R^2 - Коефіцієнт детермінації
- RMSE - Root Mean Squared Error – корінь із середньоквадратичної похибки
- sMAPE - Symmetric Mean Absolute Percentage Error – симетрична середня абсолютна процентна похибка
- TN – rue Negative) — істинно негативний результат

TP – True Positive) — істинно позитивний результат

SAS - Statistical Analysis System

SVD - Singular Value Decomposition – сингулярний розклад

VIF – Variance Inflation Factor фактор інфляції дисперсії

XGBoost - Extreme Gradient Boosting

АСОПД-КОМТЕХ - Автоматизована система оброблення пенсійної документації

ДП - Державне підприємство

ЄДРЗО - Єдиний державний реєстр застрахованих осіб.

ЄІССС - Єдина інформаційна система соціальної сфери

ІА - Інтелектуальний аналіз

ІАД - Інтелектуальний аналіз даних

ІКІС ПФУ - Інтегрована комплексна інформаційна система Пенсійного фонду України

ІС - Інформаційна система

ІТ - Інформаційні технології

МН - Машинне навчання

ПФУ - Пенсійний фонд України

СППР - Система підтримки прийняття рішень

ІІ - Штучний інтелект

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ	8
ВСТУП	11
РОЗДІЛ 1 ОСОБЛИВОСТІ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ УКРАЇНИ	
1.1 Загальна характеристика системи державного пенсійного забезпечення України та рішень, що приймаються на її рінях	22
1.2 Аналіз вітчизняного та закордонного досвіду використання інструментів прогнозування у системі пенсійного забезпечення	29
1.3 Використання Інтегрованої комплексної інформаційної системи Пенсійного фонду України (ІКІС ПФУ) для інформатизації процесів прийняття рішень у ПФУ	34
1.4 Урахування невизначеностей як засіб удосконалення підтримки прийняття рішень у системі державного пенсійного забезпечення	42
1.5 Постановка задачі дисертаційної роботи	51
Висновки до розділу 1	53
РОЗДІЛ 2 МЕТОДИЧНІ ПІДХОДИ ДО ПОБУДОВИ МОДЕЛЕЙ І ПРОГНОЗІВ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ	
2.1 Урахування особливостей системи державного пенсійного забезпечення при побудові моделей і прогнозів на певному рівні прийняття рішення	56
2.2 Удосконалення методики використання різномірних даних при побудові моделей і прогнозів у системі державного пенсійного забезпечення	63
2.3 Методи прогнозування контингенту одержувачів пенсій	83
2.4 Інформаційна технологія прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ для прийняття рішень у системі державного пенсійного забезпечення	90
Висновки до розділу 2	93
РОЗДІЛ 3 МЕТОДОЛОГІЯ ПРОГНОЗУВАННЯ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ	

3.1	Методичні підходи урахування особливостей системи державного пенсійного забезпечення при побудові моделей та прогнозів в умовах невизначеності	95
3.2	Обґрунтування застосування математичних моделей різних типів та ансамблів	101
3.3	Методика застосування байєсівських мереж, нечіткої логіки та моделей прогнозування контингенту одержувачів	110
3.4	Методи комбінування прогнозів і багатомодельні підходи	121
3.5	Критерії оцінювання якості прогнозів і точності моделей	125
	Висновки до розділу 3	137
РОЗДІЛ 4	ЗАСТОСУВАННЯ МЕТОДОЛОГІЇ ПРОГНОЗУВАННЯ У СИСТЕМІ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ	
4.1	Прогнозування факторів, що впливають на обсяг видатків Пенсійного фонду України	140
4.2	Методика узгодження прогнозів на різних рівнях прийняття рішень у системі державного пенсійного забезпечення	148
4.3	Інформаційна технологія прогнозування контингенту одержувачів пенсій за віком в умовах невизначеності	162
	Висновки до розділу 4	166
	ВИСНОВКИ	168
	СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	171
	ДОДАТКИ	187
	Додаток А	187
	Додаток Б	188
	Додаток В	190

ВСТУП

Актуальність теми дослідження. Унікальний досвід інформатизації соціальної сфери України, започаткований ще у 90-х роках ХХ-століття розробленням автоматизованих систем для виплати пенсій та допомог спочатку у системі Міністерства праці та соціального захисту населення, а потім у Пенсійному фонді України (ПФУ), не лише забезпечив спроможність та стійкість функціонування соціальної сфери в умовах сучасних викликів, а й заклав передумови для створення Єдиної інформаційної системи соціальної сфери (ЄІССС) Автоматизація ключових процесів, а саме, консультативного обслуговування населення, призначення та виплати пенсій, допомог, субсидій, веденням реєстру застрахованих осіб, планування доходів та витрат, тощо, забезпечила безперебійну роботу установ ПФУ та закладів системи соціального захисту та соціального забезпечення, навіть в умовах війни. Однак, для подолання сучасних викликів недостатньо лише оперативного реагування на події, вирішення точкових задач. Потрібні системні рішення щодо подальшої розбудови соціальної сфери країни, і особливо її провідної ланки – системи державного пенсійного забезпечення. Тому, проблема створення нового аналітичного інструментарію, який враховуючи особливості вітчизняної системи пенсійного забезпечення, забезпечував би ефективність як оперативних, так і стратегічних рішень, в умовах невизначеності, є актуальною.

Питання розроблення моделей та методів прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення, розглядається як вітчизняними та закордонними фахівцями-практиками, так і науковцями. Якщо за кордоном більш актуальними є питання удосконаленню актуарних розрахунків, прогнозування інвестицій пенсійних коштів, тощо, то вітчизняні дослідники більше уваги приділяють розробленню інформаційно-аналітичного забезпечення підтримки прийняття управлінських рішень.

Використанню математичних моделей та інструментів інтелектуального аналізу даних у системах підтримки прийняття рішень приділено увагу в

роботах: Ф. Бернштейна, О. Ф. Волошина, С. О. Довгого, В. Є. Снитюка, В. Г. Тоценка, О. М. Трофимчука, С. Холсаппла, М. З. Швиденка (*технології та системи підтримки прийняття рішень*); В. М. Азарскова, В. Я. Данилова, М. З. Згуровського, Н. Д. Панкратової (*теорія та прикладні методи системного аналізу*); П. І. Бідюка В. С. Степашка, С. Г. Удовенка, (*методи прогнозування*); Є. В. Бодянського, М. М. Корабльова, Ю. В. Крака, В. І. Литвиненка, С. А. Положаєнка (*методи інтелектуального аналізу даних*); Ю. П. Зайченка, Ю. П. Кондратенка, В. В. Кульби, О. І. Міхальова (*методи нечіткого аналізу і моделювання*); Т. Байєса, Дж. Перла, (основи ймовірнісного підходу); О. П. Гожого, Р. Г. Ковалю, О. С. Меньяйленка, С. Ф. Теленика, О. М. Терентьєва, С. В. Цюцюри (*інформаційні технології*), Т. В. Нескородевої, О. А. Павлова (*методи моделювання складних систем*), Д. В. Ланде, М. Бері, С. Аггарвала, Х. До Прадо, Е. Фемеди (*Text Mining*) та інших.

Однак, не зважаючи на значну кількість досліджень, проблема розроблення системної методології залишається невирішеною. Тому, важливою і актуальною науково-прикладною проблемою, що розв'язується в дисертаційній роботі є підвищення ефективності підтримки прийняття рішень у системі державного пенсійного забезпечення в умовах невизначеності за рахунок впровадження сучасних моделей, методів прогнозування, інформаційних технологій у роботу Пенсійного фонду України, установ соціальної сфери різного рівня та підпорядкування.

Зв'язок роботи з науковими програмами, планами й темами. Тема дисертаційної роботи повністю відповідає тематиці наукових досліджень Інституту телекомунікацій і глобального інформаційного простору НАН України. Результати, отримані в ході виконання дисертаційного дослідження знайшли відображення у науково-дослідних роботах: «Розробка та аналіз засобів теоретико-ігрового моделювання стратегій збалансованого технологічного розвитку територій» (номер державної реєстрації 0116U000796), «Розробка інформаційної технології моделювання і прогнозування розвитку соціально-еколого-економічних систем в умовах невизначеності, нестаціонарності та

ризик» (номер державної реєстрації 0121U100132), «Розробка інформаційних технологій та інструментальних засобів моделювання і прогнозування розвитку територій в умовах децентралізації» (номер державної реєстрації 0121U109211), «Математичне моделювання і прогнозування функціонування і розвитку соціально-економічної і транспортної інфраструктури територій і громад» (номер державної реєстрації 0123U104136).

Внесок автора у результати наведених науково-дослідних робіт полягає в розробці моделей, методів прогнозування, інформаційних технологій, призначених для підтримки прийняття рішень у системі державного пенсійного забезпечення, соціального захисту та соціального забезпечення, а також органами державного управління та місцевого самоврядування, що опікуються соціальними питаннями, в умовах невизначеності.

Мета й завдання дисертаційного дослідження. Мета дисертаційної роботи – створення нових моделей, методів прогнозування для виконання системних досліджень, розроблення на їх основі методологічних засад інформаційно-аналітичного забезпечення підтримки прийняття рішень у системі державного пенсійного забезпечення.

Для досягнення поставленої мети в роботі необхідно вирішити такі *завдання*:

- дослідити особливості системи державного пенсійного забезпечення України та прийняття рішень щодо пенсійного забезпечення в умовах невизначеності;
- виконати огляд та аналіз сучасних методів моделювання і прогнозування розвитку системи пенсійного забезпечення;
- удосконалити методику використання різномірних даних, отриманих з різних джерел як для опису та прогнозування процесів, що характерні для системи державного пенсійного забезпечення в умовах невизначеності, так і для забезпечення зворотного зв'язку стосовно ефективності управлінських рішень, що приймаються органами ПФУ;
- розробити системну методологію прогнозування процесів, що характерні для державного пенсійного забезпечення в умовах невизначеності, зокрема,

спричиненої військовим конфліктом;

- побудувати математичні моделі та ансамблі моделей, які описують процеси, характерні для системи пенсійного забезпечення України для прогнозування її розвитку для підтримки прийняття управлінських рішень в умовах невизначеності, зокрема, спричиненої військовим конфліктом;
- розробити та виконати апробацію запропонованих підходів до моделювання та прогнозування розвитку системи державного пенсійного забезпечення України, запропонувати підходи, що забезпечують адекватний опис причинно-наслідкових зв'язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних процесів;
- виконати аналіз адекватності побудованих моделей та сформулювати висновки щодо можливості їх застосування для розв'язання задач підтримки прийняття рішень у системі державного пенсійного забезпечення;
- на основі системного підходу побудувати інформаційну технологію, яка доповнює існуючі, поєднуючи розроблені моделі та методи прогнозування та адаптувати її для подальшого використання у підсистемі підтримки прийняття рішень Єдиної інформаційної системи соціальної сфери.

Об'єкт дослідження – процеси підтримки прийняття рішень, пов'язані із фінансуванням витрат на пенсії у системі державного пенсійного забезпечення.

Предмет дослідження – моделі, методи прогнозування, призначені для підтримки прийняття рішень у системі державного пенсійного забезпечення.

Методи дослідження. Для досягнення цілей роботи застосовано міждисциплінарний підхід, основу якого становлять загальнонаукові підходи та методи системного аналізу, інтелектуального аналізу даних, математичного моделювання, машинного навчання, підтримки прийняття рішень, розроблення інформаційних технологій. Для вивчення системи державного пенсійного забезпечення в умовах невизначеності та для узагальнення теоретичних і методологічних засад дослідження використано методи діалектичного пізнання, системного аналізу, індукції та дедукції, аналізу і синтезу, аналогій і зіставлення, формалізації й моделювання, інтелектуального аналізу даних, сценарний підхід

та елементи теорії прийняття рішень. Для збору та попередньої обробки даних застосовано методи збору та оброблення структурованих та неструктурованих даних, зокрема, великих обсягів, методи системного, інтелектуального й статистичного аналізу, експертне оцінювання. Для побудови моделей і прогнозів – регресійний аналіз, методи аналізу часових рядів, ймовірісно-статистичне моделювання, методи статистичного дослідження, економіко-математичне моделювання, машинне навчання, багатовимірні розподіли. Для розроблення методики застосування створених інформаційних технологій – методи й засоби проєктування та розроблення програмного забезпечення для систем підтримки прийняття рішень, зокрема, прикладне програмування, методи розроблення баз даних та знань.

Наукова новизна отриманих результатів:

вперше:

— розроблено новий метод прогнозування контингенту одержувачів пенсій, який відрізняється стійкістю, високою точністю, адаптивністю, за рахунок поєднання сценарного аналізу, інтелектуального аналізу даних, експертного оцінювання та багатомодельного підходу, який забезпечує врахування як багаторівневості та ієрархічності прийняття рішень (моделі макро- та мікро рівнів), так і умов невизначеності (методика прогнозування кількості новопризначених пенсій за віком у певному році, за якої використовуються часові ряди даних про трудовий стаж осіб-кандидатів, із значною кількістю неімпутованих пропусків, поєднана з використанням експертних суджень для визначення права на призначення пенсій), що підвищує якість інформаційно-аналітичного забезпечення підтримки прийняття рішень у системі державного пенсійного забезпечення;

— розроблено системну методологію побудови прогнозів для підтримки прийняття рішень щодо пенсійного забезпечення населення в умовах невизначеності, основу якої становлять математичні моделі та їх ансамблі, алгоритми машинного навчання та сценарний аналіз, методи опрацювання текстової інформації яка, відрізняється комплексністю, стійкістю результатів,

забезпечує урахування сукупного впливу різних груп чинників в процесі прийняття оперативних та стратегічних рішень, узгодженням результатів прогнозування на різних рівнях управління та етапах прийняття рішень, забезпеченням контролю виконання та оцінювання результативності рішень, забезпечуючи зворотній зв'язок у системі обслуговування населення, що дозволяє підвищити ефективність функціонування системи державного пенсійного забезпечення як складової соціальної сфери України;

удосконалено:

— інформаційну технологію прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ, за рахунок додавання рекомендаційної підсистеми, у якій поєднано використання експертних суджень із алгоритмами штучного інтелекту, що забезпечує прозорість процесу прогнозування, спрощує аналіз отриманих результатів, обґрунтованість рішень в умовах невизначеності, в тому числі, таких, що приймаються оперативно, у умовах війни;

— систему критеріїв оцінювання прогнозів, шляхом поєднання метрик точності кількісних прогнозів, показники статистичної адекватності регресійних та ймовірнісних моделей, критеріїв стійкості оцінок за неповних та/або спотворених даних, критерії верифікації прогнозів за фактичними даними і сценаріями, що забезпечує побудову адекватних моделей досліджуваних процесів та отримання високоякісних оцінок прогнозів

отримали подальший розвиток:

— метод урахування різних груп внутрішніх та зовнішніх чинників у процесі прогнозування потреби у пенсійному забезпеченні, шляхом розширення існуючих підходів опрацювання відкритої інформації та інформації з обмеженим доступом на різних етапах прийняття рішень, що забезпечує поглиблений аналіз предметної області, створюючи передумови для прийняття обґрунтованих управлінських рішень у системі державного пенсійного забезпечення в умовах невизначеності;

— методика адаптивного прогнозування нелінійних процесів різної

природи, характерних для системи державного пенсійного забезпечення за рахунок розширення набору моделей та методів аналізу даних, додавання методики оновлення оцінок за надходження нових даних і коректного поєднання статистичної інформації з експертними судженнями, що забезпечує високу точність прогнозів контингенту одержувачів пенсій та обсягів витрат на пенсійне забезпечення в умовах невизначеності.

Практичне значення отриманих результатів полягає у тому, що результати теоретико-методологічного та емпіричного досліджень доведені до використання в науково-практичній діяльності, а також апробовані та впроваджені за безпосередньою участю автора в практичну діяльність підприємств та установ, що підтверджується довідкою ТОВ «МЕДИРЕНТ», авторським правом на твір – комп'ютерну програму, яка є складовою автоматизованої системи оброблення пенсійної документації. Результати використовуються у освітньому процесі Державного торговельно-економічного університету, що підтверджено відповідною довідкою.

Особистий внесок здобувача. Усі теоретичні та практичні результати, що виносяться на захист, отримані автором самостійно. Пошук та аналіз літературних джерел за тематикою дисертаційного дослідження виконано автором особисто. Наукові положення, висновки, рекомендації, що викладені в дисертації, розроблені особисто здобувачем. У працях, виконаних у співавторстві, автору дисертації належать: постановка задачі дослідження, обґрунтування вибору моделей для прогнозування часових рядів, аналіз отриманих результатів [1]; постановка задачі дослідження, методика застосування багатомодельного підходу для прогнозування контингенту одержувачів пенсій в умовах невизначеності на рівні центрального апарату ПФУ, зокрема, метод, прогнозування кількості новопризначених пенсій за віком протягом певного року, використовуючи в якості вхідних змінних часові ряди даних про трудовий стаж, що містять значну кількість неімпутованих пропусків. Методика ідентифікації та опрацювання невизначеностей, зокрема, пов'язаних із визначенням права особи на призначення пенсії [2]; експертні рекомендації щодо

факторів, вплив яких на кількість санітарних втрат, припущення стосовно співвідношення, різних видів бойових втрат, апробація моделей, експертне оцінювання результатів прогнозування кількості інвалідів з числа санітарних втрат, формулювання висновків [4]; теоретичні та методичні засади інтеграції аналітичної підсистеми до Єдиної інформаційної системи соціальної сфери України [5], методика застосування результатів дослідження онлайн-середовища для аналізу настроїв, очікувань, потреб різних соціальних груп у пенсійному забезпеченні та соціальному захисті для прогнозування витрат на пенсії, прийняття рішень щодо подальшого розвитку системи державного пенсійного забезпечення України [6], постановка задачі дослідження, інтерпретація результатів відповідно до потреб ПФУ у зворотному зв'язку щодо ефективності рішень, що приймаються стосовно пенсійного забезпечення громадян, експертне оцінювання результатів кластеризації текстів та формування лінгвістичних правил, загальний аналіз отриманих результатів, визначення шляхів їх використання у роботі ПФУ [7, 10], визначення загальної проблеми, експертне оцінювання обґрунтованості та повноти відбору факторів, що впливають на обсяг видатків, обґрунтування підходів до побудови прогнозів у системі планування бюджету ПФУ та шляхів їх подальшого використання [8], формулювання задачі дослідження, аналіз отриманих результатів, висновки [9], визначенні загального призначення, принципів побудови Інтегрованої комплексної системи Пенсійного фонду, вимог до її загальних та низки прикладних підсистем [11], формування концепції видання, безпосереднє написання коментарів до статей Закону України «Про загальнообов'язкове державне пенсійне страхування», що стосуються принципів та структури системи загальнообов'язкового державного пенсійного страхування, прав та обов'язків та відповідальності учасників цієї системи, ведення у ній персоніфікованого обліку внесків, організаційних засад функціонування Пенсійного фонду України, суб'єктів накопичувальної системи, а також редагування коментарів до інших статей згаданого Закону, імплементація вмісту видання до використання у Інтегрованому навчально-атестаційному комплексі

[12], структура видання, інтегрування та редагування його розділів, безпосереднє написання розділів, які стосуються загальних засад функціонування Пенсійного фонду України та організації правової роботи у Пенсійному фонді України, адаптація змісту видання до використання у Інтегрованому навчально-атестаційному комплексі [13], безпосереднє написання складових посібника, що стосуються загальних організаційно-правових засад системи загальнообов'язкового державного пенсійного страхування, його соціально-економічної сутності, місця та ролі у системі соціального забезпечення, імплементація змісту видання до використання Інтегрованого навчально-атестаційного комплексу [14].

Апробація результатів дисертації. Результати дисертаційного дослідження були представлені на 3-х міжнародних та всеукраїнських науково-практичних конференціях: XXII Міжнародній науково-практичній конференції «Інформаційно-комунікаційні технології та сталий розвиток» (Київ, 14-15 листопада 2023 р.), XXIII Міжнародній науково-практичній конференції «Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення» (Київ, 12-13 листопада 2024 р.), VIII Міжнародній науково-практичній конференції «Прикладні інформаційні системи та технології в цифровому суспільстві» (Київ, 01 жовтня 2024 р.).

Публікації. Основні наукові результати дисертації опубліковані у 11 наукових працях, у тому числі, 4 статті у фахових виданнях України, 2 статті у закордонних виданнях, що індексуються у наукометричній базі Scopus, з яких одна стаття - у виданні, віднесеному до третього квартилю (Q3), 3 – у збірниках наукових праць, матеріалах і тезах міжнародних і національних конференцій, 1 – у науковому періодичному виданні України. Є 1 авторське свідоцтво на комп'ютерну програму. Додатково, наукові результати дисертації представлені у 3 навчальних посібниках.

Структура та обсяг роботи. Робота складається із вступу, 4 розділів, списку використаних джерел та додатків. Загальний обсяг роботи складає 170 сторінок, з яких основного тексту 161 сторінка, включаючи 31 рисунок, 39

таблиць, 3 додатки. Список використаних джерел налічує 152 найменування на 15 сторінках.

РОЗДІЛ 1

ОСОБЛИВОСТІ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ УКРАЇНИ

1.1 Загальна характеристика системи державного пенсійного забезпечення України та рішень, що приймаються на її рінях

Система державного пенсійного забезпечення є провідною ланкою соціальної сфери України. Вона активно розвивається адаптуючись до сучасних потреб та викликів. Формування системи пенсійного забезпечення України відбувалось у декілька етапів. Наприкінці 80-х років постало питання щодо необхідності реформування існуючої системи пенсійного забезпечення. Аналіз світової практики засвідчив успішність досвіду співпраці державних і недержавних пенсійних фондів, які акумулювали кошти і виплачували пенсії. Саме тому, й було прийнято рішення про створення спеціалізованої фінансової інституції – Пенсійного фонду СРСР. Виплату пенсій передбачалося фінансувати не з державного бюджету, а з Пенсійного фонду. Саме такий підхід фінансування пенсійних виплат і було закладено у Законі СРСР «Про пенсійне забезпечення громадян СРСР», який був затверджений у травні 1990 року і набрав чинності з 1 січня 1991 року [15]. В Україні також не було власної установи, яка б здійснювала перерахування коштів на пенсійне забезпечення, тому, був створити орган, який займався б пенсіями і всіма супутніми питаннями. Радою Міністрів УРСР і Українською Радою профспілок була ухвалена спільна постанова про створення Українського республіканського відділення Пенсійного фонду СРСР. Пенсійний фонд України був створений у 1992 році [16]. 28 січня 1992 року ухвалою постанови Кабінету Міністрів України №39 [17] про створення на базі Українського республіканського відділення Пенсійного фонду СРСР був створений Пенсійний фонд України, а на базі філій, відділень – управлінь в Автономній Республіці Крим, областях і місті Києві (управління Пенсійного фонду у місті Севастополі було утворено

згідно з Постановою Кабінету Міністрів України №о 534 від 16.09.92 [18] на базі міського відділу Кримського республіканського управління Фонду). Урядовою постановою був затверджений перший склад правління Пенсійного фонду та перше положення про Фонд, на підставі якого він почав діяти. До цього ж часу виконання функцій, покладених на Пенсійний фонд, у районах і містах продовжували виконувати уповноважені відповідних обласних управлінь. А ще за кілька тижнів щойно утворене правління Пенсійного фонду України затвердило власні нормативні документи: інструкцію зі сплати збору, положення про управління Пенсійного фонду в АР Крим, областях та м. Київ. Тобто, на початку 1992 року законодавча база функціонування Пенсійного фонду незалежної держави була повністю сформована. Вітчизняна система пенсійного забезпечення [16] суттєво відрізняється від тих, що прийняті у розвинених країнах - її основу становить солідарна система. Однак, система державного пенсійного забезпечення постійно удосконалюється, пройшла декілька етапів реформ, продовжує трансформуватись і сьогодні.

Змінюється модель формування бюджету ПФУ, система призначення та виплати пенсій. Ухвалення Закону України «Про загальнообов'язкове державне пенсійне страхування» [19] та Закону України «Про недержавне пенсійне забезпечення» [20] було проголошено перехід до трирівневої моделі: солідарної системи загальнообов'язкового державного пенсійного страхування (перший рівень), накопичувальної системи загальнообов'язкового державного пенсійного страхування (другий рівень) та системи недержавного пенсійного забезпечення (третій рівень). Із набранням чинності цими законами з 1 січня 2004 року в системі запроваджено персоніфікований облік страхових внесків [21, 22]. Саме з цієї дати відомості про страховий стаж застрахованих осіб додані у відповідну базу даних, розпорядником якої є Пенсійний фонд, ця інформація постійно оновлюється, а згодом використовується для призначення пенсій. Однак, дані персоніфікованого обліку за період до 2004 року у значній кількості застрахованих осіб, внесені не були, тому є значна кількість пропусків, які можна заповнювати тільки за наявності документів про стаж та заробітну плату.

Сьогодні повноцінно функціонують два рівні, а запровадження накопичувальної складової поки опрацьовується. Система державного пенсійного забезпечення перебуває у стані структурної трансформації, і тому, будь-який аналітичний інструментарій, що створюється для неї, повинен бути придатним до роботи не лише з поточною, а й з майбутньою конфігурацією системи. Однак, державне пенсійне забезпечення було і залишається базовою, а для більшості населення й єдиною складовою пенсійного забезпечення. Тому, питання ефективного використання коштів, які будуть акумульовані у системі державного пенсійного забезпечення, прогнозування потреби у фінансуванні пенсій є актуальною задачею. Її розв'язання потребує розроблення специфічного прогнозного інструментарію, сучасних інформаційних технологій, що відповідають потребам вітчизняної системи пенсійного забезпечення, сприяють підвищенню ефективності рішень, що приймаються, створюють передумови для запровадження оновленої системи актуарних розрахунків як важливого елемента трирівневої пенсійної системи [22, 23, 24].

Учасниками системи є застрахована особа, страхувальник (роботодавець або сама особа за договором про добровільну участь) [16, 19, 20], ПФУ як страховик, а також органи соціального захисту населення, органи державного управління та місцевого самоврядування, що ухвалюють рішення, які впливають на обсяг соціальних зобов'язань. ПФУ є центральним органом виконавчої влади, діяльність якого спрямовується і координується Кабінетом Міністрів України через Міністра соціальної політики, сім'ї та єдності, що реалізує державну політику з питань пенсійного забезпечення та ведення обліку осіб, які підлягають загальнообов'язковому державному соціальному страхуванню [16, 22]. ПФУ має чітку трирівневу ієрархічну організаційну структуру: центральний апарат, головні управління в областях і місті Києві, а також сервісні центри та відділи обслуговування громадян на місцях. Ця організаційна структура відовідає ієрархії процесів прийняття рішень [13, 14, 16]. Тобто, система прийняття рішень ПФУ є багаторівневою та ієрархічною.

Таким чином, об'єктом дослідження є процеси підтримки прийняття рішень, пов'язані із фінансуванням витрат на пенсії у системі державного пенсійного забезпечення, а предметом - моделі та методи прогнозування, призначені для підтримки прийняття рішень у цій системі.

Тобто, актуальним є питання розроблення аналітичного та прогнозного інструментарію, що дозволяє врахувати особливості прийняття рішень у системі державного пенсійного забезпечення у складі Єдиної інформаційної системи соціальної сфери, за використання існуючої апаратної платформи, був гнучким та адаптивним, відповідав би сучасним вимогам щодо збору, оброблення значних обсягів структурованої та неструктурованої інформації, захисту даних, швидкості їх оброблення та якості отриманих прогнозів [5, 25-27].

Ключовими для системи державного пенсійного забезпечення є дві групи задач, які відрізняються і за предметом рішення, і за характером необхідного прогнозного забезпечення:

- задачі прийняття рішень щодо призначення, перерахунку та виплати пенсії конкретній застрахованій особі (мікрорівень);
- задачі прийняття рішень щодо фінансування витрат на виплату пенсій, тобто формування та виконання бюджету ПФУ (макрорівень).

На мезорівні головні управління ПФУ в регіонах визначають потребу в коштах для своєї території, планують навантаження на сервісні центри та оцінюють регіональні особливості динаміки контингенту одержувачів пенсій. Виокремлення мезорівня має практичне значення: показники, що на макрорівні є агрегованими, на мезорівні набувають регіональної диференціації, яка, зокрема, в умовах воєнного стану, внутрішнього переміщення населення та тимчасової окупації окремих територій стає особливо важливою.

Принципово, що на різних рівнях відрізняється не лише масштаб, а й сама природа задачі прогнозування. На мікрорівні прогнозується подія стосовно окремої особи - настання права на призначення пенсії, тому природною формою результату є оцінка ймовірності, а вхідними даними - індивідуальні кількісні та якісні характеристики застрахованої особи: страховий стаж, сума страхових

внесків, стать, вік, наявність інвалідності, пільг, особливих умов праці. На макрорівні прогноуються агреговані показники - чисельність контингенту одержувачів та обсяг видатків, тому, доцільно використовувати моделі за часовими рядами, регресійні моделі за групами чинників і сценарний аналіз. Узагальнену характеристику задач прогнозування для різних рівнів прийняття рішень у системі державного пенсійного забезпечення наведено в табл. 1.1.

Таблиця 1.1 - Характеристика задач прогнозування для рівнів прийняття рішень у системі державного пенсійного забезпечення

Характеристика задач	Рівень прийняття рішень		
	Мікрорівень	Мезорівень	Макрорівень
Суб'єкт рішення	сервісний центр, відділ обслуговування, фахівець з призначення пенсій	головне управління ПФУ в області, м. Києві, м. Севастополі	центральный апарат ПФУ, органи державної влади
Основна задача	призначення, перерахунок, виплата пенсії застрахованій особі; консультування	визначення потреби регіону в коштах; планування навантаження на органи обслуговування	формування та виконання бюджету ПФУ; обґрунтування трансфертів; оцінювання наслідків змін законодавства
Об'єкт прогнозування	настання права на призначення пенсії за віком; очікуваний розмір пенсії	чисельність та структура контингенту одержувачів пенсій у регіоні	обсяг видатків на пенсійне забезпечення; надходження страхових внесків; загальна чисельність одержувачів
Основні джерела даних	реєстр застрахованих осіб, документи про стаж, персоніфікований облік	база даних одержувачів пенсій, регіональна статистика, дані про внутрішньо переміщених осіб	офіційна статистика, агреговані дані реєстрів, макроекономічні прогнози
Форма прогнозного результату	оцінка ймовірності, індивідуальна рекомендація	точковий та інтервальний прогноз, регіональні сценарії	сценарні прогнози, інтервальні оцінки
Горизонт прогнозування	поточний рік, 1–5 років	1–3 роки	1–3 роки (бюджетний цикл), до 10 років (стратегічні рішення)

Ієрархічність передбачає не просто співіснування рівнів, а необхідність узгодження результатів між ними. Прогноз чисельності новопризначених пенсій, отриманий на макрорівні як агрегована оцінка, має бути сумісним із сумою прогнозів, отриманих на мезорівні, а останні - узгоджуватися з індивідуальними оцінками ймовірності настання права на пенсію. Водночас необхідний і зворотний зв'язок: результати обслуговування населення, звернення громадян, виявлені під час консультування типові проблеми є джерелом інформації про якість ухвалених на вищих рівнях рішень і підставою для їх коригування. Схему ієрархії задач прийняття рішень та інформаційних потоків між рівнями прийняття рішень у системі державного пенсійного забезпечення представлено на рис. 1.1.

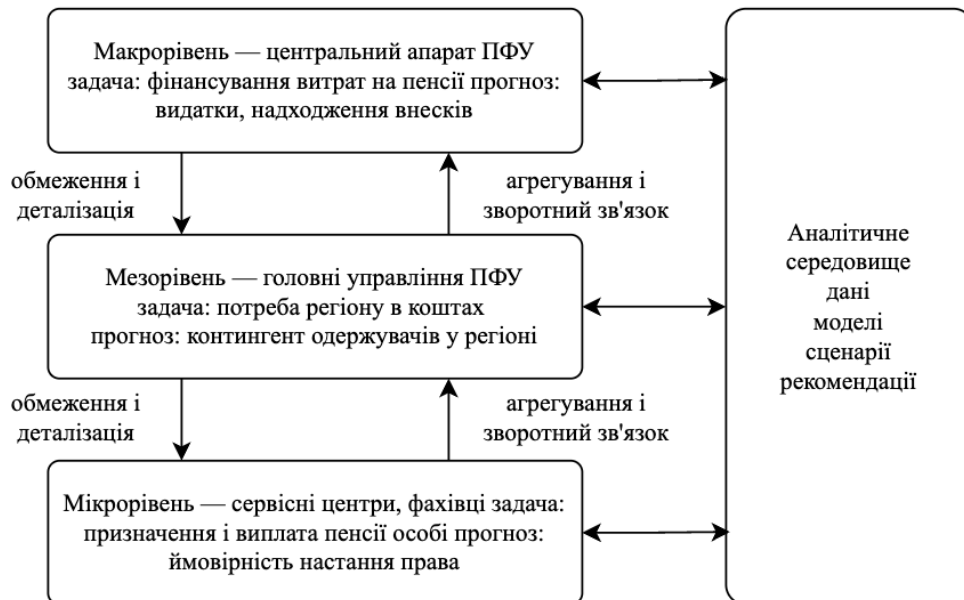


Рисунок 1.1 - Ієрархія задач прийняття рішень у системі державного пенсійного забезпечення та інформаційні потоки між рівнями

Аналіз задач прийняття рішень дав можливість виокремити особливості процесів, прогнозування розвитку яких має бути описано математичними моделями. Ці процеси є переважно нелінійними, наприклад, залежність обсягу видатків від чинників, що його визначають, не зводиться до лінійної комбінації, а сама структура залежності змінюється в часі внаслідок змін законодавства.

Вони розвиваються під сукупним впливом різнорідних груп чинників - соціально-економічних, демографічних, екологічних та суспільно-політичних, які до того ж взаємно залежні. Перебіг вказаних процесів відбувається в умовах невизначеностей різних типів. Крім того, вони є нестаціонарними: вплив воєнної агресії, міграційних процесів, змін статево-вікової структури населення призводить до структурних змін у часових рядах, тому, традиційні методи екстраполяції втрачають прийнятну точність.

Окремо слід виділити особливість, яка не характерна для більшості закордонних пенсійних систем і зумовлює особливості пропонування у роботі підходів. Дані персоніфікованого обліку про страховий стаж є повними лише за період після 2004 року; за попередні періоди трудової діяльності часові ряди містять значну кількість пропусків, які принципово не можуть бути заповнені жодною процедурою імпутації, оскільки страховий стаж підтверджується виключно документально і не може бути «відновлений» статистичною оцінкою. Отже, задача прогнозування контингенту одержувачів пенсій за віком є задачею прогнозування за свідомо неповними даними, де відсутні значення не є випадковими пропусками у статистичному сенсі, а відображають реальну невизначеність правового статусу особи.

Отже, як показало виконане інформаційне дослідження, з огляду на очікувані зміни у пенсійній системі, наслідки військової агресії, демографічні зміни, більше уваги у системі підтримки прийняття рішень державного пенсійного забезпечення, має бути приділено не лише плануванню бюджету, а прогнозуванню видатків, виходячи із всебічного дослідження ситуації, урахування сукупності соціально-економічних, суспільно-політичних, екологічних, демографічних чинників, тощо. Мають бути розглянуті різні сценарії розвитку подій, які відповідно, потребуватимуть різних підходів щодо побудови прогнозів, які мають створювати підґрунтя для прийняття управлінських рішень. Крім того, ситуація військового конфлікту значно ускладнює застосування таких традиційно використовуваних підходів, як методи екстраполяції, статистичного моделювання, тощо [11, 28, 29]. Фактично,

стало неможливо використовувати ретроспективні тенденції для дослідження поточної ситуації і побудови прогнозів. Підходи до прогнозування контингенту одержувачів пенсій, яке базуються на використанні переважно демографічних даних стає практично неможливо застосувати. Традиційно використовувані методи потребують значної кількості статистичної інформації, збір якої у повному обсязі у період війни практично не можливий, або обмежений, наприклад окремими регіонами чи громадами. Потрібні нові моделі та методи, що забезпечать отримання демографічних прогнозів, виконання актуарних розрахунків, визначення контингенту одержувачів пенсій, в умовах системної невизначеності.

1.2 Аналіз вітчизняного та закордонного досвіду використання інструментів прогнозування у системі пенсійного забезпечення

У більшості робіт як закордонних так і вітчизняних вчених, системи соціального захисту та соціального забезпечення, пенсійного страхування та пенсійного забезпечення розглядаються як окремі системи. Відповідно, відрізняється й система прийняття рішень. Зокрема, вони не враховують особливості накопичення даних про застрахованих осіб у нпшій країні, наслідки військового конфлікту, тощо. Не розглядається в них і вплив ситуації військового конфлікту, коли різко зростає як кількість осіб, що потребують соціальної підтримки, прискорюється міграційний рух населення, змінюється статево-вікова структура. Не приділено увагу й питанням урахування майбутніх соціальних витрат при формуванні державного бюджету та бюджетів територіальних громад, а також бюджету ПФУ.

Закордонні системи актуарних розрахунків потребують адаптації до вітчизняних умов. Хоча, наявні методики виконання актуарних розрахунків та актуарних прогнозів [30-35], забезпечують прийнятні результати у багатьох країнах [36-38], наприклад, як такі, що розроблені Світовим банком [39], використовуються у Європейському співтоваристві, орієнтовані більше на

накопичувальну пенсійну систему. Тому й такий закордонний досвід потребує адаптації то вітчизняних умов, як щодо урахування впливу різних груп чинників, які впливають на фінансову рівновагу пенсійної системи, найбільш вагомими з яких, на думку фахівців [40, 41], є макроекономічні, демографічні та організаційно-правові, так і щодо урахування особливостей прийняття рішень і системі державного пенсійного забезпечення. У світовій практиці для моделювання та прогнозування розвитку пенсійних систем використовується кілька класів інструментів. Узагальнену характеристику класів закордонних інструментів наведено в табл. 1.2.

Таблиця 1.2 - Класи інструментів моделювання розвитку пенсійних систем, що використовуються у закордонній практиці [42-52]

Клас інструментів	Призначення	Рівень моделювання	Основні вимоги до даних	Обмеження щодо застосування в Україні
актуарно-бюджетні моделі міжнародних організацій	оцінювання довгострокової стійкості пенсійної системи	макрорівень	демографічні проєкції, стабільні макроекономічні тренди, повні ряди за тривалий період	розраховані на усталену економіку; чутливі до якості демографічних проєкцій, побудова яких в умовах війни ускладнена
мікросимуляційні моделі	прогнозування настання права та суми виплат окремим особам з подальшим агрегуванням	мікрорівень	повні дані про застраховану особу за весь період діяльності	дані персоніфікованого обліку в Україні повні лише з 2004 року; пропуски не підлягають імпутації
моделі інвестування пенсійних активів	прогнозування дохідності та ризиків інвестицій накопичувальної складової	макро- та рівень інвестиційного портфеля	дані фінансових ринків, історія дохідності інструментів	накопичувальна складова в Україні фактично не запроваджена
демографічні і сценарні моделі	побудова варіантів розвитку демографічної ситуації як входу для пенсійних моделей	агрегований	достовірні, актуальні статистичні дані	відсутність достовірного обліку населення в умовах воєнного стану та внутрішнього переміщення

Найбільш поширеними є актуарні та бюджетні моделі, розроблені міжнародними організаціями і призначені для оцінювання довгострокової фінансової стійкості пенсійної системи. Прикладом є моделі, розроблені Світовим банком [39], які використовуються більш ніж у 90 країнах світу. Такі моделі мають детермінований, когортний характер [39], ґрунтуються на демографічних проєкціях і наборі макроекономічних припущень, а їх результатом є довгострокові (на 50–75 років) проєкції надходжень, видатків та дефіциту пенсійної системи. Другим класом є моделі мікросимуляційного типу, які реалізують прогнозування на рівні окремих осіб або домогосподарств з подальшим агрегуванням результатів. Такі моделі використовують у країнах, де накопичено тривалі й повні індивідуальні дані про трудову біографію застрахованих осіб [43, 44]. Третій клас становлять моделі, орієнтовані на прогнозування результатів інвестування пенсійних активів, актуальні насамперед для систем із розвиненою накопичувальною складовою [46, 47].

Спільною ознакою розглянутих інструментів є те, що вони розраховані на умови, коли доступна повна, регулярно оновлювана статистична інформація, а досліджувані процеси не мають різких структурних зламів. Саме тому закордонний досвід потребує адаптації до вітчизняних умов з урахуванням впливу різних груп чинників, які визначають фінансову рівновагу пенсійної системи. За узагальненням фахівців [3, 40, 41], ці чинники доцільно об'єднати у три групи - макроекономічні, демографічні та організаційно-правові, а для урахування особливостей вітчизняної системи пенсійного забезпечення, перелік розширити включивши екологічні чинники та такі, що описують вплив військового стану.

Виконаний аналіз дає підстави виокремити особливості вітчизняної системи, сукупність яких не має прямого аналога і це унеможливорює безпосереднє перенесення закордонних рішень. Це визначено низкою причин.

Однією з них є відсутність даних про страховий стаж за періодами. Оскільки персоніфікований облік застрахованих осіб як обов'язковий запроваджено лише з 2004 року, індивідуальні дані про трудову діяльність

поділяються на дві принципово різні за повнотою частини. Відомості за період до 2004 року потребують документального підтвердження, а їх відсутність в інформаційній системі не означає відсутності стажу. Це виключає застосування мікросимуляційних моделей у їх звичайному вигляді.

Інша - незавершеність структурної реформи. Одночасно існують діюча солідарна система, фактично не запроваджений другий рівень і добровільна складова. Аналітичний інструментарій має бути придатним для роботи як з поточною, так і з майбутньою трирівневою конфігурацією.

Наявність значної кількості пільгових та спеціальних умов призначення пенсій, зокрема за особливі умови праці, за вислугу років, зі зменшенням пенсійного віку, а також диференційовані вимоги щодо наявності страхового стажу для виходу на пенсію у 60, 63 та 65 років [16, 19], робить право особи на призначення пенсії окремим об'єктом прогнозування, а не детермінованою функцією віку.

Можливість набуття страхового стажу за договором про добровільну участь, тобто «докупівлі» стажу, якого не вистачає [16, 19], перетворює момент реалізації права на пенсію у ймовірнісну подію, залежну від поведінкового вибору особи [2, 40].

У період воєнної агресії різко зросла кількість осіб, які потребують соціальної підтримки, зокрема одержувачів пенсій по інвалідності [4, 41]; прискорюється міграційний рух населення; змінюється статево-вікова структура; втрачається можливість збору статистичної інформації в повному обсязі, а частина даних доступна лише за окремими регіонами або територіальними громадами.

Крім того, унікальний досвід інформатизації соціальної сфери, започаткований ще у 1990-х роках розробленням автоматизованих систем виплати пенсій та допомог, який заклав передумови для створення Єдиної інформаційної системи соціальної сфери [5, 16, 23, 25], передбачає те, що нові моделі й методи мають розроблятися не «з нуля», а з урахуванням наявної програмно-апаратної платформи, рівнів системи та процедур прийняття рішень.

У більшості робіт як закордонних, так і вітчизняних авторів інформаційні системи соціального захисту та соціального забезпечення, пенсійного страхування та пенсійного забезпечення розглядаються як окремі системи. Питання їх узгодженого функціонування у складі єдиного інформаційного простору соціальної сфери досліджені недостатньо [5, 24, 25, 27].

Не розглянуто й в роботах закордонних фахівців й вплив ситуації військового конфлікту на процеси прогнозування у системі пенсійного забезпечення, хоча саме він створює найбільш складний для моделювання тип невизначеності. Прогнозування контингенту одержувачів пенсій, побудоване переважно на демографічних даних, у цих умовах стає практично неможливим, оскільки традиційні методи потребують значних обсягів статистичної інформації, збір якої в повному обсязі у період війни обмежений.

Недостатньо уваги приділено врахуванню майбутніх соціальних витрат при формуванні державного бюджету, бюджетів територіальних громад та бюджету Пенсійного фонду України, тобто зв'язку прогнозних оцінок з реальним бюджетним процесом [53].

Не сформовано системної методології, яка поєднувала б задачі мікро- та макрорівнів, забезпечувала узгодження прогнозів за ієрархією управління та зворотний зв'язок від системи обслуговування населення щодо результативності ухвалених рішень. Незважаючи на значну кількість досліджень за окремими напрямками, проблема розроблення такої системної методології залишається невирішеною. Потребує уваги й проблема надання рекомендацій та зворотного зв'язку як у процесі прийняття індивідуальних рішень щодо застрахованих осіб, так і щодо рішень, які приймаються на мезо- та макрорівнях.

Отже, актуальним є розроблення інформаційно-аналітичного забезпечення, яке могло б бути використане у складі Єдиної інформаційної системи соціальної сфери [5, 25-27, 54] з використанням існуючої апаратної платформи, було гнучким та адаптивним, відповідало сучасним вимогам щодо збору та оброблення значних обсягів структурованої і неструктурованої

інформації, захисту даних, швидкості їх оброблення та якості отриманих прогнозів.

1.3 Використання Інтегрованої комплексної інформаційної системи Пенсійного фонду України (ІКІС ПФУ) для інформатизації процесів прийняття рішень у ПФУ

Досвід інформатизації соціальної сфери України, започаткований у 90-х роках ХХ століття, є унікальним. Спочатку автоматизовані системи для виплати пенсій та допомог створювалися в системі Міністерства праці та соціального захисту населення, а згодом - у ПФУ. Автоматизація ключових процесів, а саме консультативного обслуговування населення, призначення та виплати пенсій, допомог і субсидій, ведення реєстру застрахованих осіб, планування доходів та витрат, забезпечила безперебійну роботу установ ПФУ та закладів системи соціального захисту навіть в умовах війни [16, 23], а також заклала передумови для створення Єдиної інформаційної системи соціальної сфери (ЄІССС) [23-27].

Сучасний ПФУ є потужною організаційною структурою, яка має чи не одну з найпотужніших апаратно-технологічних платформ, що автоматизує практичні всі напрямки роботи системи державного пенсійного забезпечення - Інтегровану комплексну інформаційну систему (ІКІС ПФУ) [11, 23].

Система пенсійного забезпечення однією з перших в Україні отримала власну автоматизовану систему оброблення пенсійної документації (АСОПД-КОМТЕХ) ще у 90-х роках ХХ століття. Була створена розгалужена багаторівнева система, програмно-технічна архітектура, якої ефективно виконувала функції з пенсійного забезпечення населення [54, 55]..

Згодом, АСОПД-КОМТЕХ була розвинута у потужну автоматизовану інформаційну систему – Інформаційно-комунікаційну систему Пенсійного фонду України (ІКІС ПФУ), що забезпечила в повному обсязі виконання функцій з пенсійного забезпечення населення, державного пенсійного страхування, консультативної допомоги населенню та бізнесу. Саме в цій системі було вперше

в Україні запроваджено обмін даними між базами даних Пенсійного фонду України та інших державних установ, підприємств та організацій, забезпечено автоматизований доступ до даних державних реєстрів для використання у системі пенсійного забезпечення, налагоджена співпраця з банківським сектором, органами місцевого самоврядування, підприємствами та установами.

Система забезпечує в повному обсязі виконання всього комплексу робіт з зберігання та оброблення великих обсягів даних, виконання функцій призначення, виплати пенсій, ведення бази даних застрахованих осіб, формування звітності про використання коштів Фонду, планування його бюджету на наступний рік, тощо (рис. 1.2).

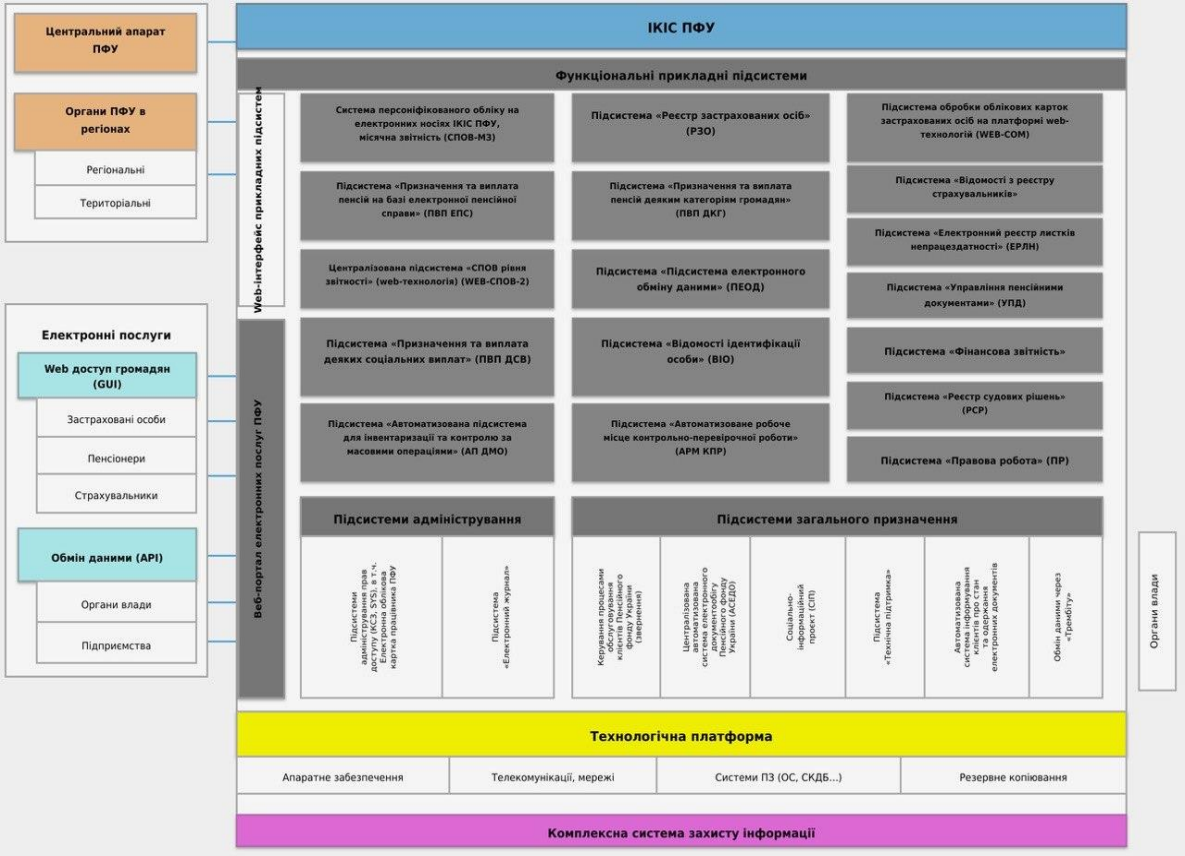


Рисунок 1.2 - Загальна архітектура ІКІС ПФУ [55]

ІКІС ПФУ призначена для уніфікації, зберігання, використання та доступу уповноважених осіб ПФУ та авторизованих користувачів інших підприємств, установ та організацій до інформаційних ресурсів, що накопичуються в ході діяльності ПФУ, відповідно до чинного законодавства, що регулює діяльність ПФУ.

ІКІС ПФУ здійснює автоматизовану обробку, передачу, зберігання та використання інформації, розпорядником якої відповідно до Закону України «Про загальнообов'язкове державне пенсійне страхування» та іншої нормативно-правової бази є Пенсійний фонд України. Основними функціональними задачами, автоматизацію яких забезпечує сьогодні ІКІС ПФУ є збирання, акумуляція та витрати пенсійних коштів, персоніфікований облік, призначення, нарахування та виплата пенсій, страхових виплат, житлових субсидій та пільг на оплату житлово-комунальних послуг тощо, облік страхувальників, застрахованих осіб та пенсіонерів; забезпечення обслуговування відвідувачів установ Пенсійного фонду України; автоматизація діловодства та контроль виконання бюджету ПФУ, бухгалтерський облік, фінансові операції, кадрове забезпечення; інформування громадян; оцифрування архівів пенсійних справ та зберігання електронних документів; аналіз та прогнозування фінансових показників діяльності ПФУ, контингенту одержувачів пенсій, субсидій, застрахованих осіб, тощо. Концептуальна структура ІКІС ПФУ представлена на рис. 1.3.

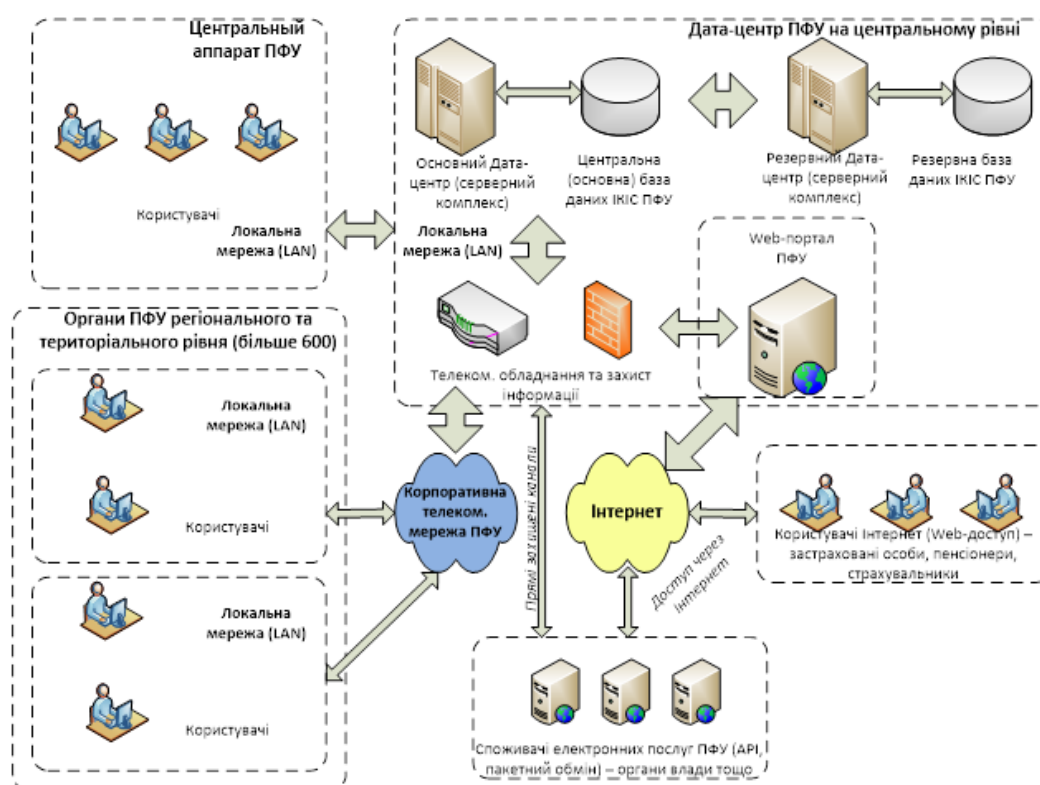


Рисунок 1.3 - Концептуальна структура ІКІС ПФУ [54, 55]

Архітектура ІКІС ПФУ відповідає організаційній структурі ПФУ, ієрархії підрозділів та складається із взаємопов'язаних технологічних рівнів, які являють собою сукупність локальних обчислювальних мереж управлінь і відділень Пенсійного фонду України, що об'єднані в суцільне територіально розподілене комунікаційне середовище за допомогою мережі оператора передачі даних. Розміщення технологічних рівнів ІКС повторює організаційно-територіальну структуру ПФУ (рис. 1.4).

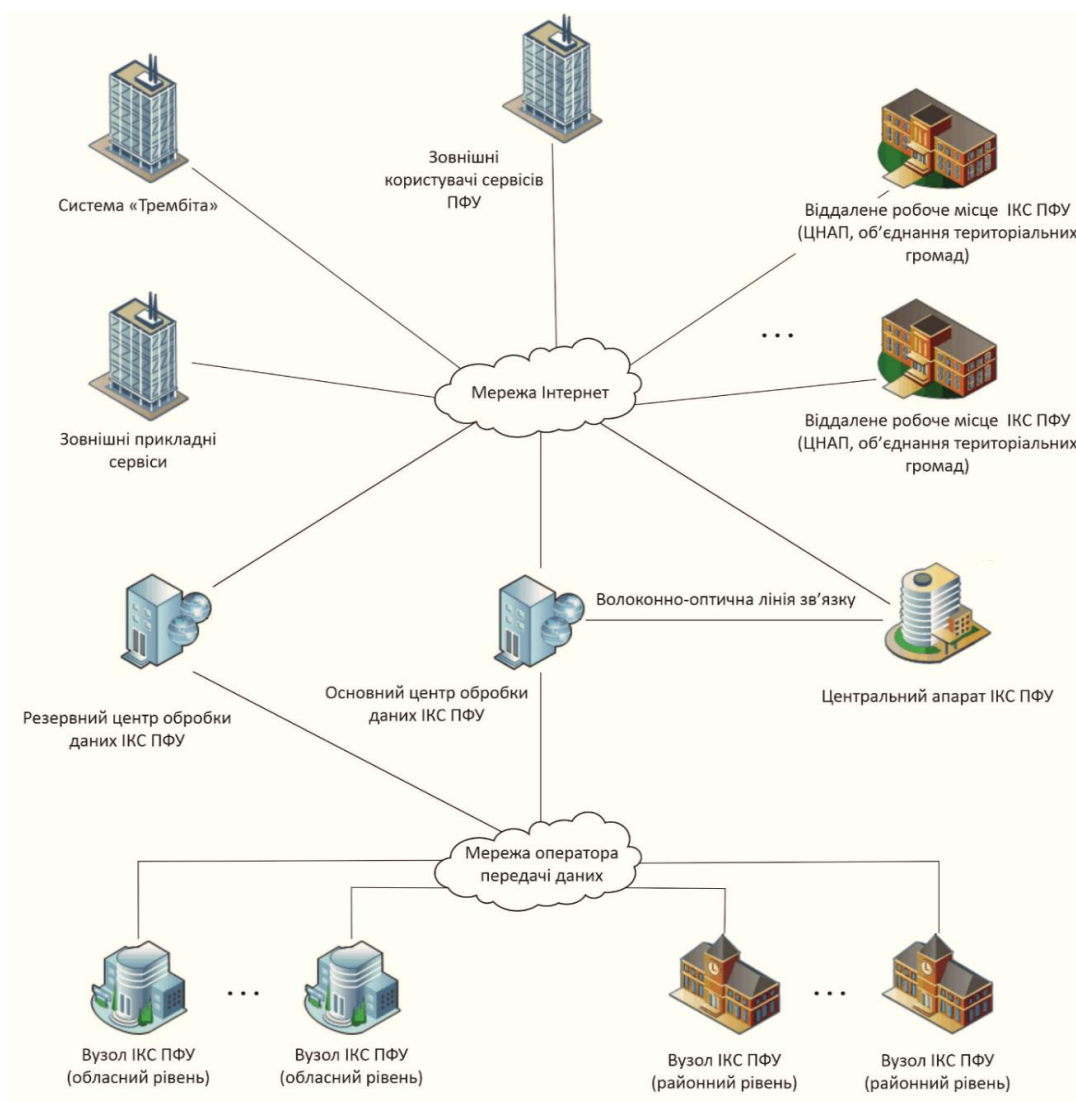


Рисунок 1.4 – Архітектура ІКІС ПФУ [55]

Організація процесів прийняття рішень значною мірою відображена у архітектурі ІКІС ПФУ. Значну увагу приділено, і продовжує приділятися створенню сучасної аналітичної складової. Наявна аналітична функціональність

орієнтована переважно на формування регламентованої звітності, облік виконання бюджету, розрахунок показників для планування видатків на наступний бюджетний рік та задоволення інформаційних запитів органів державної влади, візуалізацію статистичної інформації. Тобто, забезпечено ретроспективний та поточний аналіз, а прогнозна складова реалізована фрагментарно.

Отже, удосконалення потребують саме аналітична та прогнозна складові системи підтримки прийняття рішень, оскільки зараз вони більшою мірою орієнтовані на вирішення точкових і поточних питань забезпечення пенсійних виплат. Адже, для подолання сучасних викликів недостатньо лише оперативного реагування на події та розв'язання окремих задач - потрібні системні рішення щодо подальшої розбудови соціальної сфери, і особливо її провідної ланки, системи державного пенсійного забезпечення. Узагальнюючи виявлені обмеження наявної аналітичної складової ІКІС ПФУ, основними з них є такі:

- прогнозування здійснюється переважно методами екстраполяції наявних тенденцій, що в умовах нестаціонарності та структурних змін не забезпечує прийнятної точності;
- відсутні засоби сценарного аналізу, тобто немає можливості системно оцінити наслідки різних варіантів розвитку подій та різних управлінських рішень;
- не передбачено інструментів інтелектуального аналізу даних, зокрема опрацювання неструктурованої текстової інформації, яка є суттєвим джерелом відомостей про потреби населення;
- не формалізовано механізм поєднання результатів формальних розрахунків з експертними судженнями, хоча на практиці рішення ухвалюються саме на такій основі;
- не забезпечено узгодження прогнозних оцінок за рівнями управління та зворотний зв'язок щодо результативності ухвалених рішень;
- прогнозні оцінки не супроводжуються характеристиками якості та стійкості, що ускладнює обґрунтування рішень, ухвалених на їх основі.

Слід відзначити, що на ринку програмного забезпечення наявна значна кількість різноманітних рішень для моделювання і прогнозування. За функціональним призначенням такі засоби можна узагальнити у такі групи: засоби аналізу часових рядів та побудови моделей динаміки; засоби регресійного аналізу та класифікації; засоби загального статистичного і байєсівського аналізу, зокрема побудови мереж Байєса; засоби машинного навчання та ансамблевих методів; засоби опрацювання неструктурованої текстової інформації; засоби візуалізації та побудови інформаційних панелей.

Серед історично значущих спеціалізованих розробок слід згадати програмне забезпечення *Nonstationary Time Series Analysis and Decomposition using Time-Varying Parameter Autoregressions* Р. Прадо і М. Веста [56], призначене для аналізу часових рядів із використанням класів нестационарних авторегресійних моделей та декомпозиції рядів для вивчення структури прихованих компонент [56], а також програмне забезпечення *Autoregressive Component Modelling* Г. Уерти і М. Веста, що апроксимує авторегресійні моделі з акцентом на структурі прихованих компонент [57]. До цієї ж групи належать інтерактивна програма *JMulTi* [58], розроблена для одновимірного та багатовимірного аналізу часових рядів, відкрита система *Cronos* [59], що забезпечує оцінювання параметрів та прогнозування на основі моделей АРКС, УАРУГ, МСВ та інших, а також бібліотека *NBB*, орієнтована на вбудовування засобів аналізу часових рядів у прикладні застосування.

Окрему групу становлять засоби байєсівського аналізу даних, які характеризуються високою гнучкістю та широкими можливостями опрацювання невизначеностей різних типів. Бібліотека *Bayesian Filtering Library* [59] забезпечує застосування незалежного фреймворку для використання динамічних мереж Байєса, тобто для рекурсивного оброблення інформації та оцінювання змінних і параметрів на основі правила Байєса й методів Монте-Карло. Бібліотека *Bayes++ Bayesian Filter Classes* містить імовірнісні методи байєсівської фільтрації, бібліотека *ReBEL* є інструментарієм для послідовного байєсівського оцінювання в моделях у просторі станів, а інструментарій *Bayesian*

Model Averaging [60] реалізує методику байєсівського усереднення моделей, розроблену для врахування невизначеності, притаманної процесу вибору моделі. Відомим засобом оцінювання параметрів лінійних і нелінійних моделей є середовище BUGS та його відкрита реалізація OpenBUGS, які ґрунтуються на методі Монте-Карло для марковських ланцюгів [61, 62]; наявність відкритого коду дає можливість користувачам розширювати систему.

Сучасну практику аналізу даних визначають насамперед універсальні середовища. Програмне середовище R надає широкий інструментарій статистичних методів [63] - лінійний та нелінійний регресійний аналіз, статистичні тести, аналіз часових рядів - і є гнучкішим за спеціалізовані статистичні продукти, оскільки користувачі можуть розширювати функціональність власними функціями; сильною стороною середовища є побудова високоякісних графіків для наукових публікацій.

Мова Python з бібліотеками оброблення та аналізу даних, статистичного моделювання й машинного навчання [64, 65-67] забезпечує реалізацію всього циклу - від збору та попереднього оброблення даних до побудови ансамблевих моделей і оцінювання їх якості і має принципову перевагу з точки зору інтеграції розроблених модулів у наявну інформаційну систему.

Комерційні аналітичні платформи, зокрема ті, що містять засоби Text Mining [68, 69, 70], застосовні для опрацювання неструктурованих звернень громадян [6, 7, 9, 10]. Загальними недоліками наявних програмних систем є висока вартість; закритість для подальшого розвитку користувачами продукту; відсутність функцій, необхідних для розв'язання конкретних прикладних задач; відсутність необхідної критеріальної бази оцінювання якості прогнозів; надлишкова універсальність застосування, яка часто призводить до погіршення остаточних результатів аналізу даних.

Жоден із розглянутих засобів не може бути використаний як готове рішення для задач прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення. Проте, окремі з представлених у вказаних

додатках підходи, доцільно застосовувати як інструментальну основу для реалізації розроблюваних моделей та методів.

Тобто, адаптуючи розроблюваний модельний апарат та прогнозний інструментарій до програмно-апаратної платформи ІКІС ПФУ відповідно до існуючих процедур підтримки прийняття рішень Пенсійного фонду України. Такий підхід є принциповим: він забезпечує можливість практичного впровадження результатів без заміни наявної системи і без порушення регламентованих процедур роботи з даними, що мають обмежений доступ або є закритими. Схему розширення функціоналу аналітичної та прогнозної складових наведена на рис. 1.5.

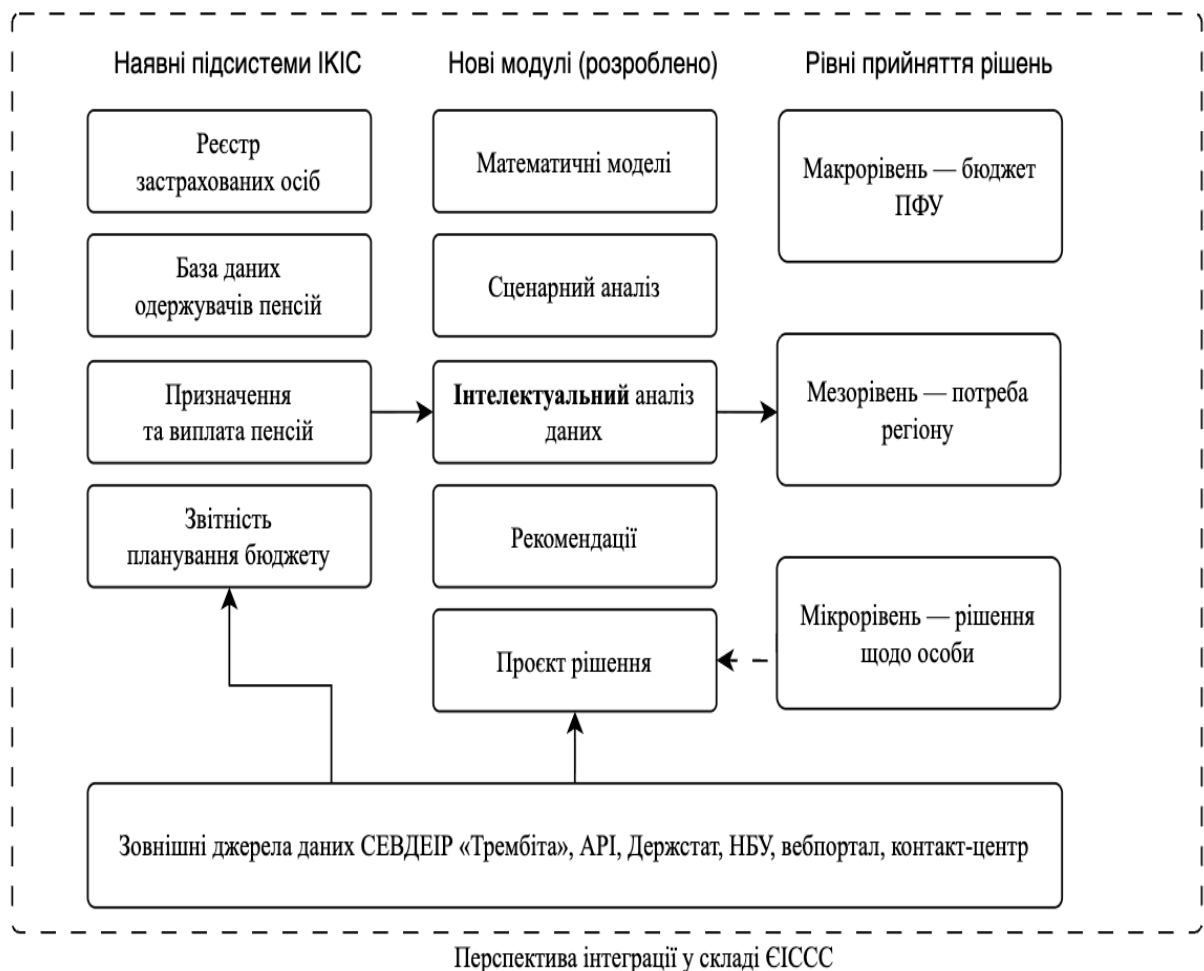


Рисунок 1.5 - Схема розширення аналітичного та прогнозного функціоналу ІКІС ПФУ

Тобто, виходячи з результатів проведеного дослідження, функціональність ІКІС ПФУ, слід розширити додаванням модулів:

- побудови математичних моделей, що реалізує відбір моделей-кандидатів, оцінювання їх параметрів та діагностику адекватності;
- сценарного аналізу, який забезпечує формування варіантів розвитку подій та розрахунок прогнозів за кожним сценарієм;
- інтелектуального аналізу даних, зокрема опрацювання неструктурованої текстової інформації, отриманої зі звернень громадян та відкритих джерел [6, 7, 9, 10];
- розроблення рекомендацій, у якому поєднано використання експертних суджень з алгоритмами штучного інтелекту;
- підготовки проєкту рішення, що формує документ для розгляду відповідальною посадовою особою.

Для інтеграції зазначених модулів у наявну систему запропоновано застосувати конвергентний підхід, за якого нові компоненти не замінюють наявні, а вбудовуються у чинні процеси оброблення інформації, використовуючи ті самі джерела даних, механізми розмежування доступу та інтерфейси взаємодії. Це дає можливість поетапного впровадження та збереження працездатності системи на кожному етапі.

1.4 Урахування невизначеностей як засіб удосконалення підтримки прийняття рішень у системі державного пенсійного забезпечення

Сучасні умови функціонування системи державного пенсійного забезпечення зміщують акценти у системі підтримки прийняття рішень з планування бюджету до прогнозування його показників, зокрема, витрат на пенсійне забезпечення. Активна увага зосереджується на розробленні методик всебічного дослідження ситуації, врахування сукупного впливу соціально-економічних, суспільно-політичних та інших чинників, розкритті невизначеностей [71-78]. Потрібен інструментарій для розроблення сценаріїв розвитку подій, які відповідно, потребуватимуть різних підходів до побудови прогнозів.

Особливості функціонування вітчизняної системи державного пенсійного забезпечення мають бути враховані під час побудови моделей та прогнозів. Зокрема, є потреба ідентифікації та врахування невизначеностей, пов'язаних зі статистичними даними, структурою досліджуваних процесів і відповідно їх моделями, оцінками параметрів та якістю прогнозів.

Негативні впливи на обчислювальні процеси виникають внаслідок неповноти або некоректності вхідних даних, наявності шуму у вимірах, впливу значних випадкових збурень, недостатньо якісного оцінювання структури і параметрів моделей, некоректного вибору обчислювальних процедур. Тому, проблема ідентифікації невизначеностей, тобто встановлення їх типу і параметрів, застосування підходів до їх розкриття, є актуальною проблемою, яка потребує практичного вирішення.

Для вирішення цієї проблеми, у роботі запропоновано метод урахування невизначеності різного типу, який передбачає сукупне використання різних груп внутрішніх та зовнішніх чинників, відрізняється розширенням опрацювання відкритої інформації та інформації з обмеженим доступом на різних етапах прийняття рішень, розширеним складом груп чинників та прив'язкою ступеня деталізації даних і способу використання результатів до рівня управління. Це забезпечує поглиблений аналіз предметної області та створює передумови для прийняття обґрунтованих управлінських рішень.

Основна увага була сконцентрована на дослідженні найбільш характерних для системи державного пенсійного забезпечення невизначеностей. Як показало проведене дослідження – це статистична, структурна та параметрична невизначеності, а в умовах війни - системна невизначеність. Остання не зводиться до жодного з традиційних типів, характеризується одночасною зміною складу доступних даних, структури досліджуваних процесів та самих правил функціонування системи, тому не може бути опрацьована засобами, розрахованими на стаціонарні умови.

Основні типи невизначеностей та методи їх опрацювання узагальнені в табл. 1.3.

Таблиця 1.3 - Типи невизначеностей, характерні для системи державного пенсійного забезпечення та методи їх опрацювання

Тип	Причини виникнення	Методи опрацювання
Статистична	неімпутовані пропуски у даних про страховий стаж за період до 2004 року; мультиколінеарність макроекономічних та демографічних показників; короткі вибірки за період після 2022 року; екстремальні значення; похибки та різні методичні підходи до збору статистичних даних	попереднє оброблення та очищення даних; скорочення простору входних ознак; відбір ознак; статистичні процедури опрацювання пропусків та аномалій; використання альтернативних розподілів; генерування синтетичних наборів даних, що відповідають структурі даних з обмеженим доступом
Структурна	неможливість розкриття всіх причинно-наслідкових зв'язків між чинниками, що впливають на обсяг видатків; зміна структури процесів внаслідок змін законодавства; нестационарність процесів; значні випадкові збурення	експертне оцінювання; статистичне тестування та перевірка гіпотез; альтернативні методи оцінювання структури моделі; кореляційний аналіз; тестування на нелінійність та нестационарність; багатомодельний підхід
Параметрична	некоректний вибір методу оцінювання параметрів; недостатній обсяг статистичних даних; невідповідність даних припущенням про нормальність розподілу; вплив неврахованої нелінійності	альтернативні методи оцінювання параметрів; метод максимальної правдоподібності; метод Монте-Карло для марковських ланцюгів; адаптивне оновлення оцінок за надходження нових даних
Ймовірнісна	складні механізми причинно-наслідкових залежностей; комплексний вплив взаємозалежних випадкових процесів; невизначеність права окремої особи на призначення пенсії; поведінковий вибір щодо «придбання» стажу	статичні та динамічні мережі Байєса, гібридні; моделі марковського типу; інші ймовірнісно-статистичні моделі; ймовірнісні фільтри
Системна	обмеження доступу до статистичної інформації або її неповнота; неконтрольований міграційний рух населення та внутрішнє переміщення; зміна статево-вікової структури; тимчасова окупація окремих територій; непрогнозовані зміни законодавства; різке зростання чисельності одержувачів пенсій по інвалідності	сценарний аналіз; експертне оцінювання; багатомодельний підхід та комбінування прогнозів; опрацювання неструктурованої текстової інформації як додаткового джерела даних; оцінювання стійкості прогнозів до неповноти й спотворення даних

Структурні невизначеності виникають у випадках, коли структура досліджуваного процесу невідома або відома лише частково. Зокрема, при застосуванні функціонального підходу до побудови моделі інформації про структуру процесу, як правило, недостатньо, тому необхідно проводити додатковий кореляційний аналіз, оцінювати взаємні ймовірності для змінних, застосовувати тестування на нелінійність та нестационарність. Для пошуку кращих структур моделей доцільно застосовувати схеми адаптивного оцінювання з використанням комплексних критеріїв, що поєднують кілька характеристик якості моделі. Існує кілька можливостей адаптивного оцінювання структури моделі: автоматизований аналіз часткових автокореляцій з метою оцінювання порядку авторегресії; оцінювання лагів екзогенних змінних; аналіз типу розподілу даних; адаптивне оцінювання параметрів моделі; оптимальне оцінювання вагових коефіцієнтів для процедур згладжування. Використання конкретної схеми адаптації залежить від обсягу і якості даних, конкретної постановки задачі, вимог до адекватності моделі та статистичних характеристик даних.

Для опрацювання статистичної невизначеності застосовують методи попереднього оброблення даних, зокрема фільтрації, оскільки виміри містять, як правило, випадкову складову. За невідомих статистичних характеристик вимірів застосовують методи оптимальної фільтрації, які дають можливість одночасно оцінювати стан досліджуваного процесу і його статистичні характеристики. У задачах прогнозування у системі державного пенсійного забезпечення більше значення має інший аспект статистичної невизначеності - наявність пропусків у даних. Їх опрацювання, за неможливості імпутації пропусків у даних про страховий стаж за період до 2004 року потребує розроблення специфічних підходів.

Параметричні невизначеності, зокрема зміщення та відсутність консистентності оцінок, виникають внаслідок низької інформативності даних або невідповідності вимірів нормальному розподілу при використанні методу найменших квадратів. Подібна ситуація може виникнути також через

мультиколінеарність незалежних змінних або вплив неврахованої нелінійності процесу. Якщо вибірка даних недостатньо повна для побудови моделі, вона може бути розширена за допомогою спеціальних процедур або імітаційного моделювання; якщо виміри не відповідають нормальному розподілу, для оцінювання параметрів доцільно застосувати метод максимальної правдоподібності або метод Монте-Карло для марковських ланцюгів [69], який не вимагає високих обчислювальних витрат за умови невеликої кількості оцінюваних параметрів.

Ймовірнісні невизначеності опрацьовуються засобами байєсівського підходу. Такий підхід має переваги при оцінюванні ймовірності настання права на призначення пенсії за віком використано апарат гібридних мереж Байєса [2, 60, 148]. Його застосування дає можливість одночасно враховувати неперервні характеристики застрахованої особи (страховий стаж за періодами, вік) та дискретні ознаки (стать, наявність пільгового стажу, право на дострокову пенсію).

Іншими інструментами опрацювання невизначеностей є нечітка логіка та нейронечіткі моделі, які дають можливість формалізувати якісні судження експертів у вигляді лінгвістичних змінних та правил виведення. Такий підхід доцільний у випадках, коли кількісні дані відсутні або недостовірні, а наявні лише експертні оцінки характеру та інтенсивності впливу чинників.

Опрацювання системної невизначеності потребує принципово іншої логіки. Оскільки в умовах коли не можна, ані отримати повні дані, ані вважати структуру процесу незмінною, єдиним прийнятним підходом є перехід від пошуку єдиної «правильної» моделі до побудови множини сценаріїв та ансамблю моделей з подальшим комбінуванням отриманих оцінок. Встановлено, що кращих результатів комбінування прогнозів можна досягти у випадку, коли дисперсії оцінок прогнозів для різних методів прогнозування суттєво не відрізняються [75-80]. Додатковим засобом є залучення альтернативних джерел інформації, зокрема неструктурованих текстових даних, які дають змогу отримати сигнали про зміну ситуації раніше, ніж вони проявляться в офіційній

статистиці. Наведене узагальнення методичних підходів до працювання невизначеностей є основою для формування методу урахування різних груп чинників у процесі підтримки прийняття рішень. Розширений склад чинників наведено в табл. 1.4 (крім джерел ПФУ).

Таблиця 1.4 - Групи чинників, що враховуються у процесі підтримки прийняття рішень у системі державного пенсійного забезпечення [3, 4, 8, 40, 41, 79, 81]

Група чинників	Приклади показників	Основні джерела даних	Рівень
Макроекономічні	валовий внутрішній продукт, індекс споживчих цін, середня та мінімальна заробітна плата, рівень безробіття, видатки зведеного бюджету	статистика Державної служби статистики, Національного банку України	мезо-, макро
Демографічні	чисельність населення, статеві-вікова структура, коефіцієнт демографічного навантаження, міграційні процеси	офіційна статистика, база даних одержувачів пенсій	мезо-, макро
Екологічні	показники стану довкілля, здоров'я населення; наслідків техногенних та екологічних катастроф	офіційна статистика, відомчі дані	мезо-, макро
Суспільно-політичні, зокрема наслідки воєнного стану	внутрішні переміщення населення, тимчасова окупація територій, дані про санітрані втрати, суспільні очікування щодо пенсійного забезпечення	статистичні дані та звітність, дані міжнародних організацій, звернення громадян, дані з відкритих джерел мережі інтернет	усі рівні
Організаційно-правові	порядок призначення пенсій, індексації та перерахунку пенсій, пільгові умови	законодавство, внутрішні регламенти	усі рівні
Індивідуальні характеристики застрахованих осіб	страховий стаж за періодами до та після 2004 року, сума страхових внесків, стать, вік, наявність інвалідності, наявність періодів роботи, що дають право на достроковий вихід на пенсію або зниження пенсійного віку, пільги	реєстр застрахованих осіб, документи про стаж	мікро

Сукупне врахування наведених груп чинників дозволить покращити обґрунтованість рішень, що приймаються на одному та різних рівнях системи

державного пенсійного забезпечення за рахунок обміну інформацією. Однак це створює методичну проблему оброблення різномірних факторів, що вимірюються в різних шкалах, надходять з різних джерел, оновлюються з різною періодичністю, а частина з них взагалі не має кількісного вираження і доступна лише у формі експертних суджень або неструктурованих текстів. у процесі прогнозування потреби у пенсійному забезпеченні на різних етапах прийняття рішень розроблена методика, наведена на рис. 1.6.

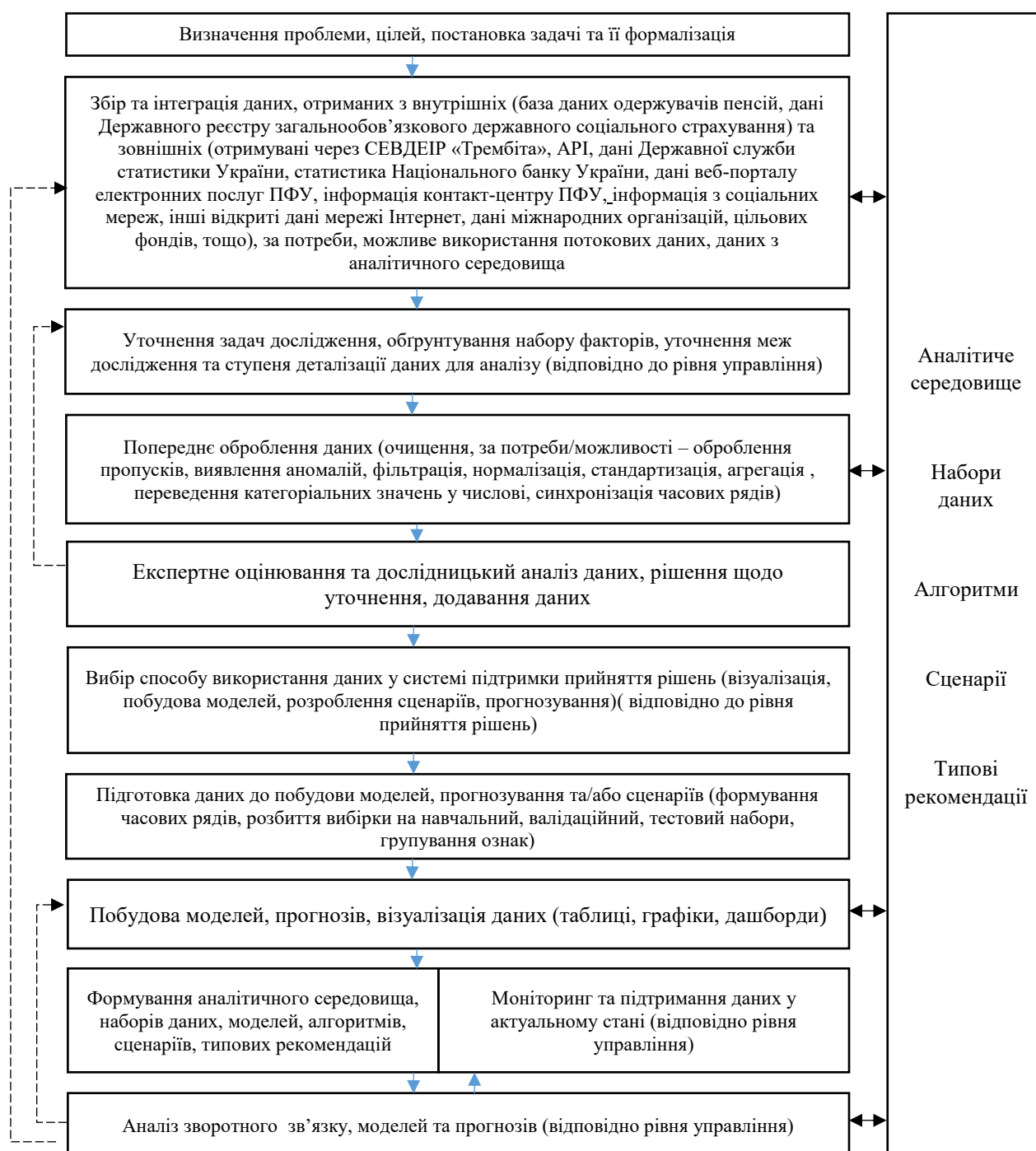


Рисунок 1.6 - Метод урахування різних груп чинників

Метод урахування різних груп чинників у процесі підтримки прийняття рішень щодо пенсійного забезпечення населення в умовах невизначеності (рис. 1.6) передбачає послідовне виконання етапів, які здійснюються на тлі використання спільного для всіх рівнів управління аналітичного середовища, що включає набори даних, моделі, алгоритми, сценарії та типові рекомендації.

Етап 1. Визначення проблеми, цілей, постановка задачі та її формалізація. Встановлюється, рішення якого саме рівня потребує прогностичного забезпечення, які показники підлягають прогнозуванню, який горизонт і форма прогнозу потрібні, які обмеження діють щодо доступу до даних.

Етап 2. Збір та інтеграція даних, отриманих з внутрішніх і зовнішніх джерел. До внутрішніх належать база даних одержувачів пенсій та дані Державного реєстру загальнообов'язкового державного соціального страхування; до зовнішніх - дані, отримувані через систему електронної взаємодії державних електронних інформаційних ресурсів та програмні інтерфейси, дані Державної служби статистики України [82], статистика Національного банку України [83], дані вебпорталу електронних послуг ПФУ [84], інформація контакт-центру ПФУ, інформація із соціальних мереж та інші відкриті дані [85] мережі Інтернет, дані міжнародних організацій і цільових фондів. За потреби можливе використання потокових даних та даних, накопичених в аналітичному середовищі. Склад джерел, форми подання даних та процедури їх інтеграції докладно розглянуто в роботі.

Етап 3. Уточнення задач дослідження, обґрунтування набору чинників, уточнення меж дослідження та ступеня деталізації даних для аналізу відповідно до рівня управління. На цьому етапі з переліку, наведеного в табл. 1.5, відбираються ті групи чинників, які є значущими для конкретної задачі, і визначається необхідний ступінь деталізації: індивідуальні дані, регіональні або агреговані показники.

Етап 4. Попереднє оброблення даних: очищення, за потреби та за наявності можливості - опрацювання пропусків, виявлення аномалій, фільтрація, нормалізація, стандартизація, агрегування, переведення категоріальних значень

у числові, синхронізація часових рядів. Принципово, що опрацювання пропусків виконується лише за умови, коли це методично допустимо; для даних про страховий стаж за період до 2004 року така операція не виконується, а невизначеність передається на подальші етапи як самостійна характеристика задачі.

Етап 5. Експертне оцінювання та дослідницький аналіз даних, ухвалення рішення щодо уточнення або додавання даних. Результат цього етапу може повернути процес до етапу 3, якщо виявиться, що відібраний набір чинників є неповним або недостовірним.

Етап 6. Вибір способу використання даних у системі підтримки прийняття рішень - візуалізація, побудова моделей, розроблення сценаріїв, прогнозування - відповідно до рівня прийняття рішень. Для оперативних рішень мікрорівня достатньо може бути візуалізації та індивідуальної оцінки ймовірності, тоді як для стратегічних рішень макrorівня потрібен сценарний прогноз.

Етап 7. Підготовка даних до побудови моделей, прогнозування та розроблення сценаріїв: формування часових рядів, розбиття вибірки на навчальний, валідаційний та тестовий набори, групування ознак [86-88].

Етап 8. Побудова моделей, прогнозів та візуалізація даних у формі таблиць, графіків та інформаційних панелей. Моделі та ансамблі моделей, застосовувані на цьому етапі, розглянуто в розділі 3, а приклади їх застосування - у розділі 4.

Етап 9. Аналіз зворотного зв'язку, моделей та прогнозів відповідно до рівня управління. Оцінюється, наскільки побудовані прогнози відповідають фактичним даним, а ухвалені на їх основі рішення - потребам населення; результати цього етапу є підставою для коригування як моделей, так і постановки задачі.

Етап 10. Моніторинг та підтримання даних в актуальному стані відповідно до рівня управління. Етап забезпечує придатність аналітичного середовища для наступних циклів прийняття рішень.

Запропонований метод є розвитком наявних підходів урахування чинників у процесі прогнозування потреби у пенсійному забезпеченні. На відміну від

наявних підходів, які обмежуються трьома групами чинників і розраховані на роботу з повними наборами даних, статистичними даними, даними з відкритих джерел, розроблений метод передбачає розширення опрацювання відкритої інформації та інформації з обмеженим доступом на різних етапах прийняття рішень, а також розширений склад груп чинників з включенням екологічних, суспільно-політичних та індивідуальних характеристик застрахованих осіб. Це забезпечує поглиблений аналіз предметної області, створюючи передумови для прийняття обґрунтованих управлінських рішень у системі державного пенсійного забезпечення в умовах невизначеності [3, 8].

Розроблений метод відрізняється прив'язкою до складу чинників, ступеня деталізації даних та способу використання результатів до рівня управління, що забезпечує узгодженість рішень за ієрархією; наявність зворотних зв'язків від етапу експертного оцінювання до етапу обґрунтування набору чинників і від етапу аналізу результатів до постановки задачі, що робить процес ітеративним; допустимість збереження невизначеності, яка не може бути коректно усунена на етапі попереднього оброблення, з її передачею на етапи моделювання; використання спільного аналітичного середовища як механізму накопичення наборів даних, моделей, сценаріїв та типових рекомендацій, що забезпечує повторне використання результатів.

1.5 Постановка задачі дисертаційної роботи

Виконаний системологічний аналіз дає підстави для формулювання науково-прикладної проблеми, що розв'язується в роботі. Автоматизація ключових процесів заклала умови для спроможності та стійкості системи державного пенсійного забезпечення, зокрема в умовах війни. Однак, для подолання сучасних викликів недостатньо оперативного реагування на події та розв'язання точкових задач. Потрібні системні рішення щодо подальшої розбудови соціальної сфери країни та її провідної ланки - системи державного пенсійного забезпечення. Тому важливою і актуальною науково-прикладною

проблемою є підвищення ефективності підтримки прийняття рішень у системі державного пенсійного забезпечення в умовах невизначеності за рахунок впровадження сучасних моделей, методів прогнозування та інформаційних технологій у роботу Пенсійного фонду України й установ соціальної сфери різного рівня та підпорядкування. Що й визначило мету та задачі дисертації.

Мета дисертаційної роботи – створення нових моделей, методів прогнозування для виконання системних досліджень, розроблення на їх основі методологічних засад інформаційно-аналітичного забезпечення підтримки прийняття рішень у системі державного пенсійного забезпечення.

Для досягнення поставленої мети в роботі необхідно вирішити такі завдання:

1. дослідити особливості системи державного пенсійного забезпечення України та прийняття рішень щодо пенсійного забезпечення в умовах невизначеності;
2. виконати огляд та аналіз сучасних методів моделювання і прогнозування розвитку системи пенсійного забезпечення;
3. удосконалити методику використання різномірних даних, отриманих з різних джерел як для опису та прогнозування процесів, що характерні для системи державного пенсійного забезпечення в умовах невизначеності, так і для забезпечення зворотного зв'язку стосовно ефективності управлінських рішень, що приймаються органами ПФУ;
4. розробити системну методологію прогнозування процесів, що характерні для державного пенсійного забезпечення в умовах невизначеності, зокрема, спричиненої військовим конфліктом;
5. побудувати математичні моделі та ансамблі моделей, які описують процеси, характерні для системи пенсійного забезпечення України для прогнозування її розвитку для підтримки прийняття управлінських рішень в умовах невизначеності, зокрема, спричиненої військовим конфліктом;
6. розробити та виконати апробацію запропонованих підходів до моделювання та прогнозування розвитку системи державного пенсійного забезпечення України, запропонувати підходи, що забезпечують адекватний опис

причинно-наслідкових зв'язків різних груп чинників та визначення можливих варіантів розвитку досліджуваних процесів;

7. виконати аналіз адекватності побудованих моделей та сформулювати висновки щодо можливості їх застосування для розв'язання задач підтримки прийняття рішень у системі державного пенсійного забезпечення;
8. на основі системного підходу побудувати інформаційну технологію, яка доповнює існуючі, поєднуючи розроблені моделі та методи прогнозування та адаптувати її для подальшого використання у підсистемі підтримки прийняття рішень Єдиної інформаційної системи соціальної сфери.

Висновки до розділу 1

1. Виконано системологічний аналіз проблеми розроблення та використання систем підтримки прийняття рішень у державному і публічному управлінні, фінансовому секторі та, зокрема, у системі пенсійного забезпечення. Встановлено, що система державного пенсійного забезпечення є складною ієрархічною системою, а процеси, що характеризують її розвиток, переважно нелінійні та розвиваються в умовах невизначеності під впливом соціально-економічних, демографічних, екологічних та суспільно-політичних чинників.

2. Відзначено багаторівневість та ієрархічність процесів підтримки прийняття рішень. Ключовими є задачі прийняття рішень щодо призначення та виплати пенсій застрахованій особі (мікрорівень) та фінансування витрат на виплату пенсій (макрорівень); обґрунтовано доцільність виокремлення проміжного, мезорівня. Показано, що на різних рівнях відрізняються не лише масштаб задачі, а й природа прогнозованої величини та форма прогнозного результату, що зумовлює необхідність поєднання моделей мікро- та макрорівнів і узгодження прогнозів за ієрархією.

3. Проаналізовано закордонний досвід моделювання розвитку пенсійних систем і встановлено, що наявні класи інструментів розраховані на умови повноти статистичної інформації та відсутності структурних зламів, тому

потребують адаптації до вітчизняних умов. Акцентовано увагу на унікальності вітчизняної системи державного пенсійного забезпечення, зокрема на неоднорідності даних про страховий стаж за періодами до та після 2004 року, незавершеності структурної реформи, наявності значної кількості пільгових умов призначення пенсій, можливості набуття стажу за договором про добровільну участь та впливі воєнної агресії.

4. Розглянуто інформаційні технології Пенсійного фонду України, значну увагу приділено Інтегрованій комплексній інформаційній системі ПФУ та її аналітичній складовій. Встановлено, що удосконалення потребують аналітична та прогнозна складові системи підтримки прийняття рішень, оскільки зараз вони більшою мірою орієнтовані на вирішення точкових і поточних питань забезпечення пенсійних виплат. Виявлено конкретні обмеження наявної аналітичної складової, зокрема відсутність засобів сценарного аналізу, інструментів інтелектуального аналізу даних, формалізованого механізму поєднання розрахунків з експертними судженнями та характеристик якості й стійкості прогнозних оцінок.

5. Виконано аналіз програмних засобів моделювання і прогнозування, встановлено їх переваги та обмеження щодо застосування у розв'язуваній задачі. Загальними недоліками наявних систем є висока вартість, закритість для подальшого розвитку користувачами, відсутність функцій, необхідних для розв'язання конкретних прикладних задач, відсутність необхідної критеріальної бази оцінювання якості прогнозів та надлишкова універсальність. Обґрунтовано, що задачу розроблення прогнозного інструментарію доцільно розв'язувати виходячи з особливостей програмно-апаратної платформи ІКІС ПФУ та відповідно до існуючих процедур підтримки прийняття рішень, розширивши функціональність системи модулями побудови математичних моделей, сценарного аналізу, інтелектуального аналізу даних, розроблення рекомендацій та підготовки проєкту рішення із застосуванням конвергентного підходу до їх інтеграції.

6. Встановлено, що найбільш характерними для системи державного пенсійного забезпечення є статистична, структурна та параметрична невизначеності, а в умовах війни - системна невизначеність. Представлено узагальнення методичних підходів до опрацювання невизначеностей різних типів із прив'язкою причин їх виникнення до особливостей предметної області.

7. Розроблено метод урахування різних груп внутрішніх та зовнішніх чинників у процесі підтримки прийняття рішень щодо пенсійного забезпечення населення в умовах невизначеності, який відрізняється розширенням опрацювання відкритої інформації та інформації з обмеженим доступом на різних етапах прийняття рішень, розширеним складом груп чинників та прив'язкою ступеня деталізації даних і способу використання результатів до рівня управління, що забезпечує поглиблений аналіз предметної області та створює передумови для прийняття обґрунтованих управлінських рішень.

8. З виконаного аналізу сучасного стану розвитку моделей, методів та інформаційних технологій підтримки прийняття рішень у системі державного пенсійного забезпечення сформульовано науково-прикладну проблему, мету та завдання дисертаційної роботи, визначено об'єкт, предмет і методи дослідження.

РОЗДІЛ 2

МЕТОДИЧНІ ПІДХОДИ ДО ПОБУДОВИ МОДЕЛЕЙ І ПРОГНОЗІВ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ

2.1 Урахування особливостей системи державного пенсійного забезпечення при побудові моделей і прогнозів на певному рівні прийняття рішення

Основними джерелами формування бюджету ПФУ є надходження від сплати єдиного соціального внеску на загальнообов'язкове державне соціальне страхування, кошти державного бюджету (цільові трансферти та фінансування для покриття виплат соціальних допомог, пенсійних програм, дефіциту ПФУ, тощо) та інші джерела (добровільні внески громадян, інвестиційні доходи, кошти від штрафів, санкцій, тощо). В умовах війни та складної соціально-економічної ситуації особливого значення набуває задача забезпечення пенсійних виплат. Проблема забезпечення своєчасного та в повному обсязі фінансування виплати пенсій стає не лише питанням функціонування ПФУ, а й однією з задач забезпечення національної стійкості. Тому, в роботі основний акцент зроблено саме на розробленні моделей та методів прогнозування для підтримки прийняття рішень щодо пенсійних витрат. Хоча, проблема розроблення моделей та методів прогнозування для забезпечення прийняття ефективних управлінських рішень у системі державного пенсійного забезпечення зазвичай, розглядається як дві взаємопов'язані складові: прогнозування видаткової частини (визначається чисельністю та структурою контингенту одержувачів пенсій і розміром виплат) та прогнозування доходної частини (визначається чисельністю платників страхових внесків, розміром заробітної плати, станом ринку праці), вибір зумовлений практичною стороною задачі: рішення, які приймаються в системі, стосуються насамперед обґрунтування потреби у коштах, а точність прогнозу видатків визначає обґрунтованість бюджетного запиту [13, 38, 41].

Особливістю методичних підходів до побудови прогнозів у системі пенсійного забезпечення, є те, що необхідно враховувати, що частина статистичної інформації у період воєнного стану недоступна або доступна з обмеженнями, рішення приймаються в умовах постійного оновлення та уточнення вхідних даних, тому прогноз не може бути одноразовим розрахунком і потребує процедури оновлення оцінок, для частини задач потрібна не точкова оцінка, а розподіл або інтервал, оскільки рішення стосується події, що має ймовірнісну природу.

Крім того, майбутній перехід до пенсійної системи, яка поєднує солідарну, накопичувальну та добровільну складові, має бути врахований при розробленні методичних підходів. У солідарній системі обсяг зобов'язань перед конкретною особою визначається її страховим стажем та сумою сплачених внесків через формулу обчислення розміру пенсії, тобто прогнозування зводиться до оцінювання чисельності контингенту та середнього розміру виплати. За наявності накопичувальної складової додається новий клас невизначеності - результати інвестування пенсійних активів, які не залежать від демографічних чинників і описуються моделями фінансових часових рядів. Крім того, з'являється задача прогнозування переходу застрахованих осіб між рівнями системи, яка має поведінковий характер і не може бути описана лише кількісними показниками [19, 20, 90-95].

Принципово, що на перехідному етапі система одночасно містить обидві логіки: для осіб, які набувають права на пенсію найближчими роками, визначальним є накопичений до цього часу страховий стаж, тоді як для молодших поколінь зростає роль накопичувальної складової. Тому методичні підходи мають бути придатними для роботи з обома конфігураціями, а прогнозна модель - допускати розширення складу чинників без перебудови всієї конструкції. Це вимога до інструментарію, врахована в роботі при розробленні відповідної інформаційної технології.

Також, перехід до трирівневої моделі змінює вимоги до складу даних. Персоніфікований облік у солідарній системі фіксує стаж і суми внесків; для

накопичувальної складової потрібні дані про стан індивідуальних накопичувальних рахунків, доходність інструментів інвестування та структуру інвестиційного портфеля. Відповідно, методика використання різномірних даних, представлена в роботі, побудована так, що склад джерел може розширюватися.

Як показало дослідження, для адекватного опису процесів у системі державного пенсійного забезпечення за сучасних умов, важливо врахувати вплив як економічних, демографічних, екологічних, суспільних та правових чинників, так і вплив воєнного стану. Має бути розглянутий розвиток системи в цілому та враховані індивідуальні характеристики осіб, застрахованих у системі загальнообов'язкового державного пенсійного страхування. Ці дві групи чинників мають різну природу і не можуть бути враховані в межах однієї моделі.

Чинники першої групи є агрегованими показниками, вимірюються на рівні країни або регіону, мають річну або квартальну періодичність і придатні для регресійного та сценарного моделювання. Чинники другої групи є характеристиками окремої особи, мають змішаний кількісно-якісний характер і придатні для ймовірнісно-статистичного моделювання та методів інтелектуального аналізу даних. Спроба звести їх до спільної шкали призводить або до втрати індивідуальної інформації при агрегуванні, або до неможливості врахувати макроекономічний контекст при роботі з індивідуальними даними.

Тому в роботі запропоновано підхід, який передбачає поєднання моделей мікро- та макрорівнів, за якого:

- на макрорівні будуються моделі обсягу видатків та надходжень за групами агрегованих чинників із застосуванням сценарного аналізу та застосовуються відповідні критерії їх адекватності;
- на мікрорівні будуються ймовірнісні моделі настання права на призначення пенсії за індивідуальними характеристиками застрахованих осіб;
- результати мікрорівня агрегуються до чисельності контингенту одержувачів пенсій і використовуються як вхідна змінна моделей макрорівня, що забезпечує узгодженість прогнозів за ієрархією управління;

– узгодження виконується з контролем відповідності агрегованих оцінок мікрорівня та безпосередніх оцінок макрорівня; розбіжність між ними є індикатором потреби в уточненні складу чинників або даних.

Таке поєднання дає можливість зберегти інформативність індивідуальних даних там, де вона визначальна (право на призначення пенсії конкретній особі), і водночас урахувати макроекономічний та демографічний контекст там, де він визначальний (обсяг видатків та потреба у фінансуванні).

Система пенсійного забезпечення потребує комплексного підходу до аналізу та прогнозування, оскільки на її функціонування впливають численні чинники різної природи. Дані, що описують ці чинники, є багатовимірними: одночасно використовуються десятки змінних - від демографічних (чисельність населення, чисельність пенсіонерів за категоріями, середня тривалість життя, коефіцієнт демографічного навантаження) до економічних (валовий внутрішній продукт, індекс споживчих цін, середня та мінімальна заробітна плата, рівень безробіття, видатки зведеного бюджету), екологічних (показники стану довкілля, що впливають на здоров'я населення) та суспільно-правових (умови призначення пенсій, пенсійний вік, вимоги до страхового стажу, пільгові умови). Принципова відмінність багатовимірних даних від одновимірних полягає не в кількості показників, а в тому, що вони описують кілька вимірів стану середовища одночасно і містять інформацію про взаємозв'язки між вимірами. Одне значення одного показника - наприклад, середня тривалість життя за окремий рік - не дає можливості оцінити, чим зумовлена його зміна; той самий показник у сукупності з економічними та екологічними змінними за той самий період уже дає таку можливість.

Соціальна сфера є складною системою з багатьма взаємопов'язаними змінними: погіршення стану довкілля може впливати на здоров'я населення і збільшувати чисельність одержувачів пенсій по інвалідності, тоді як економічні потрясіння змінюють реальний розмір виплат і потребу в їх індексації.

Узагальнений склад груп чинників макро- та мезорівня, що враховуються при побудові моделей обсягу пенсійних виплат, наведено в табл. 2.1.

Таблиця 2.1 – Групи чинників макро- та мезорівня, що визначають зміну обсягу пенсійних виплат

Група чинників	Основні показники	Періодичність	Характер впливу на обсяг виплат
Демографічні	чисельність населення, статеві-вікова структура; середня тривалість життя; демографічне навантаження; міграційний рух	річна, для окремих показників квартална	визначає чисельність контингенту одержувачів; вплив накопичувальний, проявляється з лагом
Економічні	валовий внутрішній продукт; індекс споживчих цін; середня та мінімальна заробітна плата; рівень безробіття; видатки зведеного бюджету	місячна, квартална, річна	визначає розмір виплат через індексацію та перерахунки, а також надходження страхових внесків
Екологічні	показники стану довкілля, що впливають на здоров'я населення регіону; наслідки техногенних та природних надзвичайних ситуацій	річна	опосередкований: через стан здоров'я населення впливає на чисельність одержувачів пенсій по інвалідності
Суспільні	суспільні очікування; рівень довіри до системи; готовність до добровільної участі; тематика звернень громадян	за надходження даних	впливає на поведінковий вибір застрахованих осіб
Правові	умови призначення пенсій; пенсійний вік; вимоги до страхового стажу; порядок індексації та перерахунку; пільгові умови	за змінами у законодавстві	дискретний вплив: зміна норми однооментно змінює структуру контингенту та розмір виплат
Воєнний стан	рух населення; тимчасова окупація територій; санітарні та безповоротні втрати, структурні зміни контингенту	за надходження даних, з обмеженнями	визначальний за поточних умов; змінює як чисельність, так і структуру контингенту

Перелік чинників, наведений у табл. 2.1 уточнює для кожної групи перелік показники, періодичність їх оновлення та характер впливу на обсяг пенсійних видатків. Слід підкреслити, що: правові чинники та чинники воєнного стану впливають дискретно, а не поступово. Зміна норми законодавства щодо вимог до страхового стажу однооментно переводить певну частину застрахованих осіб з однієї категорії до іншої, тобто створює у часових рядах структурний злам. Це означає, що моделі, побудовані на даних до такої зміни, втрачають адекватність,

і саме тому в роботі акцент зроблено на сценарному підході та на методиці оновлення оцінок за надходження нових даних.

Для прогнозування контингенту одержувачів пенсій на мікрорівні визначальними є характеристики окремої застрахованої особи, які встановлюють її право на призначення пенсії за віком та рік настання цього права. Це кількісні та якісні характеристики, а саме: страховий стаж, сума страхових внесків, стать, наявність інвалідності, наявність пільг, наявність особливих умов праці, вік, розмір призначеної пенсії (табл. 2.2).

Таблиця 2.2 - Індивідуальні характеристики застрахованих осіб, що враховуються при прогнозуванні контингенту одержувачів пенсій

Характеристика	Шкала вимірювання	Роль у визначенні права на призначення пенсії
страховий стаж за весь період трудової діяльності	кількісна (роки, місяці)	основна умова права на призначення пенсії за віком; визначає, у якому віці настане право - у 60, 63 або 65 років
страховий стаж до 2004 року	кількісна, з пропусками, що не імпутуються	частина загального стажу, яка підтверджується документально і не відображена в інформаційній системі повністю
страховий стаж після 2004 року	кількісна, повна	частина загального стажу, наявна в реєстрі застрахованих осіб за даними персоніфікованого обліку
сума страхових внесків	кількісна (грн)	визначає розмір призначеної пенсії; опосередковано впливає на обсяг видатків
дата народження	кількісна	визначає рік настання пенсійного віку
стать	номінальна	враховується для періодів, коли діяли різні умови щодо пенсійного віку та стажу
наявність інвалідності	номінальна	підстава для призначення пенсії по інвалідності незалежно від настання пенсійного віку
пільговий, спеціальний стаж	кількісна та номінальна	підстава для дострокового призначення пенсії або зменшення пенсійного віку
наявність права на дострокову пенсію	номінальна	змінює рік настання права на призначення пенсії
Розмір призначеної пенсії	кількісна (грн)	використовується при агрегуванні до обсягу видатків

Вказані характеристики зберігаються у реєстрі застрахованих осіб Державного реєстру загальнообов'язкового державного соціального страхування [19, 26] та в базі даних одержувачів пенсій. Оскільки інформація є конфіденційною і містить персональні дані [96], для проведення чисельних експериментів у роботі використано згенеровані набори даних, що за структурою та статистичними характеристиками відповідають наявним у базі даних персоніфікованого обліку.

Показники, що описують чинники макрорівня, часто тісно корелюють між собою. Наприклад, чисельність пенсіонерів за віком та чисельність одержувачів соціальних пенсій формально є різними показниками, однак сильно залежні, оскільки соціальні пенсії призначаються особам пенсійного віку, які не мають достатнього страхового стажу. Без урахування цього явища оцінки параметрів моделі стають нестійкими, а їх інтерпретація - некоректною.

Для виявлення мультиколінеарності використовується фактор інфляції дисперсії (Variance Inflation Factor, VIF), який показує, наскільки зростає дисперсія оцінки коефіцієнта змінної через її кореляцію з іншими регресорами:

$$VIF_i = 1 / (1 - R_i^2) \quad (2.1)$$

де R_i^2 - коефіцієнт детермінації регресії змінної X_i на всі інші регресори моделі. Значення VIF, близьке до одиниці, свідчить про практичну відсутність мультиколінеарності; значення від 1 до 5 відповідає помірній кореляції, яка зазвичай є прийнятною; значення понад 5 вказує на суттєву мультиколінеарність, що може викликати нестабільність оцінок коефіцієнтів; значення понад 10 свідчить про сильну мультиколінеарність, за якої необхідно скоротити склад змінних, об'єднати чинники або застосувати метод головних компонент. Зазначимо, що наведені межі є емпіричними правилами, а не статистичними критеріями, і мають інтерпретуватися з урахуванням обсягу вибірки та цілей моделювання [88, 97-100].

Для показників, використаних у роботі, розрахунок VIF показав наявність суттєвої мультиколінеарності. Так, для змінної «чисельність пенсіонерів»

отримано значення VIF, що перевищує 5, що свідчить про її сильну кореляцію з чисельністю населення та з показниками чисельності одержувачів окремих видів пенсій. Аналогічно, для показника валового внутрішнього продукту на душу населення в трифакторній моделі з середньою тривалістю життя та індексом споживчих цін значення VIF перевищило 6, тоді як для двох інших змінних залишилося в межах помірної кореляції. Це зумовило застосування методу головних компонент, за результатами якого значення VIF для всіх компонент, що увійшли до моделі, знизилися до рівня, нижчого за 2, тобто мультиколінеарність було усунено.

Змінні, що описують чинники, належать до різних шкал вимірювання, і це визначає допустимі операції над ними та методи їх подальшого оброблення. Кількісні змінні (середня тривалість життя, обсяг видатків, розмір пенсії) допускають усі арифметичні операції та є основою для побудови часових рядів. Порядкові змінні (наприклад, рівень пенсійного навантаження за градаціями «низький - середній - високий») мають природний порядок, однак відстань між градаціями не є визначеною, тому до них не застосовуються операції усереднення. Номінальні змінні (регіон проживання, тип пенсії, стать) використовуються лише для класифікації і перед побудовою моделей потребують переведення у числову форму. Врахування типу шкали є обов'язковою умовою коректності подальшого оброблення даних і закладене в методику, представлену в роботі.

2.2 Удосконалення методики використання різнорідних даних при побудові моделей і прогнозів у системі державного пенсійного забезпечення

Наявні підходи до підготовки даних до моделювання розраховані переважно на однорідні набори даних, отриманих з єдиного джерела. Однак, реальні умови потребують урахування особливостей предметної області, зокрема, різнорідності джерел за структурою, періодичністю та рівнем доступу, а також наявності пропусків, які принципово не підлягають заповненню. Крім того, для системи пенсійного забезпечення важливо забезпечити зворотний

зв'язок стосовно ефективності управлінських рішень, які приймаються органами ПФУ.

Для побудови моделей, застосованих у розроблюваній методиці прогнозування, для проведення чисельних експериментів використано дані з трьох груп джерел, які відрізняються структурою, форматом подання та порядком доступу.

Офіційна статистика є основним джерелом відкритих даних, що дає можливість формувати часові ряди ключових показників макро та мезо- рівнів. Дані подаються переважно у форматах *.csv, *.xls, *.xlsx та *.pdf. Державна служба статистики України [82] надає демографічні показники (чисельність населення, чисельність пенсіонерів, коефіцієнт демографічного навантаження), економічні індикатори (валовий внутрішній продукт, індекс споживчих цін, середня заробітна плата, рівень безробіття) та показники соціальних витрат (видатки державного і місцевих бюджетів на соціальний захист). Інформація Міністерства соціальної політики, сім'ї та єдності України [101] дає можливість деталізувати структуру соціальних виплат за видами допомог та статистику надання субсидій. Пенсійний фонд України публікує дані про середній розмір пенсії, кількість призначених та перерахованих пенсій, середній вік одержувачів пенсій [81]. Ці набори характеризуються повнотою, регулярною періодичністю оновлення та єдиною методологією обліку, що робить їх базовим джерелом для формування часових рядів.

Державні реєстри містять персоніфіковані або агреговані дані, які формуються державними органами для адміністрування пенсійного страхування та соціальної допомоги. Вони зберігаються у захищених форматах, що потребують спеціальних процедур доступу. Ключовими є реєстр застрахованих осіб Державного реєстру загальнообов'язкового державного соціального страхування, який містить дані про страховий стаж, розміри внесків та періоди трудової діяльності; база даних одержувачів пенсій з інформацією про категорії одержувачів, дату призначення пенсії та тривалість виплат; реєстри соціальних

допомог з даними про сімейний стан, наявність інвалідності, статус малозабезпечених та отримання пільг.

Неструктуровані дані є джерелом інформації, яка не відображена в офіційній статистиці. До них належать електронні звернення громадян до Пенсійного фонду та Міністерства соціальної політики, сім'ї та єдності України, публікації у засобах масової інформації про зміни законодавства та публічні обговорення пенсійної реформи, коментарі користувачів у соціальних мережах та на форумах. Такі дані не мають чіткої табличної форми, проте після оброблення методами аналізу тексту та обробки природної мови дають змогу виявляти настрої, проблемні тенденції та сигнали про збої у наданні соціальних послуг раніше, ніж вони проявляться у статистиці [6, 7, 9, 10]. Порівняльну характеристику груп джерел наведено в табл. 2.3.

Таблиця 2.3 - Порівняльна характеристика груп джерел даних

Група джерел	Формати подання	Періодичність	Рівень доступу	Основне призначення в задачі
Офіційна статистика	*.csv, *.xls, *.xlsx та *.pdf	місячна, квартальна, річна	відкритий	формування часових рядів чинників макро- та мезорівня
Державні реєстри	реляційні бази даних, *.XML, *.JSON	за настанням події, безперервне оновлення	обмежений, персональний і дані	формування наборів індивідуальних характеристик застрахованих осіб
Інші джерела	текст, вебсторінки, повідомлення	за надходженням	відкритий та внутрішній	виявлення суспільних чинників, зворотний зв'язок щодо ефективності рішень

Наявність кількох форматів зберігання зумовлює потребу в уніфікації структури даних на етапі їх інтеграції. Поєднання структурованих даних офіційної статистики та державних реєстрів з аналізом неструктурованих даних

дає можливість формувати комплексніші моделі, які враховують не лише кількісні характеристики, а й якісні сигнали про зміни в соціальній сфері.

Приклад структурованих даних, що надходять з офіційних статистичних джерел, наведено у табл. 2.4.

Таблиця 2.4 – Основні показники пенсійного забезпечення та ринку праці України за даними офіційної статистики, 2020–2022 рр.

Рік	Середній розмір пенсії, грн	Кількість пенсіонерів на 1 січня, тис.	Рівень безробіття за методологією МОП, %	Індекс споживчих цін, грудень до грудня попереднього року %
2020	3082,98	11335	9,5	105,0
2021	3507,51	11131	9,9	110,0
2022	3991,53	10841	обстеження не проводилось	126,6

У таблиці 2.4 наведений приклад використання структурованих даних, що представлені у офіційній статистиці. У даному прикладі поєднано дані ПФУ та Державної служби статистики. Цей приклад демонструє необхідність узгодження показників, щоб уникнути змістовних помилок. Такі набори, як, наведені у табл. 2.4 характеризуються повнотою, регулярною періодичністю оновлення та єдиною методологією обліку, що робить їх базовим джерелом для формування часових рядів у задачах прогнозування пенсійних витрат. Однак, з різних причин є періоди, у які спостереження не проводились. Державні реєстри - це персоніфіковані або агреговані дані, які формуються державними органами для адміністрування пенсійного страхування та соціальної допомоги. Вони зазвичай представлені у захищених форматах (SQL-бази, XML, JSON), що потребують спеціальних процедур доступу. Ключові джерела:

- Єдиний державний реєстр застрахованих осіб (ЄДРЗО), містить дані про страховий стаж, розміри внесків, періоди трудової діяльності кожного працівника.

- База даних отримувачів пенсій зберігає інформацію про одержувачів пенсій (за віком, по інвалідності, по втраті годувальника, тощо), дату призначення пенсії, тривалість виплат.

- Реєстри соціальних допомог (наприклад, Єдиний соціальний реєстр): дані про сімейний стан, кількість дітей, інвалідність, статус малозабезпечених, отримання допомог та пільг.

Приклад подання одного й того самого запису реєстру у форматах реляційної бази даних, XML та JSON наведено у табл. 2.5.

Таблиця 2.5 – Подання одного запису реєстру одержувачів пенсій у різних форматах структурованих даних

Приклад у SQL-базі (таблична форма):	<pre>CREATE TABLE pensioners (id INT PRIMARY KEY, full_name VARCHAR(100), birth_date DATE, pension_type VARCHAR(50), pension_amount DECIMAL(10,2), region VARCHAR(50)); INSERT INTO pensioners (id, full_name, birth_date, pension_type, pension_amount, region) VALUES (1, 'Іваненко Петро Іванович', '1958-07-15', 'за віком', 5400.00, 'Київська область');</pre>
Приклад у форматі XML:	<pre><pensioner> <id>1</id> <fullName>Іваненко Петро Іванович</fullName> <birthDate>1958-07-15</birthDate> <pensionType>за віком</pensionType> <pensionAmount>5400.00</pensionAmount> <region>Київська область</region> </pensioner></pre>
Приклад у форматі JSON:	<pre>{ "id": 1, "fullName": "Іваненко Петро Іванович", "birthDate": "1958-07-15", "pensionType": "за віком", "pensionAmount": 5400.00, "region": "Київська область" }</pre>

Як видно з табл. 2.5, наявність кількох форматів зберігання зумовлює потребу в уніфікації структури даних на етапі їх інтеграції, що реалізується засобами ETL-процесів.

Інший приклад демонструє використання даних із державних реєстрів:

- Дата народження - визначає вік одержувача пенсії, що впливає на право на пенсію за віком або інші види допомог. *Приклад:* «1958-07-15» (означає вік 65 років станом на 2023 рік).
- Стать - категоріальна змінна, яка може впливати на пенсійний вік і тривалість виплат. *Приклад:* «Чоловік» або «Жінка».
- Регіон проживання - географічна прив'язка, що важлива для аналізу регіональних відмінностей у потребі соцвиплат. *Приклад:* «Київська область», «Львівська область».
- Сума щомісячної пенсії - кількісна змінна, що показує розмір пенсії для кожного одержувача. *Приклад:* 5400 грн.
- Категорія пенсії - визначає тип нарахованої пенсії, що впливає на тривалість і обсяг виплат. *Приклад:* «За віком», «По інвалідності», «По втраті годувальника».
- Кількість років страхового стажу - ключовий показник для права на пенсію та розрахунку її розміру. *Приклад:* 35 років.
- Дані про пільги - інформація про наявність спеціальних статусів, що впливають на право на дострокову пенсію або підвищений розмір виплат. *Приклад:* «Чорнобилець 1 категорії», «Ветеран війни», «Особа з інвалідністю II групи».
- Тип останнього місця роботи - відображає сферу зайнятості перед виходом на пенсію, що може визначати пільги чи специфіку пенсійних нарахувань. *Приклад:* «Державна служба», «Приватний сектор», «Сільське господарство».

Перелік основних джерел неструктурованих текстових даних із прикладами відповідного контенту наведено у табл. 2.6.

Такі дані не мають чіткої табличної форми, проте після оброблення методами аналізу текстів (Text Mining) та засобами оброблення природної мови (NLP) дають змогу виявляти настрої, проблемні тенденції та сигнали про збої у наданні соціальних послуг [6, 7, 9, 10].

Таблиця 2.6 – Джерела неструктурованих текстових даних для аналізу потреб населення у пенсійному забезпеченні

Джерело	Приклад даних
Електронні звернення громадян до ПФУ, Мінсоцполітики, Уряду: скарги, питання щодо перерахунку пенсій, відгуки про якість обслуговування.	«Чому мені не нарахували пенсію за липень 2023 року?» «Прошу надати роз'яснення щодо доплати за стаж понад 35 років.» «Скарга на відмову в призначенні пенсії за віком через помилки в базі даних.»
Публікації у засобах масової інформації: новини про зміни законодавства, публічні обговорення пенсійної реформи.	Заголовок новини: «Уряд схвалив підвищення мінімальної пенсії з січня 2024 року» Фрагмент статті: «Нові правила нарахування пенсій обговорили в Міністерстві соціальної політики. Головна зміна - збільшення мінімального стажу для виходу на пенсію.»
Соціальні мережі та форуми: коментарі користувачів про затримки виплат, проблеми при оформленні допомог, думки про проблеми у ПФУ.	Коментар у Facebook: «Мені досі не прийшла пенсія за цей місяць. У когось ще є така проблема?» Допис на форумі: «Хто оформлював субсидію цього року? Чи правда, що процедуру ускладнили?» Обговорення в Telegram-чаті: «Кажуть, пенсії затримують через технічні проблеми в Пенсійному фонді.»

Поєднання структурованих даних офіційної статистики та державних реєстрів із аналізом неструктурованих даних дозволяє формувати комплексні та більш повні моделі прогнозування, які враховують не лише кількісні характеристики, а й якісні сигнали про зміни в соціальній сфері. Такий підхід дає змогу виявляти малопомітні тенденції та «тонкі» сигнали у запитах і настроях населення, які не завжди відображаються у традиційних статистичних показниках. Крім того, інтеграція різних джерел даних забезпечує можливість швидкої адаптації прогнозних моделей до нових умов і дозволяє оперативно реагувати на зміни у потребах населення шляхом сценарного моделювання та підготовки обґрунтованих управлінських рішень.

Процес збору даних передбачає отримання інформації з різномірних джерел, її об'єднання, перевірку якості та приведення до єдиного подання, придатного для моделювання. Для цього використовуються процеси видобування, перетворення та завантаження даних (Extract - Transform - Load,

ETL) [138], у межах яких інформація приводиться до єдиного формату, узгоджується за ключами та завантажується у цільове сховище. Архітектуру процесу інтеграції наведено на рис. 2.1.

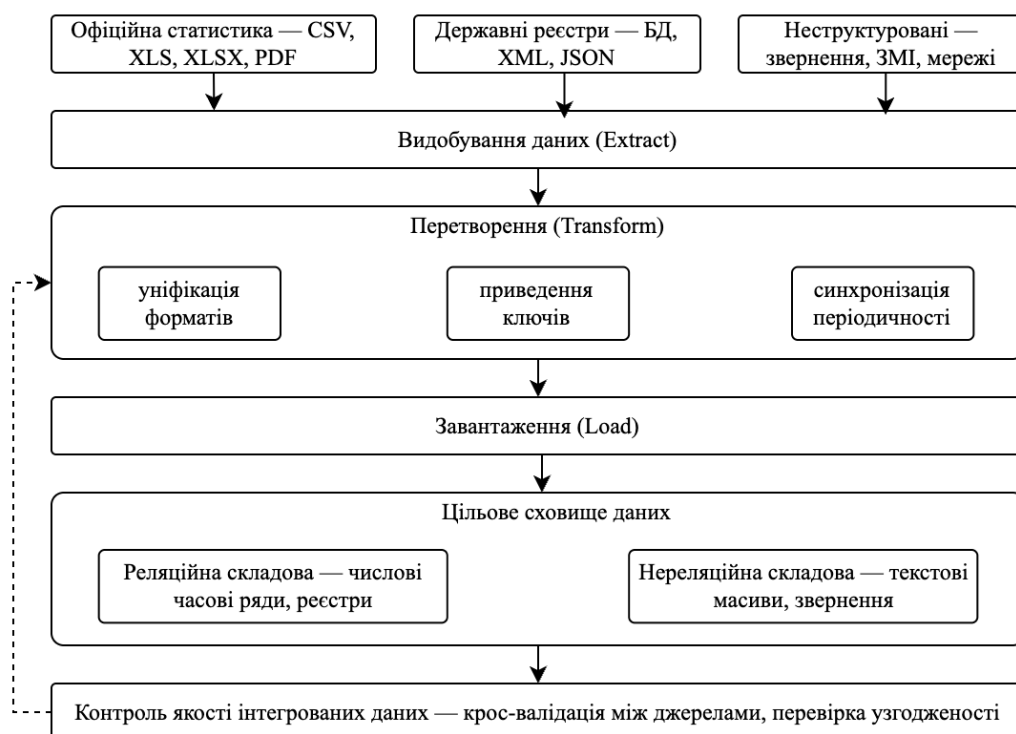


Рисунок 2.1 -Процес інтеграції різномірних даних у системі державного пенсійного забезпечення

Істотною особливістю задачі є різна періодичність збору даних: пенсійні нарахування формуються помісячно, показники валового внутрішнього продукту – поквартально, демографічні показники – щорічно. Для об'єднання таких показників в одному наборі даних виконується синхронізація часових рядів. Для квартальних даних застосовується дисагрегація шляхом інтерполяції або експертного оцінювання; для річних - кумуляція місячних та квартальних значень або екстраполяція трендів; для показників, що за своєю природою є річними, значення приймається постійним для кожного місяця відповідного року. Вибір способу синхронізації визначається природою показника: для поточкових величин (видатки, надходження) допустима кумуляція, для показників стану (чисельність, тривалість життя) - інтерполяція.

Об'єднання структурованих і неструктурованих даних потребує створення єдиної інформаційної моделі [102-108], яка включає структури для зберігання

кількісних змінних, таблиці для категоріальних ознак та текстові поля для неструктурованих даних. Для цього формується сховище з реляційними та нереляційними елементами, у якому числові часові ряди зберігаються у реляційній частині, а текстові масиви – у частині, оптимізованій для повнотекстового пошуку. Зв'язок між частинами забезпечується спільними ключами – періодом та регіоном.

Для перевірки якості об'єднаних даних застосовуються процедури крос-валідації, які дають можливість порівнювати дані з різних джерел за однаковими періодами і категоріями, перевіряти внутрішню узгодженість наборів та оцінювати коректність інтеграції неструктурованих даних. Останнє виконується шляхом порівняння частоти ключових тем у зверненнях громадян з офіційною статистикою відповідних проблем: наприклад, зростання частоти звернень щодо затримки виплат за відсутності відповідних відомостей у звітності є підставою для перевірки повноти даних. Розбіжність між джерелами понад установлений порог є сигналом про можливу помилку інтеграції та підставою для повернення до етапу перетворення.

Наприклад, здійснюється завантаження даних про щомісячні середні пенсії з таблиць Державної служби статистики України та даних про кількість пенсіонерів із бази даних ПФУ для формування єдиного представлення (табл. 2.7): [81, 82]

Таблиця 2.7 – Результат інтеграції даних Державної служби статистики України та ПФУ у єдиний часовий ряд

Місяць	Середній розмір пенсії (дані ПФУ), грн	Чисельність пенсіонерів (дані ПФУ), тис.	Індекс споживчих цін (дані Державної служби статистики України), %
01.01.2019	2645,70	11470	109,8
01.01.2020	3082,98	11535	104,1
01.01.2021	3507,51	11131	105,0
01.01.2022	3991,53	10841	110,0

У таблиці 2.7 у одному наборі об'єднано моментні (значення показників на 1 січня) та інтервальні – значення показника, що характеризує вплив інфляції. Часто дані мають різну періодичність збору - пенсійні нарахування (помісячно), показники ВВП (поквартально), демографічні зміни (щорічно). Щоб об'єднати їх в одному наборі для аналізу, потрібна синхронізація:

- для квартальних даних використовується дисагрегація (розбиття на місяці шляхом інтерполяції або експертного оцінювання),
- для річних - кумуляція місячних/квартальних значень або екстраполяція трендів.

Наприклад, якщо річний показник середньої тривалості життя - 71,2 роки, його значення використовується як постійне для кожного місяця відповідного року в місячному часовому ряді. Об'єднання структурованих і неструктурованих даних потребує створення єдиної інформаційної моделі, яка включає:

- структури для зберігання числових змінних (кількість пенсіонерів, обсяг видатків),
- таблиці для категоріальних ознак (тип пенсії, регіон проживання),
- текстові поля для неструктурованих даних (звернення громадян).

Для цього формується єдина база даних із реляційними та нереляційними елементами (наприклад, PostgreSQL + Elasticsearch). Наприклад, таблиця pensions із полями date, region, avg_pension, num_pensioners, та зв'язана таблиця complaints із текстом звернень, категоріями проблем (затримка, нарахування, помилки в стажі тощо).

Попереднє оброблення є етапом, від якості якого безпосередньо залежить адекватність побудованих моделей. Основними операціями є очищення даних, виявлення та усунення аномалій, приведення змінних до єдиної шкали.

Очищення даних включає вилучення дублікатів записів, виправлення помилок форматів та узгодження одиниць вимірювання. Дублікати виникають, коли інформація надходить з різних підсистем: наприклад, у даних щодо одержувачів пенсій можуть з'являтися кілька однакових записів для однієї особи. Формально запис вважається дублікатом, якщо значення ключових полів двох

рядків збігаються; ключем є ідентифікатор особи або його комбінація з датою призначення пенсії. Процедура передбачає визначення ключових полів, порівняння значень ключів між записами, позначення груп однакових записів та збереження одного запису з кожної групи. Формально умова дублікатності записів має вигляд

$$K_i = K_j, i \neq j \Rightarrow Duplicate(j) = 1 \quad (2.2)$$

де K_i, K_j - значення ключа або конкатенації кількох ключових полів для записів з індексами i та j ; $Duplicate(j)$ - індикатор віднесення запису j до дублікатів.

Виявлення аномалій виконується статистичними та відстаневими методами [88, 95]. Аномалією вважається значення, яке істотно відхиляється від характерного для відповідного показника діапазону. Для часових рядів застосовується перевірка на відхилення від локального тренду, для перерізів даних - правило міжквартильного розмаху

$$x \in [Q_1 - 1,5 \cdot IQR; Q_3 + 1,5 \cdot IQR], \quad IQR = Q_3 - Q_1 \quad (2.3)$$

де Q_1 і Q_3 - перший і третій квартилі вибірки; IQR - міжквартильний розмах.

Значення, що виходять за межі зазначеного інтервалу, позначаються як потенційні аномалії. Для часових рядів додатково оцінюються стандартизовані залишки відносно локального тренду. Принципово, що не кожне аномальне значення є помилкою: різка зміна показника може відображати реальну подію, зокрема зміну законодавства або наслідки воєнних дій. Тому виявлені аномалії не вилучаються автоматично, а передаються на етап експертного оцінювання, передбачений методом (рис. 1.6), де приймається рішення про їх вилучення, коригування або збереження як інформативної особливості ряду.

Приведення змінних до єдиної шкали є необхідним, оскільки показники, що входять до моделей, відрізняються за порядком величин на кілька десяткових розрядів: чисельність пенсіонерів вимірюється мільйонами осіб, індекс споживчих цін — відсотками, розмір пенсії — гривнями. Застосовуються стандартизація, за якої змінна центрується за середнім та нормується за стандартним відхиленням,

$$z_{ij} = (x_{ij} - \bar{x}_j) / s_j \quad (2.4)$$

та нормалізація до заданого діапазону

$$\tilde{x}_{ij} = (x_{ij} - \min_j) / (\max_j - \min_j) \quad (2.5)$$

де x_{ij} - значення j -ї змінної для i -го спостереження; $\bar{x}_j$ і s_j - середнє та стандартне відхилення j -ї змінної; $\min_j$ і $\max_j$ - її мінімальне та максимальне значення у вибірці. Для методу головних компонент стандартизація є обов'язковою умовою, оскільки без неї компоненти визначаються переважно змінними з найбільшою дисперсією. Категоріальні змінні переводяться у числову форму: номінальні - індикаторним кодуванням, порядкові - з урахуванням природного порядку градацій.

Виконання процедури вилучення дублікатів ілюстровано у табл. 2.8 і 2.9. У вихідному наборі (табл. 2.8) наявні два записи з однаковими значеннями ключових полів, а після оброблення (табл. 2.9) збережено лише один із них.

Таблиця 2.8 – Фрагмент набору даних до вилучення дублікатів

pensioner_id	assignment_date	pension_amount
12345	2022-01-15	4500
12345	2022-01-15	4500

Таблиця 2.9 – Фрагмент набору даних після вилучення дублікатів

pensioner_id	assignment_date	pension_amount
12345	2022-01-15	4500

Для видалення дублікатів записів застосовано перевірку однакових значень ключових полів (ID пенсіонера та дати призначення пенсії). При виявленні дублікатів зберігається лише останній за часом оновлення. Це забезпечує коректність кількісних оцінок і запобігає завищенню прогнозованих витрат.

Цей етап методики має принципове значення для роботи і є тим, що відрізняє її від типових (загальноприйнятих) процедур підготовки даних. У

прогнозуванні витрат на пенсійне забезпечення найпоширенішими є дані у формі часових рядів, для заповнення пропусків у яких придатні такі методи: просте усереднення за умови малої кількості пропусків; генерування оцінок за моделями, побудованими на наявних вимірах; генерування даних з розподілів, форма і параметри яких визначаються за наявними даними; експоненціальне згладжування; застосування оптимізаційних процедур, зокрема алгоритму максимізації очікування [88, 95, 108-111]; імітаційне моделювання. Загальні підходи до аналізу даних з пропусками та умови коректності імпутації розглянуто в [112-114].

Однак для даних про страховий стаж застрахованих осіб жоден із наведених методів не є застосовним. Причина не статистична, а правова: страховий стаж є юридичним фактом, який підтверджується документально, і за період до 2004 року [16] відомості про нього не відображені в інформаційній системі повністю, оскільки персоніфікований облік запроваджено з 1 січня 2004 року [2]. Відсутність значення в такому часовому ряді не означає відсутності стажу - вона означає, що стаж не підтверджений документами, наявними в системі. Заповнення такого пропуску статистичною оцінкою призвело б до присвоєння особі права, якого вона юридично може не мати, і навпаки. Отже, пропуски цього типу не підлягають імпутації за жодних умов.

Відповідно, у методиці передбачено окрему процедуру: пропуски, які не можуть бути коректно заповнені, не вилючаються і не заповнюються, а зберігаються як самостійна характеристика задачі та передаються на етапи моделювання. На етапі побудови моделей така невизначеність опрацьовується ймовірнісними засобами: замість детермінованого значення стажу оцінюється розподіл можливих значень, а замість детермінованого висновку про право на пенсію – ймовірність настання цього права. Практично це означає, що загальний обсяг страхового стажу застрахованої особи за весь період трудової діяльності розглядається як комбінація двох складових: стажу до 2004 року, який містить пропуски і оцінюється за розподілом, та стажу після 2004 року, який відомий за даними персоніфікованого обліку. Формально це записується так:

$$S = S_1 + S_2, \quad S_1 \sim N(\mu_1, \sigma_1^2), \quad S_2 \sim N(\mu_2, \sigma_2^2) \quad (2.6)$$

де S – страховий стаж за весь період трудової діяльності; S_1 – стаж за період до 2004 року, який містить пропуски і оцінюється за розподілом; S_2 – стаж за період після 2004 року, відомий за даними персоніфікованого обліку; μ_1, σ_1 та μ_2, σ_2 – математичні сподівання та стандартні відхилення відповідних складових. Ймовірність достатності стажу для настання права на призначення пенсії у віці a обчислюється як:

$$P_s(a) = P(S \geq s(a)) \quad (2.7)$$

де $s(a)$ - нормативна вимога до страхового стажу для призначення пенсії у віці a , встановлена законодавством для віку 60, 63 та 65 років. Кожна складова описується власними статистичними характеристиками, поєднання яких дає розподіл загального стажу для відповідної групи застрахованих осіб. Порівняльну характеристику типів пропусків та способів їх опрацювання наведено в табл. 2.10.

Таблиця 2.10 - Типи пропусків у даних та способи їх опрацювання

Тип пропуску	Приклад	Причина виникнення	Спосіб опрацювання
випадковий пропуск у часовому ряду агрегованого показника	відсутнє значення індексу споживчих цін за певний місяць	збій збору або публікації даних	інтерполяція, експоненціальне згладжування, короткостроковий прогноз за наявними вимірами
систематична відсутність показника за період	відсутність статистики за окремими регіонами	обмеження збору даних в умовах воєнного стану	використання даних за доступними регіонами, сценарні припущення, експертне оцінювання
пропуск у даних про страховий стаж до 2004 року	відсутні відомості про роботу застрахованої особи	дані в базі даних наявні з 2004 року; стаж підтверджується документально	імпутація не виконується; невизначеність передається на етап моделювання
Відсутність значення категоріальної ознаки	не зазначено наявності пільгового стажу	неповнота первинного документа	окрема градація «не визначено», врахування у моделі як окремого стану

Приклад заповнення пропуску в часовому ряді наведено у табл. 2.11 і 2.12: у вихідному ряді (табл. 2.11) відсутнє значення за лютий 2021 року, яке після оброблення (табл. 2.12) відновлено за наявними сусідніми спостереженнями.

Таблиця 2.11 – Фрагмент часового ряду середнього розміру пенсії з пропущеним значенням

На дату	Середній розмір пенсії, грн
01.01.2019	2645,70
01.01.2020	NaN
01.01.2021	3507,51

Оброблення: середнє по наявними сусідніми спостереженнями: $(2645,70 + 3507,51) / 2 = 3076,61$; пропущене значення у 01.01.2020 замінюється на 3076,61.

Таблиця 2.12 – Фрагмент часового ряду після заповнення пропущеного значення

Дата зрізу	Середній розмір пенсії, грн	Джерело даних
01.01.2019	2645,70	статистичні дані
01.01.2020	3076,61	Відновлено як середнє сусідніх спостережень
01.01.2021	3507,51	статистичні дані

Для оброблення пропусків кількісних змінних застосовувалося заповнення середніми або медіанними значеннями по вибірці, що дозволяє зберегти середню тенденцію ряду без суттєвого спотворення статистичних характеристик. Для даних із вираженою сезонністю або трендом використовувалася екстраполяція за часовим рядом на основі лінійної регресії, що забезпечує більш точне відновлення пропущених спостережень відповідно до динаміки показника [88].

Застосування лінійної інтерполяції для відновлення пропущеного спостереження ілюструють табл. 2.13 і табл. 2.14: відсутнє значення за серпень 2021 року обчислено як півсуму значень за суміжні місяці [88, 114, 115].

Таблиця 2.13 – Фрагмент часового ряду чисельності пенсіонерів із пропущеним значенням

Дата зрізу	Чисельність одержувачів пенсій, тис. осіб
01.01.2019	11470
01.01.2020	NaN
01.01.2021	11131

А якщо виконати заповнення пропущених значень використовуючи лінійну інтерполяцію, то отримано фрагмент ряду, представлений у табл. 2.14.

Таблиця 2.14 – Фрагмент часового ряду чисельності одержувачів пенсій після заповнення пропуску

Дата зрізу	Чисельність одержувачів пенсій, тис. осіб	Джерело даних
01.01.2019	11470	статистичні дані
01.01.2020	11300,5	відновлено лінійною інтерполяцією
01.01.2021	11131	статистичні дані

При роботі з багатовимірними даними у сфері пенсійного забезпечення одночасно використовується кілька десятків змінних. Така кількість ознак за обмеженої довжини часових рядів призводить до нестійкості оцінок параметрів моделей та до перенавчання. Тому в методиці передбачено етап зменшення розмірності, для якого застосовуються метод головних компонент та процедури відбору ознак.

Метод головних компонент [100] дає можливість звести набір вихідних змінних до меншої кількості нових змінних - головних компонент, які є

лінійними комбінаціями початкових і впорядковані за часткою поясненої варіації. Процедура передбачає центрування та стандартизацію матриці даних, обчислення кореляційної матриці, знаходження її власних векторів і власних значень, формування компонент як лінійних комбінацій вихідних змінних з коефіцієнтами, що дорівнюють складовим власних векторів. Послідовність обчислень має вигляд. Спочатку виконується центрування матриці даних X розмірності $n \times p$

$$\tilde{X} = X - I \cdot \bar{x}^T \quad (2.8)$$

далі обчислюється кореляційна (або коваріаційна) матриця

$$C = (1 / (n - 1)) \cdot \tilde{X}^T \tilde{X} \quad (2.9)$$

для якої розв'язується задача на власні значення

$$C \cdot v_j = \lambda_j \cdot v_j, \quad j = 1, \dots, p \quad (2.10)$$

Частка варіації, яку пояснює j -та компонента, визначається як

$$w_j = \lambda_j / (\lambda_1 + \lambda_2 + \dots + \lambda_p) \quad (2.11)$$

а сама компонента є лінійною комбінацією центрованих і стандартизованих вихідних змінних

$$K_j = v_{j1} \cdot \tilde{x}_1 + v_{j2} \cdot \tilde{x}_2 + \dots + v_{jp} \cdot \tilde{x}_p \quad (2.12)$$

де v_j - власний вектор матриці C , що відповідає власному значенню λ_j ; v_{ji} - його i -та складова.

Власні значення визначають частку варіації, яку пояснює кожна компонента, що є критерієм для вибору кількості компонент, які залишаються в моделі: залишаються компоненти, сукупна частка поясненої варіації яких перевищує заданий рівень.

Для набору показників чисельності одержувачів пенсій за категоріями - за віком, по інвалідності, соціальних пенсій - які є сильно корельованими, застосування методу дало можливість перейти до однієї агрегованої компоненти без істотної втрати інформативності. Після цього перерахунок фактора інфляції

дисперсії для моделі з новими ознаками показав значення, нижчі за 2 для всіх компонент, тобто мультиколінеарність було усунено.

Відбір ознак є альтернативним підходом, за якого зменшення розмірності досягається не побудовою нових змінних, а виключенням малоінформативних вихідних. Розрізняють три групи процедур [88, 100, 115, 116]: фільтраційні, за яких ознаки оцінюються незалежно від моделі за статистичними критеріями зв'язку з цільовою змінною; обгорткові, за яких склад ознак підбирається за якістю моделі, побудованої на кожному підмножині, зокрема методом рекурсивного вилучення ознак; вкладені, за яких відбір виконується безпосередньо в процесі оцінювання параметрів моделі, наприклад регуляризацією за методом LASSO [117], за якою оцінки параметрів визначаються з умови

$$\beta^* = \arg \min \{ \sum (y_i - x_i^T \beta)^2 + \alpha \cdot \sum |\beta_j| \} \quad (2.13)$$

де y - вектор значень цільової змінної; X - матриця значень ознак; β - вектор параметрів; α - параметр регуляризації, за зростання якого частина складових вектора β обертається на нуль, що і реалізує відбір ознак.

Альтернативою є відбір за оцінками важливості ознак ансамблевих алгоритмів [88, 118, 119].

Вибір між побудовою головних компонент та відбором ознак визначається постановкою задачі. Головні компоненти забезпечують більш повне збереження інформації, однак ускладнюють інтерпретацію, оскільки компонента є комбінацією різнорідних показників. Відбір ознак зберігає інтерпретованість, що істотно для обґрунтування управлінських рішень, однак може призвести до втрати частини інформації. У роботі застосовано обидва підходи: метод головних компонент для груп технічно корельованих показників однієї природи, відбір ознак - для формування остаточного складу чинників моделі.

Приклад на реальних даних. Умовний набір даних, на прикладі якого показано застосування методу головних компонент, наведено у табл. 2.15. Наведені змінні характеризують чисельність одержувачів пенсій за трьома

категоріями і є сильно корельованими між собою, що робить доцільним переходити до меншої кількості агрегованих компонент.

Таблиця 2.15 – Умовний набір даних про чисельність одержувачів пенсій за категоріями для демонстрації методу головних компонент

Дата зрізу	Кількість одержувачів пенсій (окремі види), тис осіб		
	x ₁ (пенсії за віком)	x ₂ (пенсії по інвалідності)	x ₃ (пенсії по втраті годувальника)
01.01.2016	9408	1400	726
01.01.2017	9116	1395	695
01.01.2018	8922	1401	663
01.01.2019	8693	1415	646
01.01.2020	8535	1454	649
01.01.2021	8330	1479	641
01.01.2022	8052	1491	638

Умовний набір даних, на прикладі якого проілюстровано стратегію фільтра, наведено у табл. 2.16: для кожної ознаки обчислюється коефіцієнт кореляції Пірсона з цільовою змінною, а ознаки зі значенням, меншим за встановлений поріг, вилучаються з моделі.

Таблиця 2.16 – Результати застосування методу головних компонент до даних табл. 2.15

Компонент	Власне значення	Частка поясненої дисперсії, %	Накопичена частка, %	Навантаження		
				x ₁	x ₂	x ₃
K ₁	3,157	90,19	90,19	0,605	–0,564	0,561
K ₂	0,329	9,39	99,58	–0,016	0,697	0,717
K ₃	0,015	0,42	100,00	0,796	0,443	–0,413

Таблиця 2.17 – Умовний набір даних для демонстрації відбору ознак за коефіцієнтом кореляції

Номер спостереження	чисельність одержувачів пенсій (X_1), тис. осіб	випускники університетів (X_3), тис. осіб	частка витрат на соціальні виплати (Y), %
1	10,5	30	14,0
2	10,2	32	13,7
3	10,0	31	15,0

Обчислення кореляції:

$R = X_1, Y=0,85$ (сильний зв'язок),

$R = X_2, Y=0,7$ (сильний зв'язок),

$R = X_3, Y=0,1$ (слабкий зв'язок).

Отже, ознака X_3 (кількість випускників університетів) має слабкий зв'язок із часткою витрат на соціальні виплати, тому її можна виключити.

Заключним етапом методики є формування вхідних масивів для моделювання. Будуються часові ряди з узгодженою періодичністю. Враховуються часові лаги впливу чинників. Вибірki розбиваються на навчальний, валідаційний та тестовий набори, групуються ознаки за групами чинників. Врахування лагів є обов'язковим, оскільки вплив демографічних чинників на обсяг видатків проявляється не одночасно зі зміною самого чинника: зростання чисельності осіб, які досягають пенсійного віку, впливає на видатки того ж року, тоді як зміна показників стану здоров'я населення впливає на чисельність одержувачів пенсій по інвалідності з лагом у кілька років. У загальному вигляді модель з урахуванням лагів записується як [95]

$$y_t = f(x_{1,t-\tau_1}, x_{2,t-\tau_2}, \dots, x_{m,t-\tau_m}) + \varepsilon_t \quad (2.14)$$

де y_t - значення прогнозованого показника у момент t ; $x_{k,t-\tau_k}$ - значення k -го чинника з лагом τ_k ; ε_t - випадкова складова. Значення лагів визначається аналізом часткових автокореляцій та кореляційних функцій із затримкою [43, 111].

Складовою, якою методику розширено порівняно з наявними підходами, є забезпечення зворотного зв'язку стосовно ефективності управлінських рішень,

що приймаються органами ПФУ. Для цього неструктуровані дані використовуються не лише як джерело інформації про суспільні чинники на вході моделі, а й як засіб оцінювання результативності рішень на виході. Звернення громадян, отримані через вебпортал електронних послуг та контакт-центр, після кластеризації за тематикою [7, 10] дають можливість встановити, які саме рішення викликали найбільшу кількість запитів і скарг, і в які терміни після їх ухвалення. Це формує контур зворотного зв'язку, передбачений методом, розробленим в роботі: результат прийнятого рішення оцінюється не лише за фактичним значенням прогнозованого показника, а й за реакцією одержувачів послуг. Порівняння частоти тематичних кластерів звернень з періодами набрання чинності змінами законодавства дає можливість вчасно виявити проблеми та скоригувати як подальші рішення, так і склад чинників моделі.

Таким чином, удосконалена методика забезпечує формування інформаційно повної, узгодженої та придатної до моделювання вибірки, що відображає комплексний вплив чинників різної природи, коректно опрацьовує пропуски, які не підлягають імпутації, та містить контур зворотного зв'язку щодо результативності рішень.

2.3 Методи прогнозування контингенту одержувачів пенсій

Зростання чисельності одержувачів пенсій є одним із ключових чинників, що впливають на збільшення видатків на державне пенсійне забезпечення, тому прогнозування контингенту одержувачів є однією з основних задач підтримки прийняття рішень у системі. Найчисленнішою категорією є одержувачі пенсій за віком, і саме динаміка цієї категорії визначає основну частину видаткової частини бюджету ПФУ.

Традиційні підходи до розв'язання цієї задачі побудовані на тому, що є демографічна статистика, і, кількість осіб, які набудуть права на пенсію в певному році, оцінюється за чисельністю відповідної вікової групи з поправкою на смертність та міграцію. В умовах реформування системи призначення пенсій,

військової агресії, достовірні дані обліку населення, зокрема, міграції та статевовікової структури в період воєнного стану, недоступні в повному обсязі. Крім того, досягнення пенсійного віку в чинній системі не є достатньою умовою призначення пенсії: право виникає лише за наявності необхідного страхового стажу, та вимоги до нього диференційовані для віку 60, 63 та 65 років [19]. Отже, чисельність вікової групи не визначає чисельність новопризначених пенсій.

Тому в роботі розроблено метод прогнозування контингенту одержувачів пенсій, який враховує індивідуальні характеристики осіб, застрахованих у системі загальнообов'язкового державного пенсійного страхування, що визначають їх право на призначення пенсій за віком та рік настання цього права. Склад цих характеристик наведено в табл. 2.2. Задача формулюється так: за наявними в реєстрі застрахованих осіб відомостями про страховий стаж, дату народження та інші характеристики осіб певного року народження оцінити чисельність осіб, яким пенсія за віком буде призначена у відповідному календарному році, з розподілом за віком настання права - 60, 63 або 65 років.

Схему методу прогнозування кількості новопризначених пенсій за віком протягом одного календарного року із використанням ймовірнісного моделювання наведено на рис. 2.2.

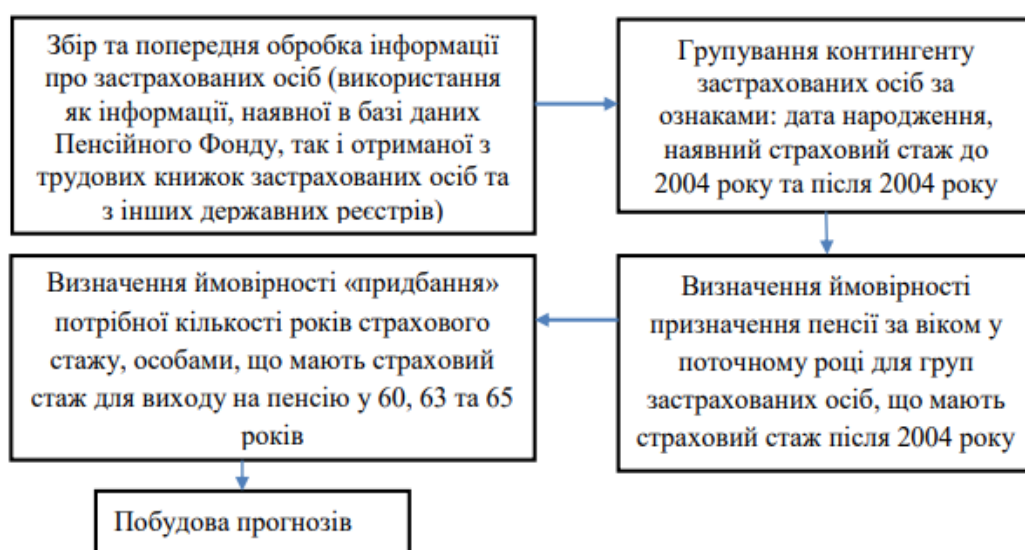


Рисунок 2.2 - Схема прогнозування кількості новопризначених пенсій за віком у певному році із використанням ймовірнісного моделювання

Етап 1. Збір та попереднє оброблення інформації про застрахованих осіб. Використовується як інформація, наявна в базі даних Пенсійного фонду України, так і отримана з трудових книжок застрахованих осіб та з інших державних реєстрів. Останнє істотно: частина відомостей про стаж за період до 2004 року існує лише у документальній формі і надходить до системи в міру звернення особи, тому набір даних на момент побудови прогнозу є свідомо неповним.

Етап 2. Групування контингенту застрахованих осіб за ознаками: дата народження, наявний страховий стаж до 2004 року та наявний страховий стаж після 2004 року. Групування за датою народження визначає рік настання пенсійного віку. Розділення стажу за періодами є ключовою особливістю методу: воно дає можливість розмежувати частину стажу, відому за даними персоніфікованого обліку, та частину, яка містить пропуски, що не підлягають імпутації. Для кожної групи оцінюються статистичні характеристики розподілу стажу - математичне сподівання та стандартне відхилення - окремо для стажу за весь період трудової діяльності, для стажу до 2004 року та для стажу після 2004 року. Розподіл загального стажу розглядається як комбінація розподілів двох складових.

Етап 3. Визначення ймовірності набуття потрібної кількості років страхового стажу особами, які мають страховий стаж, достатній для виходу на пенсію у 60, 63 або 65 років. Цей етап враховує передбачену законодавством можливість укладення договору про добровільну участь у системі загальнообов'язкового державного пенсійного страхування, за яким особа може набути стаж, якого їй не вистачає [19, 97]. Оскільки рішення про це є поведінковим вибором особи, подія має ймовірнісний характер, а ймовірність суттєво залежить від обсягу стажу, який необхідно набутися: чим більше років потрібно, тим нижча ймовірність такого рішення. Для кожної групи оцінюється ймовірність набуття від одного до п'яти років стажу; обсяг понад п'ять років у методі не розглядається як практично реалізований протягом одного року.

Етап 4. Визначення ймовірності призначення пенсії за віком у поточному році для груп застрахованих осіб. На цьому етапі для кожної групи обчислюється

ймовірність настання права на призначення пенсії з урахуванням наявного стажу, ймовірності його набуття за договором про добровільну участь, а також якісних ознак - статі, наявності пільгового стажу за особливі умови праці та наявності права на дострокове призначення пенсії або зменшення пенсійного віку. Для сумісного врахування неперервних характеристик (стаж за періодами, вік) та дискретних ознак застосовано апарат гібридних мереж Байєса [120-124], у якому умовний розподіл цільової вершини залежить від значень батьківських вершин, причому умовний розподіл цільової вершини має вигляд

$$p(y | d, z) = N(\mu_d + w_{d1} \cdot z_1 + \dots + w_{dk} \cdot z_k, \sigma_d^2) \quad (2.15)$$

де y - значення цільової вершини; d - конфігурація значень дискретних батьківських вершин (стать, наявність пільгового стажу, право на дострокову пенсію); $z_1, \dots, z_k$ - значення неперервних батьківських вершин (стаж за періодами, вік); μ_d і σ_d^2 - математичне сподівання та дисперсія розподілу для конфігурації d ; w_{di} - вагові коефіцієнти, що враховують ступінь впливу неперервних батьківських вершин.

Етап 5. Побудова прогнозів. Повна ймовірність призначення пенсії у віці a для групи g обчислюється з урахуванням як наявного стажу, так і можливості його набуття за договором про добровільну участь:

$$P_a(g) = P(S \geq s(a)) + \sum P(s(a) - \Delta \leq S < s(a)) \cdot P_{доб}(\Delta), \Delta = 1, \dots, 5 \quad (2.16)$$

де $P_{доб}(\Delta)$ - ймовірність набуття Δ років страхового стажу за договором про добровільну участь. Прогнозна чисельність новопризначених пенсій у віці a становить

$$N_a = \sum n_g \cdot P_a(g) \quad (2.17)$$

а загальна прогнозна чисельність новопризначених пенсій за віком у році T .

$$N(T) = N_{60} + N_{63} + N_{65} \quad (2.18)$$

де n_g - чисельність застрахованих осіб у групі g , сформованій за датою народження та обсягом стажу за періодами. Результати агрегуються до загальної

чисельності контингенту і передаються як вхідна змінна до моделей макрорівня, що реалізує узгодження прогнозів за ієрархією управління.

Особливістю пропонованого підходу є те, що як вхідні дані використані часові ряди, що містять інформацію про страховий стаж застрахованих осіб в інформаційній системі загальнообов'язкового державного пенсійного страхування і містять значну частину пропусків у період до 2004 року, які не можна заповнювати без підтверджувальних документів.

Метод опрацьовує цю невизначеність не шляхом її усунення, а шляхом переходу від детермінованої постановки до ймовірнісної. Замість висновку «особа має або не має права на пенсію» оцінюється ймовірність настання права, причому невизначеність щодо стажу за період до 2004 року враховується через розподіл його можливих значень, побудований за частотним аналізом наявних даних та за припущеннями щодо рівня зайнятості населення у відповідні періоди. Такі припущення формулюються експертно і є частиною методу: невизначеність, яку неможливо усунути даними, опрацьовується поєднанням статистичної інформації з експертними судженнями. Це визначає і форму результату. Прогноз подається не єдиним числом, а розподілом чисельності контингенту за категоріями: особи, яким пенсія буде призначена з настанням пенсійного віку за наявного стажу; особи, яким пенсія буде призначена за умови набуття стажу за договором про добровільну участь; особи, яким пенсія буде призначена пізніше - у 63 або 65 років. Така форма безпосередньо відповідає потребам прийняття рішень: вона показує не лише очікувану чисельність, а й те, яка її частина залежить від поведінкового вибору застрахованих осіб, і, отже, які саме групи потребують адресного консультування.

Розроблений метод прогнозування контингенту одержувачів пенсій відрізняється стійкістю, високою точністю та адаптивністю за рахунок поєднання сценарного аналізу, інтелектуального аналізу даних, експертного оцінювання та багатомодельного підходу. Він забезпечує врахування як багаторівневості та ієрархічності прийняття рішень через побудову моделей мікро- та макрорівнів, так і умов невизначеності - за рахунок методики

прогнозування кількості новопризначених пенсій за віком у певному році, за якої використовуються часові ряди даних про трудовий стаж осіб-кандидатів із значною кількістю неімпутованих пропусків, поєднана з використанням експертних суджень для визначення права на призначення пенсій.

Відмінними ознаками методу порівняно з наявними підходами є: перехід від демографічної основи прогнозування до основи, що спирається на індивідуальні дані персоніфікованого обліку, що робить прогноз незалежним від доступності демографічної статистики; розділення страхового стажу за періодами до та після 2004 року як окремих джерел інформації з різним рівнем повноти; врахування набуття стажу за договором про добровільну участь як ймовірнісної події; сумісне опрацювання кількісних та якісних характеристик застрахованих осіб засобами гібридних мереж Байєса; формування результату у вигляді розподілу за категоріями, а не точкової оцінки. Метод є універсальним для розв'язання задач опису та прогнозування розвитку системи державного пенсійного забезпечення та її складових в умовах інформаційної та статистичної невизначеностей [2].

Другою за впливом на видаткову частину бюджету категорією одержувачів, чисельність якої в поточних умовах зростає найшвидше, є одержувачі пенсій по інвалідності. Зростання кількості постраждалих, зокрема з числа військовослужбовців, внаслідок активної фази військового конфлікту потребує значних матеріальних та грошових витрат, тому актуальною є задача прогнозування чисельності цієї категорії. Особливість задачі полягає в тому, що традиційні підходи, побудовані на екстраполяції наявних тенденцій захворюваності та інвалідизації, для неї непридатні: чисельність визначається не демографічними процесами, а перебігом бойових дій.

Тому метод побудовано на іншій основі: чисельність осіб, які набудуть право на пенсію по інвалідності внаслідок військової агресії, оцінюється як частина санітарних втрат, а самі санітарні втрати прогнозуються за чинниками, що визначають їх обсяг і структуру. Ланцюг оцінювання має вигляд: чинники бойової обстановки та медичного забезпечення - обсяг і структура санітарних

втрат - чисельність осіб, яким встановлено інвалідність, з розподілом за групами - чисельність одержувачів пенсій по інвалідності та обсяг відповідних видатків [4].

Для прогнозування санітарних втрат застосовано апарат мереж Байєса, вершинами якої є чинники, вплив яких на кількість санітарних втрат за оцінками експертів є найсуттєвішим: інтенсивність бою, тип застосовуваної зброї, ступінь тяжкості поранення, рівень захисту особового складу, час евакуації, стан медичного забезпечення та логістики, характер ушкоджень. Вибір цього апарату зумовлений тим, що для більшості зазначених чинників кількісні статистичні дані відсутні або мають обмежений доступ, а наявні відомості є експертними оцінками у формі розподілів ймовірностей за градаціями. Мережа Байєса дає можливість коректно поєднати такі оцінки та врахувати умовні залежності між чинниками: спільний розподіл усіх змінних мережі факторизується за структурою графа [4, 88. 120-124]

$$P(X_1, X_2, \dots, X_n) = \prod P(X_i | pa(X_i)) \quad (2.19)$$

де $pa(X_i)$ - множина батьківських вершин вершини X_i . Чисельність осіб, яким буде встановлено інвалідність з числа санітарних втрат, оцінюється як

$$I = L \cdot \sum P(t_k) \cdot P(inv | t_k) \quad (2.20)$$

де L - прогнозований обсяг санітарних втрат; $P(t_k)$ - ймовірність k -го ступеня тяжкості поранення, отримана за мережею; $P(inv | t_k)$ - умовна ймовірність встановлення інвалідності за відповідного ступеня тяжкості.

Припущення щодо співвідношення різних видів бойових втрат сформульовано за наявною інформацією щодо втрат у попередні періоди. Санітарні втрати становлять переважну більшість втрат, причому їх кількість коливається залежно від інтенсивності та характеру бойових дій, рівня захисту особового складу і своєчасності евакуації.

Метод реалізовано у складі інтелектуальної системи підтримки прийняття рішень, яка є модульним програмним комплексом, призначеним для опрацювання даних про санітарні втрати з урахуванням специфіки предметної області. Особливістю комплексу є наявність окремих підсистем попереднього оброблення та збору вхідних даних, оскільки дані про санітарні та безповоротні втрати надходять з різних джерел, з різною періодичністю та з різним рівнем повноти. Розроблені методи прогнозування є мікрорівневою складовою прогнозного інструментарію та реалізовані у відповідній інформаційній технології, інтегрованій до ІКІС ПФУ [4, 23].

2.4 Інформаційна технологія прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ для прийняття рішень у системі державного пенсійного забезпечення

Методи, представлені в роботі, реалізовані в інформаційній технології прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ. Технологія забезпечує послідовне виконання етапів, передбачених методом на рис. 1.6, стосовно двох взаємозалежних задач - прогнозування чисельності та структури контингенту одержувачів пенсій і прогнозування показників бюджету, що визначаються цією чисельністю.

Функціонально технологія включає складові збору та інтеграції різнорідних даних, попереднього оброблення з опрацюванням пропусків та аномалій, побудови моделей мікро- та макрорівнів, сценарного аналізу, оцінювання адекватності моделей та якості прогнозів, візуалізації результатів, а також розмежування доступу до інформаційних ресурсів. Останнє є обов'язковою вимогою, оскільки технологія працює з персональними даними застрахованих осіб [23, 96]: для побудови моделей використовуються агрегати та згенеровані набори, що відповідають структурі реальних даних, тоді як доступ до персоніфікованих відомостей обмежений відповідно до чинних регламентів.

У роботі запропоновано удосконалення зазначеної інформаційної технології. Підставою для нього стали виявлені потреби користувачів Пенсійного фонду України в адресному консультуванні: результати прогнозування, отримані на мікрорівні, показують не лише очікувану чисельність контингенту, а й те, які групи застрахованих осіб не набудуть права на пенсію без набуття додаткового стажу. Ця інформація має практичну цінність лише за умови, що вона доведена до конкретної особи у формі зрозумілої рекомендації, а не залишається агрегованою оцінкою. Складові технології та їх відповідність задачам наведено в табл. 2.18.

Таблиця 2.18 - Складові інформаційної технології та їх призначення

Складова	Призначення	Задача, яку обслуговує
збір та інтеграція даних	отримання даних з внутрішніх і зовнішніх джерел, приведення до єдиного подання, синхронізація періодичності	обидві задачі
попереднє оброблення	очищення, виявлення аномалій, стандартизація, опрацювання пропусків без імпутації там, де вона недопустима	обидві задачі
моделі мікрорівня	оцінювання ймовірності настання права на призначення пенсії за індивідуальними характеристиками	прогнозування контингенту
моделі макрорівня	оцінювання обсягу видатків і надходжень за групами агрегованих чинників	прогнозування показників бюджету
сценарний аналіз	формування варіантів розвитку подій та розрахунок прогнозів за кожним сценарієм	обидві задачі
оцінювання якості	перевірка адекватності моделей та якості оцінок прогнозів за системою критеріїв	обидві задачі
рекомендаційна підсистема	формування рекомендацій та проекту рішення, адресне консультування	обидві задачі
візуалізація та розмежування доступу	подання результатів у формі таблиць, графіків та інформаційних панелей з урахуванням прав доступу	обидві задачі

Тому технологію доповнено рекомендаційною підсистемою, адаптованою до ІКІС ПФУ. У ній поєднано використання експертних суджень з алгоритмами штучного інтелекту. Експертні судження задають правила, що відображають норми законодавства та практику їх застосування, тобто визначають, які саме рекомендації є допустимими для кожної ситуації. Алгоритми штучного інтелекту забезпечують віднесення конкретного випадку до відповідної ситуації, зокрема опрацювання формулювань звернення особи та підбір релевантних норм і роз'яснень. Таке поєднання є принциповим: правила гарантують правову коректність рекомендації, а алгоритми - швидкість її формування. Функціонально підсистема виконує такі дії: за характеристиками застрахованої особи визначає, у якому віці за наявного стажу настане право на призначення пенсії; оцінює, який обсяг додаткового стажу потрібен для настання права у більш ранньому віці; формує перелік допустимих варіантів дій особи з посиланням на відповідні норми; готує проект відповіді або консультації для розгляду відповідальною посадовою особою. Проект рішення не є остаточним: підсистема готує його, а рішення ухвалює посадова особа, що відповідає принципу збереження відповідальності людини.

Практичний ефект удосконалення полягає в тому, що, завдяки поєднанню експертних суджень з алгоритмами штучного інтелекту, забезпечується оперативність прийняття рішення, зменшується вплив людського фактору (особливо під час аналізу законодавства), спрощується процедура аналізу та інтерпретації отриманих результатів прогнозування, та обґрунтованість рішень в умовах невизначеності, у тому числі таких, що приймаються в умовах війни.

Адаптація технології до ІКІС ПФУ виконана відповідно до існуючих регламентів та нормативів. Нові складові не замінюють наявні підсистеми, а вбудовуються у чинні процеси оброблення інформації з використанням тих самих джерел даних, механізмів розмежування доступу та інтерфейсів взаємодії. Це дає можливість поетапного впровадження зі збереженням працездатності системи на кожному етапі. Технологія відрізняється гнучкістю та здатністю до масштабування, що забезпечує її подальше використання як у ПФУ, так і в

системах підтримки прийняття рішень органів державного управління та місцевого самоврядування, що опікуються соціальними питаннями. Теоретичні та методичні засади інтеграції аналітичної підсистеми до Єдиної інформаційної системи соціальної сфери викладено в [5]; відповідно, розроблені моделі та методи прогнозування у подальшому мають бути адаптовані до використання у складі ЄІССС [23, 25, 130].

Висновки до розділу 2

1. Розглянуто методичні підходи до побудови моделей та прогнозів у системі державного пенсійного забезпечення, проаналізовано їх використання для підтримки прийняття рішень за сучасних умов та за переходу до трирівневої пенсійної моделі. Виявлено, що в умовах складної фінансової ситуації в країні актуальним питання є забезпечення фінансування пенсійних виплат, тому основний акцент зроблено на чинниках, що здійснюють найбільший вплив на зміну пенсійних виплат.

2. Установлено, що важливо врахувати вплив як економічних, демографічних, екологічних, суспільних та правових чинників, а також вплив воєнного стану, визначальних для розвитку системи державного пенсійного забезпечення в цілому, так і індивідуальних характеристик осіб, застрахованих у системі загальнообов'язкового державного пенсійного страхування. Оскільки ці групи чинників мають різну природу і не можуть бути враховані в межах однієї моделі, запропоновано поєднання моделей мікро- та макрорівнів з агрегуванням результатів мікрорівня як вхідної змінної моделей макрорівня.

3. Удосконалено методику використання різномірних даних, отриманих з різних джерел, як для опису та прогнозування процесів, характерних для системи державного пенсійного забезпечення в умовах невизначеності, так і для забезпечення зворотного зв'язку стосовно ефективності управлінських рішень, що приймаються органами ПФУ. Методику розширено процедурою

опрацювання пропусків, які не підлягають імпутації, та складовою оцінювання результативності рішень за результатами кластеризації звернень громадян.

4. Обґрунтовано принципову неможливість заповнення пропусків у даних про страховий стаж за період до 2004 року: відсутність значення в таких часових рядах є не статистичним пропуском. Запропоновано зберігати таку невизначеність як самостійну характеристику задачі та опрацьовувати її ймовірнісними засобами на етапі моделювання.

5. Розроблено метод прогнозування контингенту одержувачів пенсій, який враховує індивідуальні характеристики застрахованих осіб, що визначають їх право на призначення пенсій за віком та рік настання цього права. Метод відрізняється стійкістю, високою точністю та адаптивністю за рахунок поєднання сценарного аналізу, інтелектуального аналізу даних, експертного оцінювання та багатомодельного підходу, забезпечує врахування багаторівневості прийняття рішень та умов невизначеності і є універсальним для задач опису та прогнозування розвитку системи державного пенсійного забезпечення в умовах інформаційної та статистичної невизначеностей.

6. Розроблено метод прогнозування зростання контингенту одержувачів пенсій по інвалідності внаслідок військової агресії, побудований на оцінюванні санітарних втрат засобами мереж Байєса, що дає можливість отримати прогноз за умов, коли кількісні статистичні дані або мають обмежений доступ.

7. Удосконалено інформаційну технологію прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ за рахунок додавання рекомендаційної підсистеми, у якій поєднано використання експертних суджень з алгоритмами штучного інтелекту, що забезпечує прозорість процесу прогнозування, спрощує аналіз отриманих результатів та обґрунтованість рішень в умовах невизначеності, у тому числі таких, що приймаються оперативно, в умовах війни. Технологія адаптована до ІКІС ПФУ за конвергентним підходом і відрізняється гнучкістю та здатністю до масштабування.

РОЗДІЛ 3

МЕТОДОЛОГІЯ ПРОГНОЗУВАННЯ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ

3.1 Методичні підходи урахування особливостей системи державного пенсійного забезпечення при побудові моделей та прогнозів в умовах невизначеності

Моделювання та прогнозування витрат на пенсійне забезпечення, яке є найбільш вагомою частиною витрат на соціальне забезпечення є особливо важливою задачею в умовах демографічної ситуації, що має негативні тенденції, макроекономічної нестабільності, війсьового конфлікту. Управління бюджетними ресурсами Пенсійного фонду потребує не лише точкових оцінок чисельності одержувачів пенсій і планування потреби у витратах на виплату пенсій на наступний місяць, а прогнозування потреби у фінансових ресурсах на перспективу. Таким чином, формалізуючи, дану проблему можна представити наступним чином.

Нехай у момент часу t відомі значення вектора факторних ознак $X_t = (X_{1t}, X_{2t}, \dots, X_{pt})^T$ та попередні значення цільового показника $Y_t, Y_{t-1}, \dots$. Тоді, задача прогнозування на горизонті h полягає у побудові відображення F , яке ставить у відповідність наявній інформації оцінку майбутнього значення цільового показника [96, 115]:

$$\hat{Y}_{t+h} = F(Y_t, \dots, Y_{t-q}; X_t, \dots, X_{t-r}; \theta) \quad (3.1)$$

де $\hat{Y}_{t+h}$ — прогнозна оцінка цільового показника на h кроків уперед; q — глибина врахування передісторії цільової змінної; r — максимальний лаг факторних ознак; θ — вектор параметрів моделі, що підлягає оцінюванню; F — функція, вигляд якої визначається обраним класом моделей.

Оптимальна модель обирається з деякої множини допустимих моделей $\mathcal{F}$ за критерієм мінімуму очікуваних втрат від помилки прогнозу [96, 115]:

$$F^* = \arg \min \mathbb{E}[\mathcal{L}(Y_{t+h} - \hat{Y}_{t+h})], F \in \mathcal{F} \quad (3.2)$$

де $\mathcal{A}(\cdot)$ - функція втрат (квадратична, абсолютна або асиметрична); $E[\cdot]$ - оператор математичного сподівання. Вибір асиметричної функції втрат є методологічно обґрунтованим для бюджетного планування: недооцінка потреби у фінансуванні спричиняє касові розриви у виплаті пенсій, тоді як переоцінка призводить лише до тимчасового відволікання ресурсів, тобто ці помилки мають різну управлінську вагу.

Постановка задачі охоплює такі складові, як визначення цільових змінних, формування факторної множини, урахування часових лагів, трендів і сезонності.

Формалізуються показники, що відображають потребу у фінансуванні: обсяг витрат на пенсійне забезпечення (у грошовому вимірі та як частка ВВП); частка витрат на соціальну допомогу у ВВП; чисельність одержувачів пенсій і соціальних допомог за видами з урахуванням трендової та сезонної складових.

На основі багатокритеріальних даних, підготовлених за методикою запропонованою у роботі, визначаються групи факторів: демографічні (чисельність населення, чисельність пенсіонерів, очікувана тривалість життя, коефіцієнт демографічного навантаження); економічні (ВВП на душу населення, індекс споживчих цін, рівень безробіття, середня заробітна плата); соціальні (кількість одержувачів допомог за видами, частка внутрішньо переміщених осіб); екологічні (за потреби - показники, що впливають на стан здоров'я населення).

Вплив факторів на цільову змінну має відкладений характер: зміна індексу споживчих цін позначається на витратах через механізм індексації пенсій з лагом у декілька кварталів. Тому, до моделі включаються лагові змінні $X_{j,t-k}$, а сезонні коливання (зокрема зростання виплат при проведенні масових перерахунків пенсій) описуються сезонними компонентами або фіктивними змінними.

Вимоги до побудованих моделей, що визначають їх придатність для практичного застосування, - це стійкість до мультиколінеарності, інтерпретованість, точність.

Стійкість до мультиколінеарності є першою вимогою, оскільки між факторними ознаками існують щільні зв'язки. Наприклад, між чисельністю

пенсіонерів і кількістю одержувачів мінімальних пенсій. Формальним індикатором мультиколінеарності слугує коефіцієнт інфляції дисперсії (3.3) [114].

$$VIF_j = \frac{1}{1 - R^2_j} \quad (3.3)$$

де R^2_j — коефіцієнт детермінації допоміжної регресії j -ї факторної ознаки на решту ознак. Значення $VIF_j > 10$ свідчить про суттєву мультиколінеарність, що потребує застосування методу головних компонент або відбору ознак.

Друга вимога - інтерпретованість, тобто можливість пояснити внесок кожного фактора у прогнозоване значення, що є обов'язковою умовою використання моделі для обґрунтування бюджетних рішень.

Третя вимога — точність, яка перевіряється на ретроспективних даних за різними метриками, в тому числі, запропонованими в роботі.

Емпіричною базою дослідження слугують агреговані показники, наприклад, за 2018–2022 рр. (табл. 3.1).

Таблиця 3.1 – Вихідні дані для побудови моделей прогнозування витрат на пенсійне забезпечення

Рік	Чисельність одержувачів пенсій X_1 , млн осіб	Індекс споживчих цін X_2 , %	Індекс споживчих цін попереднього року $X_2(-1)$, %	ВВП на душу населення X_3 , тис. грн	Частка витрат на пенсійне забезпечення Y , %
2010	13,738	109,1	112,3	23,6	9,3
2011	13,821	104,6	109,1	28,8	7,8
2012	13,640	99,8	104,6	30,9	8,6
2013	13,533	100,5	99,8	32,0	9,5
2014	12,174	124,9	100,5	35,8	8,7
2015	12,296	143,3	124,9	46,2	8,9
2016	11,956	112,4	143,3	55,9	10,8
2017	11,725	113,7	112,4	70,2	9,6
2018	11,470	109,8	113,7	84,2	8,7
2019	11,335	104,1	109,8	94,6	8,1
2020	11,131	105,0	104,1	100,4	8,2
2021	10,841	110,0	105,0	131,9	6,5

Вибір періоду зумовлено тим, що дванадцять спостережень забезпечують мінімально достатню кількість ступенів свободи для перевірки статистичної значущості чотирьох параметрів, а крім того, включення 2022 р. порушує співставність показників в умовах воєнного стану.

Таким чином, задача моделювання полягає у побудові багатofакторних моделей, які поєднують різнорідні змінні, що відображають часову динаміку процесів і забезпечують кількісне оцінювання потреби у фінансуванні пенсійного забезпечення в умовах дії демографічних, економічних і соціальних чинників.

У роботі запропоновано методику, яка реалізує безперервний адаптивний цикл прогнозування (рис. 3.1) [1, 76, 95, 125, 126],

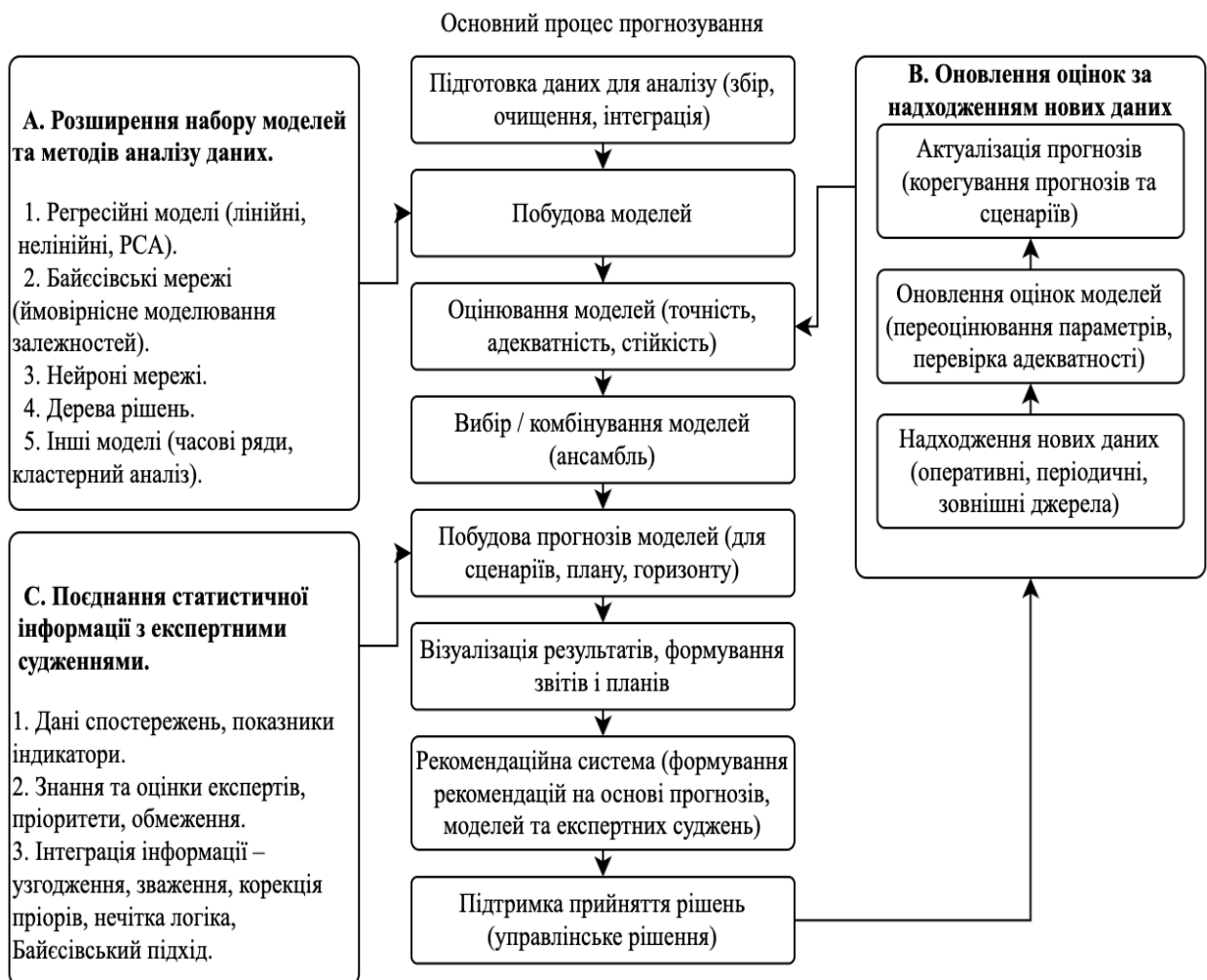


Рисунок 3.1 - Методика адаптивного прогнозування нелінійних процесів різної природи в системі державного пенсійного забезпечення

На першому етапі здійснюється підготовка даних для аналізу, що охоплює їх збір, очищення та інтеграцію. На другому етапі формується набір моделей для прогнозування. При цьому відповідно до принципу розширення модельного інструментарію можуть використовуватися регресійні моделі, байєсівські та нейронні мережі, дерева рішень, часові ряди та інші моделі.

На третьому етапі здійснюється оцінювання побудованих моделей за системою критеріїв, що враховує точність, адекватність і стійкість отриманих результатів. На основі результатів оцінювання на четвертому етапі виконується вибір або комбінування моделей у вигляді ансамблю. На п'ятому етапі формуються прогнози для визначених сценаріїв, планових періодів та прогнозних горизонтів.

Результати прогнозування на шостому етапі візуалізуються та використовуються для формування аналітичних звітів і планів. На сьомому етапі рекомендаційна система формує рекомендації на основі прогнозів, моделей та експертних суджень, після чого на восьмому етапі забезпечується підтримка прийняття управлінських рішень.

Адаптивність методики забезпечується трьома взаємопов'язаними складовими. Блок А забезпечує розширення набору моделей і методів аналізу даних та, відповідно, можливість вибору найбільш адекватного модельного інструментарію для конкретного процесу. Блок В реалізує механізм оновлення оцінок за надходження нових даних: після отримання оперативної, періодичної або зовнішньої інформації здійснюється переоцінювання параметрів моделей, перевірка їх адекватності та актуалізація прогнозів, зокрема за окремими сценаріями. Блок С забезпечує поєднання статистичної інформації з експертними судженнями шляхом їх узгодження, зважування та корекції, з використанням, зокрема, нечіткої логіки та байєсівського підходу.

Таким чином, зворотні зв'язки між основним процесом прогнозування, надходженням нових даних та інтеграцією статистичної і експертної інформації формують безперервний адаптивний цикл, у межах якого результати прогнозування можуть бути уточнені внаслідок зміни даних, оцінок моделей,

сценарних умов або експертних суджень. Це забезпечує адаптацію прогнозного інструментарію до динаміки досліджуваних нелінійних процесів та умов невизначеності. Розроблено системну методологію побудови прогнозів для підтримки прийняття рішень щодо пенсійного забезпечення населення в умовах невизначеності (рис. 3.2)

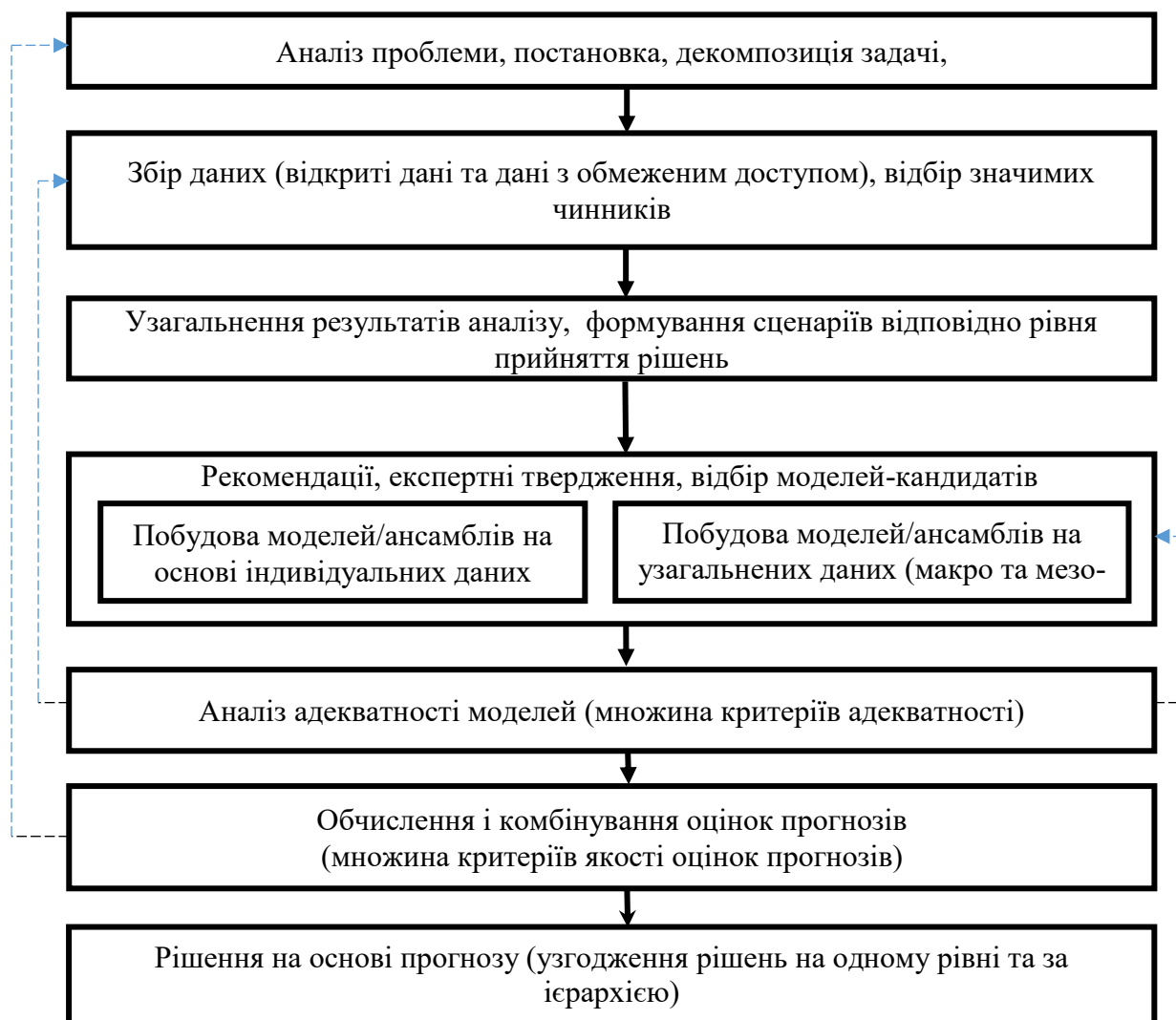


Рисунок 3.2 - Методологія побудови прогнозів для підтримки прийняття рішень щодо пенсійного забезпечення населення в умовах невизначеності (узагальнено)

У даній методології поєднані математичні моделі та їх ансамблі, алгоритми машинного навчання та сценарний аналіз, методи опрацювання текстової інформації. Їх використання є комплексним, забезпечує стійкість результатів, урахування сукупного впливу різних груп чинників в процесі прийняття

оперативних та стратегічних рішень. Результати прогнозування узгоджуються на різних рівнях управління та етапах прийняття рішень. Порівняння із експертними оцінками, бюджетними показниками, тощо створює умови для контролю виконання. Для оцінювання результативності рішень, особливо тих, що приймаються на мікро- рівні, у системі обслуговування населення, реалізовано зворотний зв'язок що дозволяє підвищити якість послуг, що надаються та потреби населення у пенсійному забезпеченні, вчасно виявити проблеми та скоригувати рішення. Таким чином, використання розробки підвищить ефективність функціонування системи державного пенсійного забезпечення як складової соціальної сфери України.

3.2 Обґрунтування застосування математичних моделей різних типів та ансамблів

Для прогнозування витрат на пенсійне забезпечення застосовуються різні класи математичних моделей, що дозволяють враховувати лінійні, нелінійні та ймовірнісні залежності між факторами і цільовими змінними. Лінійна багатофакторна регресія є базовим інструментом кількісного опису залежностей між факторами та цільовим показником. Її перевагою є безпосередня інтерпретованість коефіцієнтів: кожен з них показує приріст цільової змінної за зміни відповідного фактору на одиницю за незмінних інших умов. У загальному випадку, модель має вигляд (3.4):

$$Y_t = \beta_0 + \sum_{j=1}^p \beta_j X_{jt} + \varepsilon_t \quad (3.4)$$

де Y_t - цільова змінна (частка витрат на пенсійне забезпечення у ВВП); X_{jt} — значення j -ї факторної ознаки у момент t ; β_j — оцінювані коефіцієнти, що відображають силу та напрям впливу; β_0 — вільний член; ε_t — випадкова похибка, для якої приймаються стандартні припущення: $E[\varepsilon_t] = 0$, $Var(\varepsilon_t) = \sigma^2$, $Cov(\varepsilon_t, \varepsilon_s) = 0$ при $t \neq s$.

У матричній формі модель записується як $Y = X\beta + \varepsilon$, а оцінки параметрів за методом найменших квадратів визначаються співвідношенням [127]:

$$\beta = (X^T X)^{-1} X^T Y \quad (3.5)$$

де X — матриця плану розмірності $n \times (p+1)$, перший стовпець якої складається з одиниць; Y — вектор спостережень цільової змінної; β — вектор оцінок параметрів.

Статистична значущість окремого коефіцієнта перевіряється за t -статистикою (3.6):

$$t_j = \frac{\beta_j}{SE(\beta_j)}, \quad SE(\beta_j) = \sqrt{\hat{\sigma}^2_{jj}} \quad (3.6)$$

де $SE(\beta_j)$ — стандартна похибка оцінки j -го коефіцієнта; σ^2 — оцінка дисперсії залишків; індекс jj позначає відповідний діагональний елемент матриці $(X^T X)^{-1}$. Коефіцієнт вважається статистично значущим, якщо відповідне p -значення не перевищує прийнятого рівня значущості (як правило, 0,05) [114, 115].

За даними табл. 3.1 побудовано модель залежності частки витрат на соціальний захист від трьох факторів. Оцінювання виконано методом найменших квадратів, результати наведено у табл. 3.2.

Таблиця 3.2 – Результати оцінювання лінійної багатфакторної регресії

Ознака	X_1	X_2	$X_2(-1)$	X_3	Y	VIF_j
X_1 чисельність одержувачів пенсій, тис осіб	1,000	-0,193	-0,126	-0,902	0,270	9,93
X_2 ІСЦ, %	-0,193	1,000	0,373	-0,099	0,129	1,82
$X_2(-1)$, ІСЦ з лагом 1 рік, %	-0,126	0,373	1,000	-0,052	0,586	1,22
X_3 , ВВП на душу населення, тис грн	-0,902	-0,099	-0,052	1,000	-0,525	9,63
Y , частка витрат на пенсійне забезпечення, %	0,270	0,129	0,586	-0,525	1,000	—

Перевірка першої з трьох сформульованих вимог до моделей – стійкості до мультиколінеарності виконується за коефіцієнтом інфляції дисперсії (3.3). Порогове значення $VIF_j > 10$ формально не перевищено, проте для чисельності

пенсіонерів і ВВП на душу населення коефіцієнти становлять 9,93 і 9,63, тобто впритул наближаються до нього. Причиною є високий за модулем коефіцієнт парної кореляції між цими ознаками ($-0,902$): обидві є носіями спільної часової тенденції.

Практичним наслідком є нестійкість оцінок окремих коефіцієнтів до складу вибірки, що обмежує змістовну інтерпретацію їх величин, хоча й не перешкоджає використанню моделі для прогнозування. Це обґрунтовує застосування методу головних компонент для задачі прогнозування витрат на пенсійне забезпечення, де факторна множина є ширшою і мультиколінеарність набуває критичного характеру.

Використовуючи дані, наведені у табл. 3.1, побудовано лінійну багатофакторну модель (3.4). Оцінювання моделі виконано методом найменших квадратів, результати наведено у табл. 3.3.

Таблиця 3.3 – Результати оцінювання лінійної багатофакторної регресії

Змінна	Назва показника	Коефіцієнт	Стандартна похибка	t-статистика	p-значення	Висновок
β_0	вільний член	28,7312	9,1926	3,125	0,0167	значущий
X_1	чисельність пенсіонерів	-1,3523	0,5124	-2,639	0,0335	значущий
X_2	індекс споживчих цін	-0,0448	0,0199	-2,246	0,0595	значущий на рівні 0,10
$X_2(-1)$	ІСЦ попереднього року	0,0448	0,0163	2,743	0,0288	значущий
X_3	ВВП на душу населення	-0,0552	0,0159	-3,475	0,0103	значущий

Модель загалом є статистично значущою ($F = 7,171$ за $p = 0,0127$), коефіцієнт детермінації становить 0,804, скоригований – 0,692. Змістовна інтерпретація коефіцієнтів така. Підвищення індексу споживчих цін попереднього року на 1 в. п. збільшує частку видатків поточного року на 0,0448 в. п. – це прямий кількісний вимір механізму індексації. Водночас підвищення

поточного індексу споживчих цін на ту саму величину зменшує частку на 0,0448 в.п.,: поточна інфляція збільшує показник ВВП (тобто знаменник відношення), швидше, ніж встигають зрости самі витрати. Практично однакові за модулем і протилежні за знаком оцінки цих двох коефіцієнтів показують, що інфляційний чинник діє на цільовий показник двома каналами, які майже повністю компенсують один одного протягом дворічного інтервалу.

Коефіцієнт при змінній «ВВП на душу населення» від’ємний ($-0,0552$) з тієї самої причини – цей показник входить до знаменника цільового відношення. Коефіцієнт при змінній «чисельність одержувачів пенсій» також від’ємний ($-1,3523$), що не має самостійного змістовного тлумачення й пояснюється щільним зв’язком цієї ознаки з показником «ВВП на душу населення». (рис. 3.3, рис. 3.4))

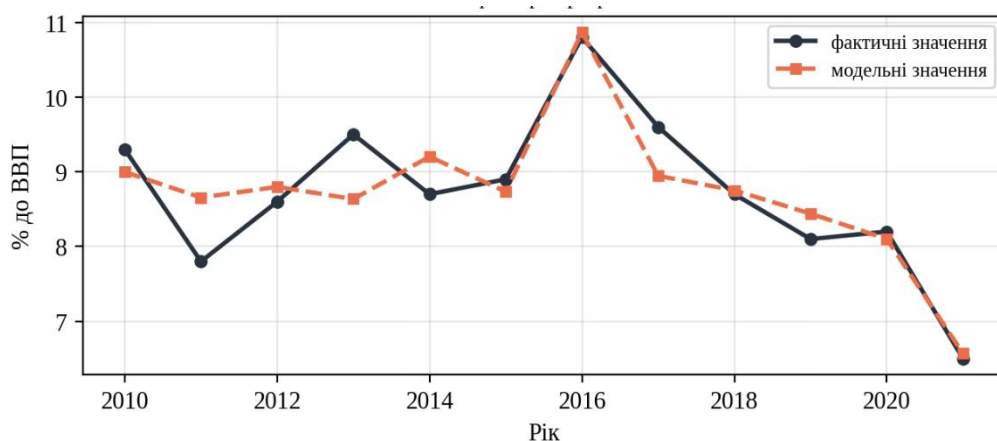


Рисунок 3.3 – Зіставлення фактичних і модельних значень частки витрат на пенсійне забезпечення (лінійна регресія)

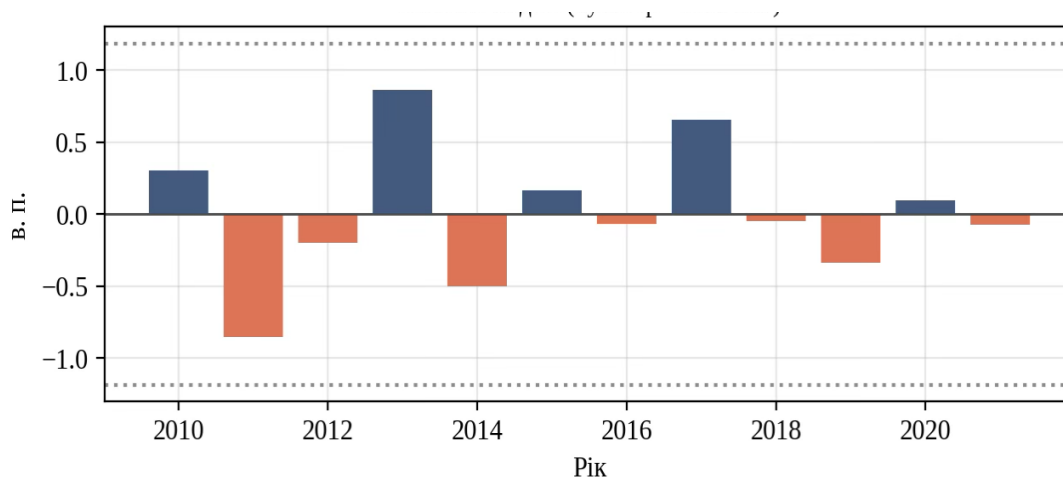


Рисунок 3.4 – Залишки моделі лінійної багатofакторної регресії

Водночас високе значення R^2 на вибірці з п'яти спостережень і трьох предикторів не може розглядатися як самостійне свідчення якості моделі: за такого співвідношення числа спостережень і параметрів існує суттєвий ризик перенавчання. Тому показник R^2 в роботі доповнено скоригованим R^2 , інформаційними критеріями та процедурою крос-валідації на ковзному вікні.

Нелінійна регресія дозволяє описати ефекти прискореного або уповільненого впливу факторів. Це відповідає економічній природі процесу: за помірної інфляції механізм індексації компенсує зростання цін поступово, тоді як за високої — витрати зростають випереджально. Квадратична специфікація має вигляд:

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t \quad (3.7)$$

де X — індекс споживчих цін.

Гранична (маржинальна) дія фактора у цій моделі не є сталою, а лінійно залежить від його рівня:

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X \quad (3.8)$$

Модель відтворює як загальний рівень показника, так і його локальний максимум 2016 р. (10,8 %) та різке зниження 2021 р. (6,5 %). Найбільші за модулем залишки припадають на 2013 р. (+0,864 в. п.) і 2011 р. (−0,856 в. п.), проте всі вони лежать у межах $\pm 2\sigma$, а їх знаки чергуються без вираженої серії.

Таблиця 3.4 – Результати оцінювання квадратичної регресії за фактором «ВВП на душу населення»

Параметр	Оцінка	p-значення	Гранична дія $\partial Y / \partial X_3$ за значення X_3
β_0	7,3235	< 0,001	$X_3 = 25$ тис. грн $\rightarrow +0,0404$
β_1	0,07003	0,0501	$X_3 = 60$ тис. грн $\rightarrow -0,0012$
β_2	−0,000593	0,0195	$X_3 = 100$ тис. грн $\rightarrow -0,0486$
			$X_3 = 130$ тис. грн $\rightarrow -0,0842$

Нелінійна специфікація (3.7) дозволяє описати ефекти прискореного або уповільненого впливу фактора. Оцінювання квадратичної моделі за фактором

«ВВП на душу населення» дало результати, наведені у табл. 3.4. Обидва коефіцієнти при факторі є статистично значущими, від’ємний знак β_2 вказує на спадну віддачу.

Гранична дія фактора, обчислена за (3.8), змінює знак у точці $X_3 = 59,0$ тис. грн. Доки номінальний ВВП на душу населення залишався нижчим за цей рівень (рис. 3.5, рис. 3.6), його зростання супроводжувалося підвищенням частки видатків на соціальний захист; після досягнення цього рівня, що припало на 2016–2017 рр., подальше зростання знаменника почало переважати, і частка видатків стала знижуватися.

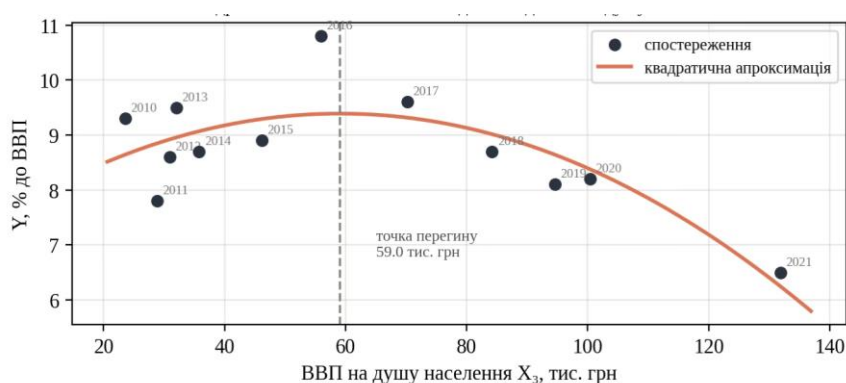


Рисунок 3.5 – Квадратична залежність частки видатків на соціальний захист від ВВП на душу населення

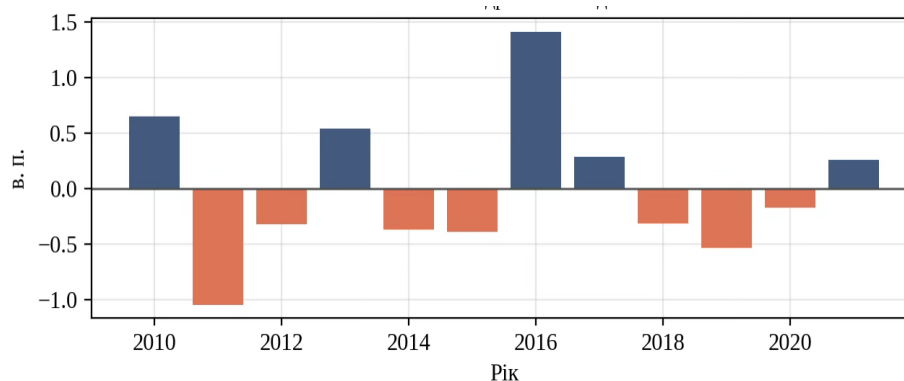


Рисунок 3.6 – Залишки моделі квадратичної регресії

Побудована парабола відтворює прискорений характер залежності: за низьких значень ІСЦ приріст частки витрат є помірним, тоді як у діапазоні понад 115 % крива помітно крутішає, причому найбільше відхилення спостерігається в точці $X_2 = 110$ %, де фактичне значення перевищує модельне. Залишки чергують знак без вираженого тренду, що свідчить про відсутність систематичного

зміщення, однак за п'яти спостережень формальна перевірка гомоскедастичності та автокореляції залишків є неможливою, тому висновок про адекватність специфікації має попередній характер.

Ансамблеві методи будують прогноз як композицію багатьох базових моделей, що дозволяє описувати складні нелінійні взаємодії факторів без явного задання функціональної форми. Випадковий ліс (Random Forest) [130] усереднює прогнози B дерев регресії, кожне з яких навчається на бутстреп-вибірці з випадковим підмножинним відбором ознак у вузлах:

$$\hat{Y}^{RF}(x) = \frac{1}{B} \sum_{b=1}^B h_b(x) \quad (3.9)$$

де $h_b(x)$ — прогноз b -го дерева; B — кількість дерев в ансамблі; x — вектор значень факторних ознак.

Відносна важливість j -ї ознаки визначається як нормоване сумарне зменшення критерію неоднорідності по всіх вузлах, у яких ця ознака використовується для розбиття:

$$I_j = \frac{\sum_b \sum_{v \in V_{jb}} \Delta Q(v)}{\sum_k \sum_b \sum_{v \in V_{kb}} \Delta Q(v)} \quad (3.10)$$

де V_{jb} — множина вузлів b -го дерева, у яких розбиття виконано за j -ю ознакою; $\Delta Q(v)$ — зменшення критерію неоднорідності (для задач регресії — суми квадратів відхилень) у вузлі v .

Градiєнтний бустинг будує адитивну композицію послідовно, причому кожна наступна базова модель наближає антиградієнт функції втрат:

$$F_M(x) = F_0(x) + \sum_{m=1}^M \nu \gamma_m h_m(x) \quad (3.11)$$

$$r_{im} = - \left[\frac{\partial L(Y_i, F(x_i))}{\partial F(x_i)} \right], \quad F = F_{m-1} \quad (3.12)$$

де $F_M(x)$ — підсумкова модель після M ітерацій; $F_0(x)$ — початкове наближення; $h_m(x)$ — базова модель m -ї ітерації, навчена на псевдозалишках r_{im} ; γ_m — оптимальний крок; $\nu \in (0; 1]$ — коефіцієнт швидкості навчання, що обмежує внесок кожної ітерації та стримує перенавчання.

Наприклад, для даних табл. 3.1 побудовано обидва ансамблі. Результати оцінювання важливості ознак і якості наведено у табл. 3.5,

Таблиця 3.5 – Порівняння ансамблевих моделей за важливістю ознак і точністю

Ознака	Назва показника	Випадковий ліс	Гرادієнтний бустинг	Коментар
X_1	чисельність пенсіонерів	0,254	0,346	демографічний чинник
X_2	індекс споживчих цін	0,123	0,031	поточний інфляційний чинник
$X_2(-1)$	ІСЦ попереднього року	0,401	0,450	відкладений інфляційний чинник
X_3	ВВП на душу населення	0,222	0,173	опосередкований вплив через знаменник
MSE на навчальній вибірці		0,1217	0,0000	нижче – не завжди краще
R^2 на навчальній вибірці		0,883	1,000	ознака перенавчання бустингу

Два ансамблі – випадковий ліс (500 дерев) і градієнтний бустинг [131, 132] (300 ітерацій, коефіцієнт швидкості навчання 0,05, максимальна глибина дерева 2) побудовано за даними табл. 3.1. У моделях узгоджено визначають ІСЦ попереднього року як найважливішу ознаку (0,401 і 0,450), що незалежно підтверджує результат регресійного аналізу щодо відкладеного характеру впливу інфляційного чинника. Розподіл важливості ознак у моделі випадкового лісу є відносно рівномірним: жодна з ознак не домінує беззастережно, а різниця між першою та третьою позиціями не перевищує 0,16 (рис. 3.7).

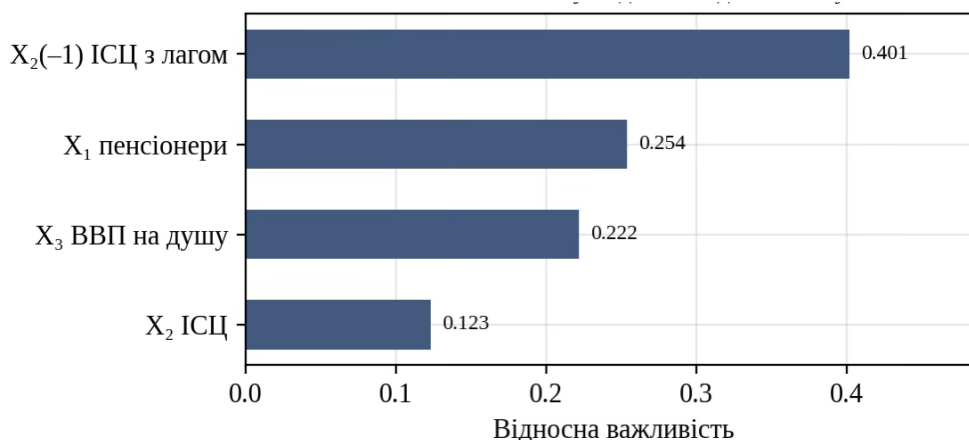


Рисунок 3.7 – Відносна важливість ознак у моделі випадкового лісу

Градiєнтний бустинг дає помітно контрастніший розподіл і змінює порядок ознак — на перше місце виходить індекс споживчих цін, тоді як внесок ВВП на душу населення стає майже нульовим (рис. 3.8).

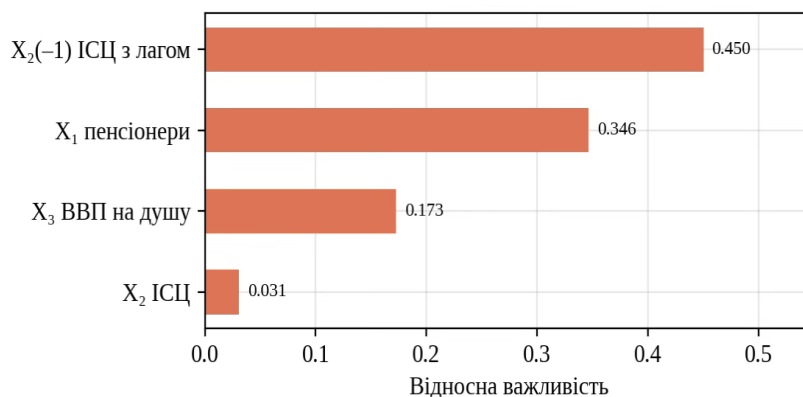


Рисунок 3.8 – Відносна важливість ознак у моделі градiєнтного бустингу

Обидві моделі узгоджено визначають чисельність пенсіонерів та індекс споживчих цін як домінантні фактори, що відповідає результатам регресійного аналізу (табл. 3.2). Водночас, між моделями є суттєва розбіжність в оцінці ролі ВВП на душу населення: випадковий ліс приписує цій ознаці 24 % важливості, градiєнтний бустинг — лише 4 %. Причиною є відомий недолік показника важливості, обчисленого через зменшення неоднорідності: він зміщений на користь ознак з більшою кількістю унікальних значень. Тому, такі оцінки слід розглядати як допоміжні й перевіряти за допомогою перестановочної важливості на відкладеній вибірці.

Значно нижче значення MSE градiєнтного бустингу (0,1450 проти 0,8781) не є свідченням його переваги, оскільки обидва значення обчислено на навчальній вибірці. За обсягу вибірки $n = 5$ бустинг здатний практично точно відтворити навчальні дані, тобто наведена різниця характеризує радше ступінь перенавчання, ніж прогностичну здатність. Коректне порівняння моделей потребує оцінювання на незалежних даних за процедурою оцінювання прогнозів.

3.3 Методика застосування байєсівських мереж, нечіткої логіки та моделей прогнозування контингенту одержувачів

Байєсівська мережа є ймовірнісною графовою моделлю [133], що подається орієнтованим ациклічним графом, вершини якого відповідають випадковим величинам, а дуги — відношенням безпосередньої ймовірнісної залежності. Основна властивість такої моделі — факторизація спільного розподілу за локальними умовними розподілами:

$$P(V_1, \dots, V_n) = \prod_{i=1}^n P(V_i | Pa(V_i)) \quad (3.13)$$

де V_i — i -та випадкова величина мережі; $Pa(V_i)$ — множина її батьківських вершин. Ця факторизація радикально скорочує кількість параметрів, які необхідно оцінити: замість повної таблиці спільного розподілу достатньо задати умовні розподіли для кожної вершини за її батьками.

Оновлення оцінок за появи нових спостережень виконується за формулою Байєса [133]

$$P(H | E) = \frac{P(E | H) P(H)}{\sum_{H'} P(E | H') P(H')} \quad (3.14)$$

де H — гіпотеза (наприклад, стан «висока потреба у фінансуванні»); E — спостережуване свідчення (фактичні значення демографічних та економічних індикаторів); $P(H)$ — апіорна ймовірність; $P(H|E)$ — апостеріорна ймовірність. Ймовірність цільової величини за наявних свідчень обчислюється маргіналізацією за прихованими змінними:

$$P(F | E) = \sum_{D,I,N,S} P(F, D, I, N, S | E) \quad (3.15)$$

Структуру мережі, побудованої для оцінювання потреби у фінансуванні, наведено на рис. 3.9.

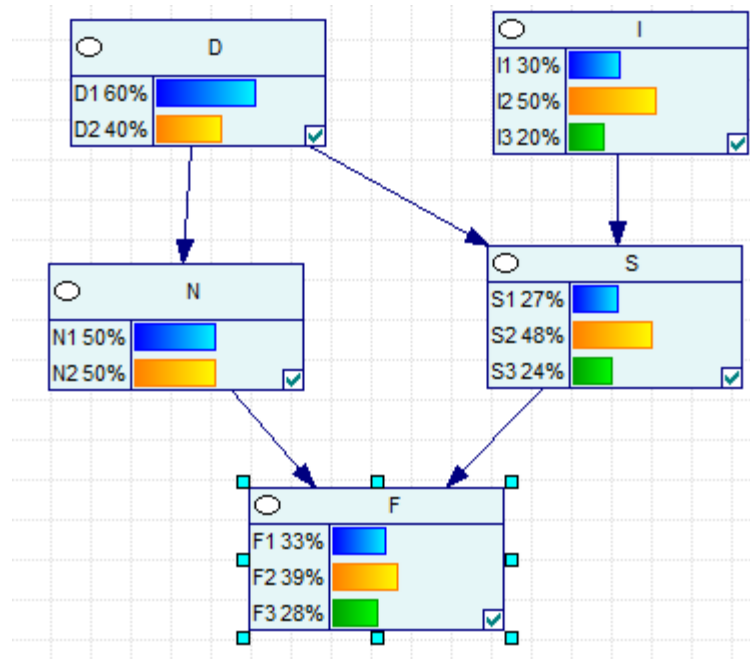


Рисунок 3.9 - Топологія мережі Байєса (побудована в програмі Genie 2 [133])

Таблиця 3.6 – Опис мережі Байєса

Вершина	Опис вершини	Стани верши та значення апріорних ймовірностей
D	Демографічний режим	$P(\text{Стабільний}) = 0,6$ $P(\text{Прискорене старіння}) = 0,4$
I	Інфляційний режим	$P(\text{Низький}) = 0,3$ $P(\text{Помірний}) = 0,5$ $P(\text{Високий}) = 0,2$
N	Чисельність одержувачів	$P(\text{Низька}) = 0,5$ $P(\text{Висока}) = 0,5$
S	Середній розмір виплати	$P(\text{Низький}) = 0,275$ $P(\text{Середній}) = 0,483$ $P(\text{Високий}) = 0,242$
F	Потреба у фінансуванні	$P(\text{Низька}) = 0,325$ $P(\text{Середня}) = 0,399$ $P(\text{Висока}) = 0,276$

Спільний розподіл мережі обчислюється за формулою:

$$P(D, I, N, S, F) = P(D) \cdot P(I) \cdot P(N|D) \cdot P(S|D, I) \cdot P(F|N, S)$$

В свою чергу формула для обчислення значень ймовірностей вершини F має вигляд:

$$P(F = F_i) = \sum_n \sum_s P(F = F_i | n, s) \cdot P(n) \cdot P(s)$$

Мережа містить дві екзогенні вершини — демографічний режим D та інфляційний режим I; дві проміжні — чисельність одержувачів N і середній розмір виплати S; цільову вершину — потреба у фінансуванні F (табл. 3.7-3.8).

Таблиця 3.7 - Значення апріорних ймовірностей вершини D (демографічний режим)

D	Опис стану	P(D)
D1	Стабільний	0,6
D2	Прискорене старіння	0,4

Таблиця 3.8 - Значення апріорних ймовірностей вершини S (середній розмір виплати)

S	Стан	P(I)
S1	Низький	0,3
S2	Помірний	0,5
S3	Високий	0,2

Таблиця 3.9 – Таблиця умовних ймовірностей вершини N (чисельність одержувачів)

D	N = N1 низька	N = N2 висока
D1=Стабільний	0,7	0,3
D2=Прискорене старіння	0,2	0,8

Таблиця 3.10 – Таблиця умовних ймовірностей вершини I (інфляційний режим)

Демографічний режим D	Інфляція I	S = S1 Низький	S = S2 Середній	S = S3 Високий
D1 = Стабільний	I1=Низький	0,5	0,4	0,1
D1 = Стабільний	I2=Помірний	0,25	0,55	0,2
D1 = Стабільний	I3=Високий	0,1	0,45	0,45
D2 = Прискорене старіння	I1=Низький	0,45	0,45	0,1
D2 = Прискорене старіння	I2=Помірний	0,2	0,55	0,25
D2 = Прискорене старіння	I3=Високий	0,05	0,35	0,6

Таблиця 3.11 - Таблиця умовних ймовірностей вершини F (потреба у фінансуванні)

N	S	F=F1 Низька	F=F2 Середня	F=F3 Висока
N1 = Низька	S1 = Низький	0,85	0,14	0,01
N1 = Низька	S2 = Середній	0,45	0,45	0,1
N1 = Низька	S3 = Високий	0,15	0,55	0,3
N2 = Висока	S1 = Низький	0,4	0,45	0,15
N2 = Висока	S2 = Середній	0,1	0,5	0,4
N2 = Висока	S3 = Високий	0,02	0,18	0,8

Наприклад, демографічна ситуація набуває станів «стабільний» і «прискорене старіння» з апіорними ймовірностями 0,6 і 0,4 відповідно. Умовні ймовірності, що чисельність одержувачів пенсій буде значною, задано таблицею (табл 3.7): $P(N = \text{висока} \mid D = \text{стабільний}) = 0,3$; $P(N = \text{висока} \mid D = \text{старіння}) = 0,8$. Тоді, безумовна ймовірність значної чисельності одержувачів пенсій становить $P(N = \text{висока}) = 0,6 \cdot 0,3 + 0,4 \cdot 0,8 = 0,50$. Якщо за результатами моніторингу зафіксовано значну чисельність одержувачів, апостеріорна ймовірність режиму прискореного старіння за формулою (3.14) дорівнює $P(D = \text{старіння} \mid N = \text{висока}) = (0,8 \cdot 0,4) / 0,50 = 0,64$, тобто зростає з 0,40 до 0,64. Це дозволяє переглянути сценарні ваги у розрахунках бюджету ПФУ без повного перебудування моделі.

Практична цінність байєсівського підходу для задач прийняття рішень у системі пенсійного забезпечення полягає у тому, що, його застосування дає розподіл прогнозованої величини, а не лише точкову оцінку; дозволяє коректно поєднувати статистичні дані з експертними судженнями через апіорні розподіли; [41, 95. 133] допускає послідовне оновлення оцінок по мірі надходження нової звітності. У ситуаціях, коли межі між станами факторів не піддаються чіткому кількісному визначенню, а експертні оцінки формулюються лінгвістично («низька», «середня», «висока» потреба у фінансуванні), доцільним є . Кожному терму лінгвістичної змінної відповідає функція приналежності; найпоширенішою є трикутна [135, 136]:

$$\mu(x; a, b, c) = \max\left(0, \min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right)\right) \quad (3.16)$$

де a та c — межі носія нечіткої множини; b — точка, у якій приналежність досягає одиниці. Функції приналежності лінгвістичних змінних «чисельність одержувачів пенсій» і «індекс споживчих цін» наведено на рис. 3.10 та рис. 3.11.

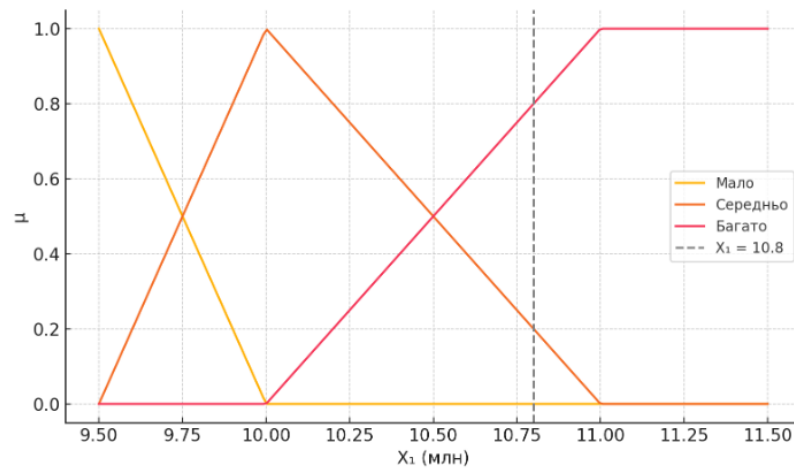


Рисунок 3.10 – Функції приналежності лінгвістичної змінної «чисельність одержувачів пенсій»

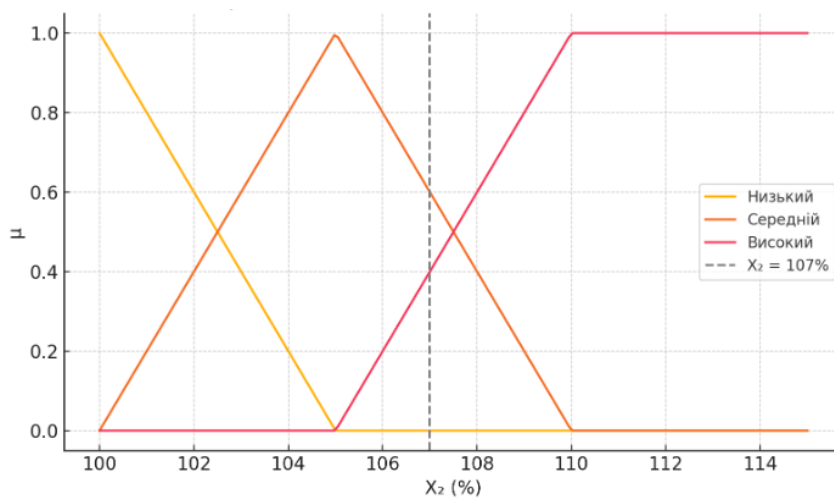


Рисунок 3.11 – Функції приналежності лінгвістичної змінної «індекс споживчих цін»

Активність k -го правила бази знань за кон'юнктивного зв'язку передумов обчислюється за операцією мінімуму [136, 137]:

$$\alpha_k = \min\left(\mu_{A_k}(x_1), \mu_{B_k}(x_2)\right) \quad (3.17)$$

де $\mu_{A_k}(x_1)$ та $\mu_{B_k}(x_2)$ — ступені приналежності поточних значень факторів до термів, що фігурують у передумові k -го правила. Правила з однаковим висновком агрегуються за операцією максимуму:

$$\mu_j = \max_{k \in K_j} \alpha_k \quad (3.18)$$

де K_j - множина правил, висновком яких є j -й терм цільової змінної. Дефазифікація за методом центру ваги для синглетонних висновків дає остаточну кількісну оцінку:

$$y^* = \frac{\sum_j \mu_j c_j}{\sum_j \mu_j} \quad (3.19)$$

де c_j — числове значення, що відповідає j -му терму висновку. Важливим є такий методологічно суттєвий момент. Якщо активним є одне правило, то згідно з (3.19) $y^* = c_j$ незалежно від значення активності μ_j . Тому база правил, побудована так, що для типових значень факторів спрацьовує лише одне правило, є виродженою — вона не використовує механізм нечіткого виведення й фактично зводиться до звичайної таблиці відповідностей. Коректна база правил має забезпечувати одночасну активацію щонайменше двох правил із різними висновками.

Наприклад, нехай $X_1 = 10,8$ млн осіб і $X_2 = 107$ %. За функціями приналежності (рис. 3.10, 3.11) отримуємо: $\mu(X_1 = \text{«середньо»}) = 0,2$; $\mu(X_1 = \text{«багато»}) = 0,8$; $\mu(X_2 = \text{«середній»}) = 0,6$; $\mu(X_2 = \text{«високий»}) = 0,4$. Базу правил та результати обчислень наведено у табл. 3.12.

Таблиця 3.12 - База нечітких правил і результати виведення

№	Передумова за X_1	Передумова за X_2	Висновок	Активність α_k
1	середньо (0,2)	середній (0,6)	потреба середня	$\min(0,2; 0,6) = 0,2$
2	багато (0,8)	середній (0,6)	потреба висока	$\min(0,8; 0,6) = 0,6$
3	багато (0,8)	високий (0,4)	потреба висока	$\min(0,8; 0,4) = 0,4$
4	середньо (0,2)	високий (0,4)	потреба середня	$\min(0,2; 0,4) = 0,2$
Агрегація за (3.18)	$\mu(\text{середня}) =$ $\max(0,2; 0,2) = 0,2$	$\mu(\text{висока}) =$ $\max(0,6; 0,4) = 0,6$	$c(\text{середня}) = 50$; $c(\text{висока}) = 80$	

Підстановка агрегованих ступенів приналежності у формулу (3.19) дає:

$$y^* = \frac{0,2 \cdot 50 + 0,6 \cdot 80}{0,2 + 0,6} = \frac{58}{0,8} = 72,5 \quad (3.20)$$

Отже, прогнозована потреба у фінансуванні пенсійних витрат становить 72,5 умовних одиниць, що відповідає рівню між «середньою» та «високою» з переважанням останньої. Результат дефазифікації наведено на рис. 3.12. На відміну від виродженого випадку з одним активним правилом, отримане значення є чутливим до зміни вхідних факторів: зменшення X_1 до 10,5 млн осіб змінює співвідношення ступенів приналежності й знижує оцінку до 65,0 ум. од.

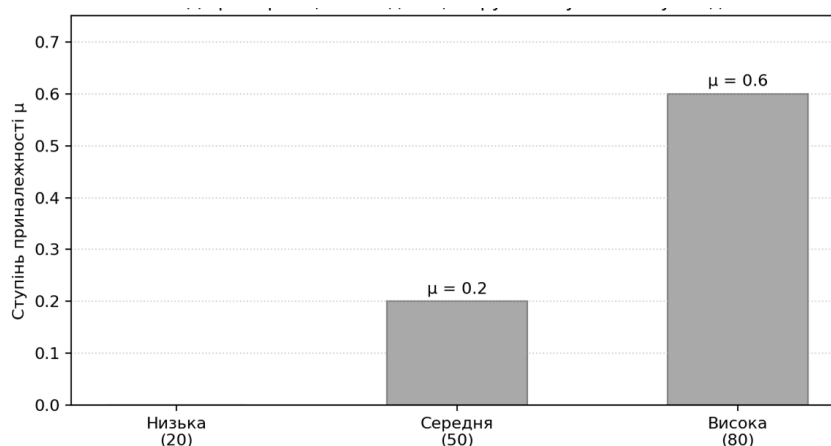


Рисунок 3.12 – Дефазифікація методом центру ваги за агрегованою базою правил

На рис. 3.12 видно, що центр ваги зміщений до терму «висока», проте не збігається з його синглетоном 80: внесок терму «середня» з ненульовим ступенем приналежності 0,2 утримує підсумкову оцінку на рівні 72,5 ум. од. Саме така поведінка й відрізняє повноцінне нечітке виведення від таблиці відповідностей — вихідна величина неперервно змінюється разом із вхідними факторами, а не стрибками переходить між дискретними значеннями 20, 50 і 80.

Для прогнозування кількості нових призначень пенсій доцільно застосовувати й моделі авторегресії – інтегрованого ковзного середнього.

Позначимо через B оператор зсуву назад ($BY_t = Y_{t-1}$), тоді оператор різниці порядку d записується як $(1 - B)^d$. Модель $ARIMA(p, d, q)$ має вид [78, 114]:

$$\varphi(B)(1 - B)^d Y_t = c + \theta(B)\varepsilon_t \quad (3.21)$$

де $\varphi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p$ — авторегресійний поліном; $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ — поліном ковзного середнього; d — порядок інтегрування (кількість різницювань, потрібних для досягнення стаціонарності); c — константа; ε_t — білий шум. За наявності сезонності періоду s модель розширюється до $SARIMA(p, d, q)(P, D, Q)s$:

$$\varphi(B)\Phi(B^s)(1 - B)^d(1 - B^s)^D Y_t = \theta(B)\Theta(B^s)\varepsilon_t \quad (3.22)$$

де $\Phi(B^s)$ і $\Theta(B^s)$ — сезонні авторегресійний поліном і поліном ковзного середнього; D — порядок сезонного інтегрування; s — довжина сезонного циклу (для квартальних даних $s = 4$, для місячних $s = 12$). Порядки моделі обираються мінімізацією інформаційного критерію Акаїке [114, 128].

Розглянемо помісячну кількість нових призначень пенсій за 12 місяців (тис осіб): 25, 28, 30, 35, 27, 31, 34, 40, 32, 35, 39, 48 (рис. 3.13).

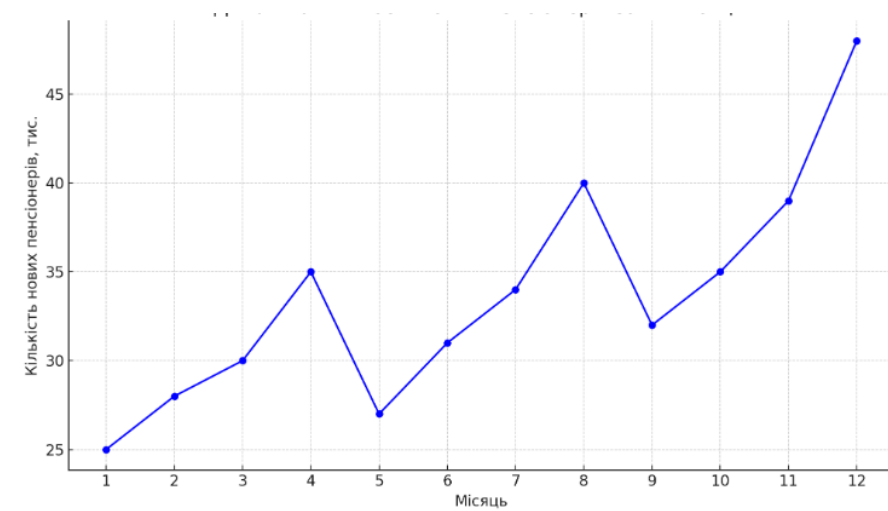


Рисунок 3.13 — Динаміка кількості нових призначень пенсій за 12 місяців

Ряд є нестационарним із вираженим зростаючим трендом і сезонним компонентом періоду 4, що зумовлює вибір $d = 1$. Прогноз за моделлю $ARIMA$

на чотири наступні періоди становить 47, 45, 49 і 55 тис осіб (рис. 3.14); модель коректно відтворює як трендову складову, так і локальні спади.

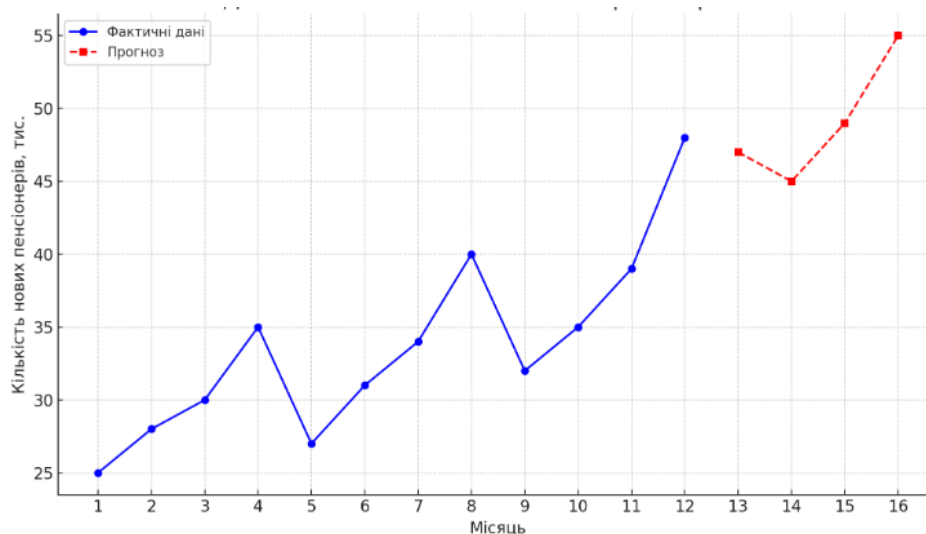


Рисунок 3.14 – Прогноз кількості нових призначень пенсій за моделлю ARIMA на 4 періоди

Важливим обмеженням є те, що за обсягу вибірки у 12 спостережень оцінки параметрів сезонної складової мають низьку надійність, а довірчі інтервали прогнозу є широкими. Для практичного застосування необхідні ряди довжиною щонайменше 4–5 повних сезонних циклів.

Сценарний підхід описує чисельність контингенту одержувачів як функцію заданого темпу приросту, який відображає гіпотезу про розвиток демографічної ситуації [32, 138,]:

$$P_t^{(s)} = P_0(1 + g^{(s)})^t \quad (3.23)$$

де P_0 — чисельність пенсіонерів у базовому році; $g(s)$ — середньорічний темп приросту за сценарієм s ; t — номер року прогнозного горизонту. Балансове рівняння, що деталізує механізм зміни контингенту, має вигляд:

$$P_{t+1} = P_t + N_t - D_t + M_t \quad (3.24)$$

де N_t — кількість нових призначень пенсій; D_t — вибуття внаслідок смертності; M_t — сальдо міграції одержувачів.

За базової чисельності $P_0 = 10,00$ млн осіб розглянуто три сценарії: базовий ($g = 0,5 \%$ на рік), сценарій прискореного старіння населення ($g = 0,8 \%$) і сценарій міграційного впливу ($g = 0,3 \%$). Результати розрахунку за формулою (3.23) наведено у табл. 3.13 та на рис. 3.15.

Таблиця 3.13 - Результати сценарного прогнозування чисельності пенсіонерів, млн осіб

Сценарій	Роки				
	1	2	3	4	5
Базовий ($g = 0,5 \%$)	10,05	10,10	10,15	10,20	10,25
Прискорене старіння ($g = 0,8 \%$)	10,08	10,16	10,24	10,32	10,40
Міграційний вплив ($g = 0,3 \%$)	10,03	10,06	10,09	10,12	10,15

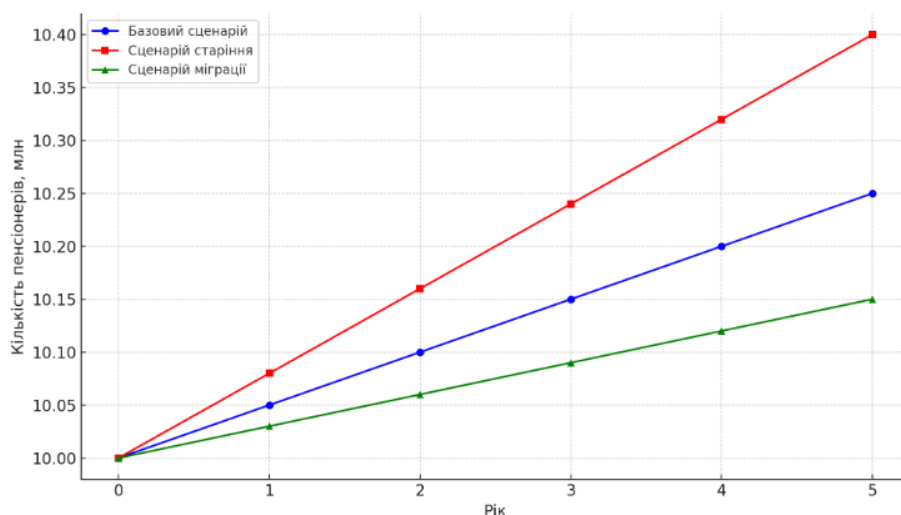


Рисунок 3.15 - Сценарне прогнозування чисельності пенсіонерів на п'ятирічному горизонті

Розкид між крайніми сценаріями на п'ятому році становить 0,25 млн осіб, або 2,4 % базової чисельності. За середнього річного розміру пенсійної виплати

це відповідає суттєвій різниці у потребі фінансування, що обґрунтовує доцільність формування бюджетного резерву на рівні верхнього сценарію.

Для оцінювання ймовірності виходу окремої особи на пенсію протягом прогнозного періоду застосовується логістична регресія [34, 114, 140-145]:

$$P(Y = 1 | x) = \frac{1}{1 + \exp(-\beta_0 - \sum_{j=1}^p \beta_j x_j)} \quad (3.25)$$

де $Y = 1$ відповідає настанню події (призначення пенсії) у прогнозному періоді; x_j — характеристики особи (вік, страховий стаж, наявність пільгових підстав). Оцінювання параметрів виконується максимізацією логарифмічної правдоподібності, еквівалентною мінімізації логістичної функції втрат:

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^n [y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i)] \quad (3.26)$$

Для однофакторної моделі із предиктором «вік» отримано оцінку $\beta_1 \approx 1,48$, що відповідає відношенню шансів $e^{1,48} \approx 4,4$: кожен додатковий рік віку в передпенсійному діапазоні збільшує шанси виходу на пенсію приблизно у 4,4 раза. Прогнозовані ймовірності зростають від 0,02 у віці 55 років до 0,99 у віці 61 року (рис. 3.16), тобто модель відтворює різкий перехід у вузькому віковому інтервалі.

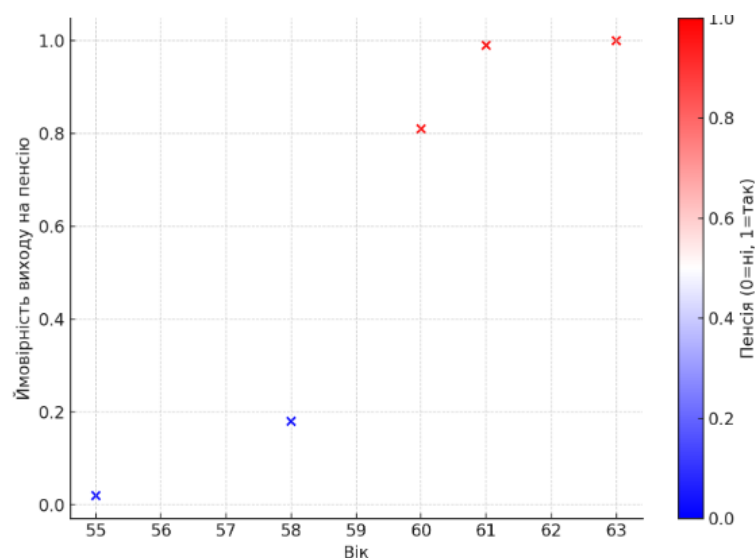


Рисунок 3.16 - Прогнозовані ймовірності виходу на пенсію залежно від віку (логістична модель)

Практично весь перехід від нульової до одиничної ймовірності вкладається в інтервал 58–61 рік, тоді як за його межами модель дає майже детерміновані оцінки, не чутливі до подальшої зміни віку. Це означає, що прогностична цінність моделі зосереджена саме у вузькій передпенсійній групі, і для оцінювання очікуваної чисельності нових призначень достатньо застосовувати її до контингенту цього вікового діапазону. Водночас оцінки, отримані на вибірці з п'яти спостережень, мають ілюстративний характер: широкі довірчі інтервали коефіцієнта не дають змоги надійно локалізувати точку перегину логістичної кривої.

3.4 Методи комбінування прогнозів і багатомодельні підходи

Комбінування прогнозів кількох моделей дозволяє зменшити похибку порівняно з найкращою окремою моделлю за рахунок часткової взаємної компенсації різноспрямованих помилок. Лінійна комбінація прогнозів має вигляд: [145 -150]

$$\hat{Y}_t^{comb} = \sum_{i=1}^M w_i \hat{Y}_t^{(i)}, \quad \sum_{i=1}^M w_i = 1, \quad w_i \geq 0 \quad (3.27)$$

де $\hat{Y}_t(i)$ — прогноз i -ї базової моделі у момент t ; w_i — вага i -ї моделі; M — кількість моделей. Обмеження невід'ємності ваг і їх нормування забезпечують незміщеність комбінованого прогнозу за незміщених базових. Оптимальні ваги визначаються з задачі мінімізації середньоквадратичної помилки:

$$w^* = \arg \min \frac{1}{n} \sum_{t=1}^n \left(Y_t - \sum_{i=1}^M w_i \hat{Y}_t^{(i)} \right)^2 \quad (3.28)$$

Для випадку двох моделей задача (3.28) має аналітичний розв'язок (оптимальні ваги Бейтса — Гренджера) [92]:

$$w_1^* = \frac{\sigma_2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}, \quad w_2^* = 1 - w_1^* \quad (3.29)$$

де σ^2_1 та σ^2_2 — дисперсії помилок першої та другої моделей; σ_{12} — коваріація їх помилок. З (3.29) випливає ключовий висновок: чим слабше корельовані помилки моделей, тим більший вииграш дає комбінування; за $\sigma_{12} \rightarrow \sigma^2_1 = \sigma^2_2$ комбінування втрачає сенс.

Нехай прогнози двох моделей кількості нових пенсіонерів на чотири квартали становлять: модель А — 45, 47, 48, 50 тис. осіб; модель В — 43, 46, 49, 52 тис. осіб. Фактичні значення: 44, 48, 50, 51 тис. осіб. Помилки моделей: $e_A = (1; -1; -2; -1)$, $e_B = (-1; -2; -1; 1)$. Звідси $\sigma^2_1 = 1,75$; $\sigma^2_2 = 1,75$; $\sigma_{12} = 0,50$. За формулою (3.29) $w_1^* = (1,75 - 0,50)/(1,75 + 1,75 - 2 \cdot 0,50) = 1,25/2,50 = 0,50$, тобто оптимальним є рівновагове комбінування. Комбінований прогноз становить 44,0; 46,5; 48,5; 51,0 тис. осіб, а його MSE дорівнює 1,125 проти 1,750 для кожної з базових моделей, тобто похибка зменшилася на 35,7 % (рис. 3.17).

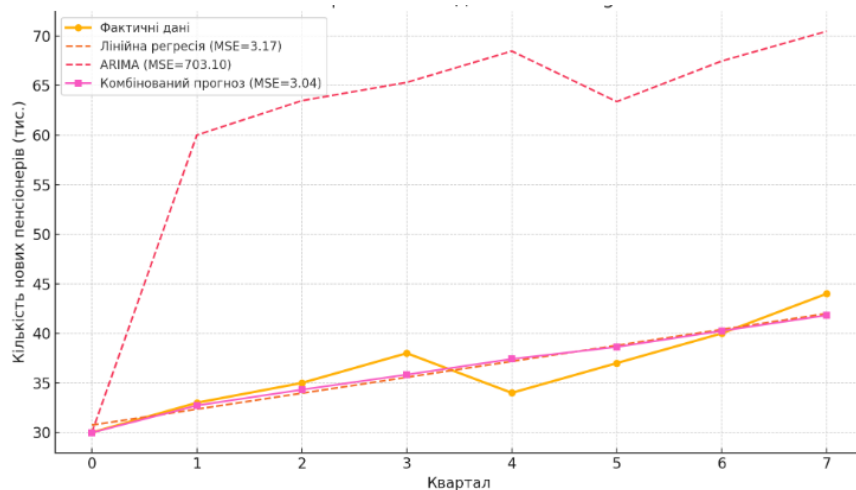


Рисунок 3.17 – Порівняння базових моделей і комбінованого прогнозу (stacking)

Складнішою формою комбінування є стекінг, у якому прогнози базових моделей виступають вхідними ознаками мета-моделі:

$$\hat{Y}_t^{stack} = g(\hat{Y}_t^{(1)}, \dots, \hat{Y}_t^{(M)}; \lambda) \quad (3.30)$$

де $g(\cdot)$ - мета-модель (найчастіше лінійна регресія або неглибоке дерево рішень); λ — вектор її параметрів. Принципова вимога коректного стекінгу — навчання

мета-моделі на прогнозах, отриманих поза навчальною вибіркою базових моделей (за схемою крос-валідації), інакше мета-модель систематично переоцінює найбільш перенавчену базову модель [94, 152].

Для щоквартального ряду кількості новопризначених пенсій (30, 33, 35, 38, 34, 37, 40, 44 тис. осіб) як базові використано лінійну регресію на тренді та модель ARIMA(1,1,0), як мета-модель — лінійну регресію. Результати наведено у табл. 3.14 і на рис. 3.18.

Таблиця 3.14 – Порівняння точності базових моделей і комбінованого прогнозу

Модель	MSE	Характеристика
Лінійна регресія на тренді	3,17	відтворює трендову складову, ігнорує коливання
ARIMA(1,1,0)	703,10	нестійка за короткого ряду, накопичує похибку
Комбінований прогноз (стекінг)	3,04	найменша похибка серед розглянутих

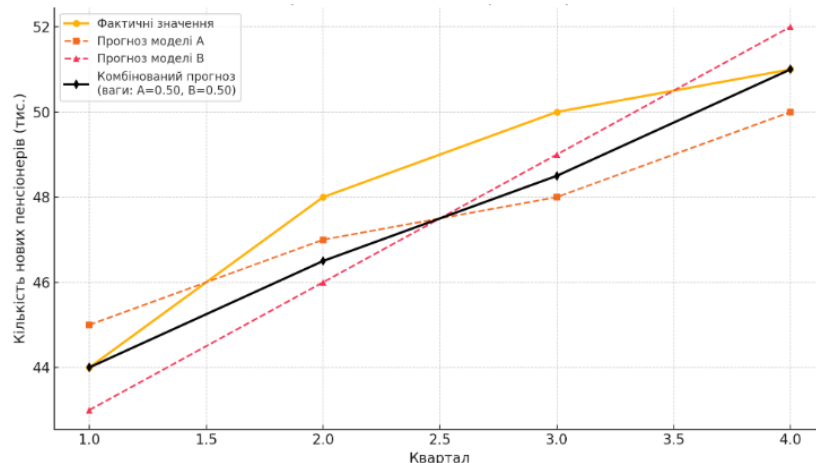


Рисунок 3.18 – Вагове комбінування прогнозів двох моделей з оптимальними вагами

Наведений приклад демонструє важливу властивість стекінгу: мета-модель здатна практично повністю знівелювати внесок неякісної базової моделі.

Водночас він виявляє й обмеження підходу - вигаиш комбінованого прогнозу порівняно з найкращою базовою моделлю становить лише 4,1 % (3,04 проти 3,17), тобто за наявності однієї явно домінантної моделі ускладнення схеми комбінування є не виправданим.

Окремим класом є комбінування моделей різної природи, зокрема схема «ARIMA + регресія на залишках». Спочатку часова модель описує трендову й сезонну складові ряду, після чого на її залишках будується регресія за факторними змінними [94, 114]:

$$e_t = Y_t - \hat{Y}_t^{ARIMA}, \quad \hat{e}_t = \gamma_0 + \sum_{j=1}^p \gamma_j X_{jt} \quad (3.31)$$

$$\hat{Y}_t^{comb} = \hat{Y}_t^{ARIMA} + \hat{e}_t \quad (3.32)$$

де e_t — залишки часової моделі; γ_j — коефіцієнти регресії залишків на факторні ознаки.

Такий підхід дозволяє поєднати переваги обох класів моделей: часова компонента враховує інерційність процесу, а регресійна — вплив зовнішніх чинників, насамперед інфляційного (рис. 3.19).

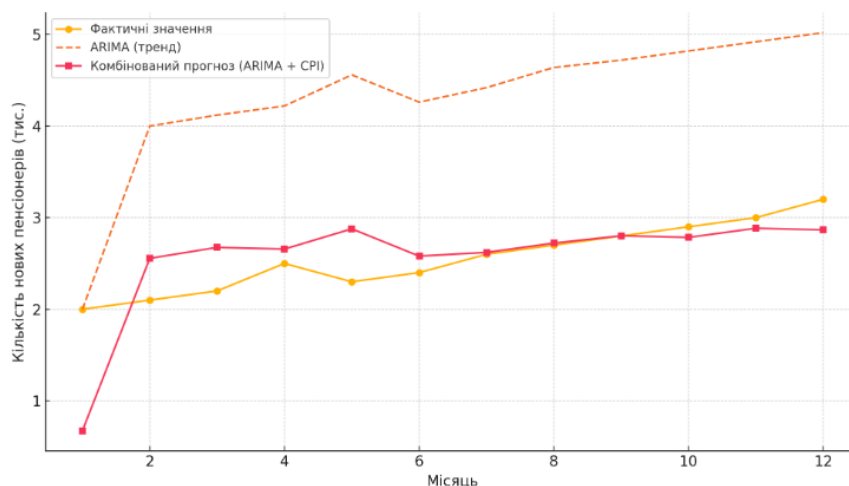


Рисунок 3.19 – Комбінування ARIMA та регресії на залишках

Для оцінювання комбінованих моделей застосовується критерій Дібольда-Маріано [112].

$d_t = \mathcal{L}(e_{1t}) - \mathcal{L}(e_{2t})$ — ряд різниці втрат двох моделей у момент t , e_{1t} та e_{2t} - помилки прогнозу першої та другої моделі у час t , $\mathcal{L}(\cdot)$ — функція втрат, тоді статистика критерію має вигляд:

$$DM = \frac{\bar{d}}{\sqrt{\frac{2\pi\hat{f}_d(0)}{n}}} \quad (3.33)$$

де $\bar{d}$ — вибіркове середнє значення ряду d_t за весь період прогнозування (n спостережень), показує, яка модель у середньому була точнішою. різниці втрат d_t ; $\hat{f}_d(0)$ — оцінка спектральної щільності ряду d_t у нульовій частоті; n - довжина вибірки (кількість прогнозів).

За нульової гіпотези про однакову прогностичну точність статистика DM асимптотично має стандартний нормальний розподіл. Застосування цього критерію є обов'язковим для обґрунтування висновку про перевагу комбінованої моделі, оскільки саме по собі зменшення MSE на кількох спостереженнях може бути наслідком випадкових коливань чи стійкою закономірністю [112, 113].

3.5 Критерії оцінювання якості прогнозів і точності моделей

Оцінювання якості моделей виконується за системою метрик, склад якої визначається типом моделі та характером прогнозованої величини. Це як «традиційні» метрики, так і запропоновані для даної роботи.

Середня абсолютна похибка характеризує відхилення прогнозу від факту в одиницях вимірювання цільової величини [94, 114, 128]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (3.34)$$

де n — кількість спостережень; Y_i — фактичні значення; $\hat{Y}_i$ — прогнозовані.

Корінь із середньоквадратичної похибки надає більшої ваги великим відхиленням, що доцільно у бюджетному плануванні, де значні одиничні помилки є критичнішими за низку дрібних [94, 114, 128]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (3.35)$$

де n — кількість спостережень; Y_i — фактичні значення; $\hat{Y}_i$ — прогнозовані.

Середня абсолютна відносна похибка у відсотках дозволяє зіставляти точність моделей для показників різного масштабу [94, 114, 128]:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \quad (3.36)$$

де n — кількість спостережень; Y_i — фактичні значення; $\hat{Y}_i$ — прогнозовані.

Оскільки MAPE не визначена за $Y_i = 0$ і несиметрична щодо знака помилки, для рядів із малими значеннями застосовується масштабована похибка, нормована на похибку наївного прогнозу [94, 114, 128]:

$$MASE = \frac{MAE}{\frac{1}{n-1} \sum_{i=2}^n |Y_i - Y_{i-1}|} \quad (3.37)$$

де n — кількість спостережень; Y_i — фактичні значення; $\hat{Y}_i$ — прогнозовані. Значення $MASE < 1$ означає, що модель точніша за наївний прогноз «останнє спостережене значення».

Коефіцієнт детермінації показує частку дисперсії цільової змінної, поясненої моделлю, а скоригований коефіцієнт додатково штрафує за надмірну кількість предикторів [94, 114, 128]:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (3.38)$$

де n — кількість спостережень; Y_i — фактичні значення; $\hat{Y}_i$ — прогнозовані.

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1} \quad (3.39)$$

де $\bar{Y}$ — середнє фактичних значень; p — кількість факторних змінних. Для порівняння моделей різної складності застосовуються інформаційні критерії [94, 114, 128]:

$$AIC = 2k - 2 \ln \hat{L}, \quad (3.40)$$

$$BIC = k \ln n - 2 \ln \hat{L} \quad (3.41)$$

де $\hat{L}$ — максимальне значення функції правдоподібності; k — кількість оцінюваних параметрів; n — обсяг вибірки.

Менше значення критерію відповідає кращому балансу точності та складності; BIC штрафує складність суворіше, тому надає перевагу економнішим моделям.

Для інтервальних прогнозів обчислюється частка фактичних значень, що потрапили у прогнозний інтервал [147]:

$$PICP = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(L_i \leq Y_i \leq U_i) \quad (3.42)$$

де L_i та U_i — нижня і верхня межі прогнозного інтервалу; $\mathbb{1}(\cdot)$ — індикаторна функція. Емпіричне покриття PICP має відповідати номінальному рівню довіри: істотне відхилення свідчить про некоректне оцінювання невизначеності.

Для моделей, що прогнозують настання події, використовується матриця невідповідностей із позначеннями: TP — істинно-позитивні, FP — хибно-позитивні, FN — хибно-негативні, TN — істинно-негативні рішення. На її основі обчислюються [66]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.43)$$

$$Precision = \frac{TP}{TP + FP} \quad (3.44)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.45)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.46)$$

Через залежність спостережень у часі стандартна k -кратна крос-валідація з випадковим розбиттям є неприйнятною: вона допускає використання майбутньої інформації для прогнозування минулого. Застосовується крос-валідація на ковзному вікні, за якої модель послідовно перенавчається на розширюваній навчальній вибірці, а оцінка усереднюється за всіма зрізами [94]:

$$CV = \frac{1}{K} \sum_{k=1}^K \mathcal{L}(Y_{n_k+h}, \hat{Y}_{n_k+h}^{(k)}) \quad (3.47)$$

де K — кількість зрізів; n_k — момент завершення навчальної вибірки на k -му зрізі; $\hat{Y}(k)$ — прогноз моделі, навченої на даних до моменту n_k включно.

Для фактичних значень частки витрат (12,5; 13,0; 14,0; 15,5 %) і прогнозних (12,7; 13,2; 13,8; 15,0 %) обчислено метрики точності.

Помилки становлять $-0,2$; $-0,2$; $+0,2$; $+0,5$ в. п.

Звідки $MAE = (0,2 + 0,2 + 0,2 + 0,5)/4 = 0,275$;

$$RMSE = \sqrt{(0,04 + 0,04 + 0,04 + 0,25)/4} = \sqrt{0,0925} = 0,304;$$

$MAPE = 1,95$ %.

Графічне зіставлення наведено на рис. 3.20.

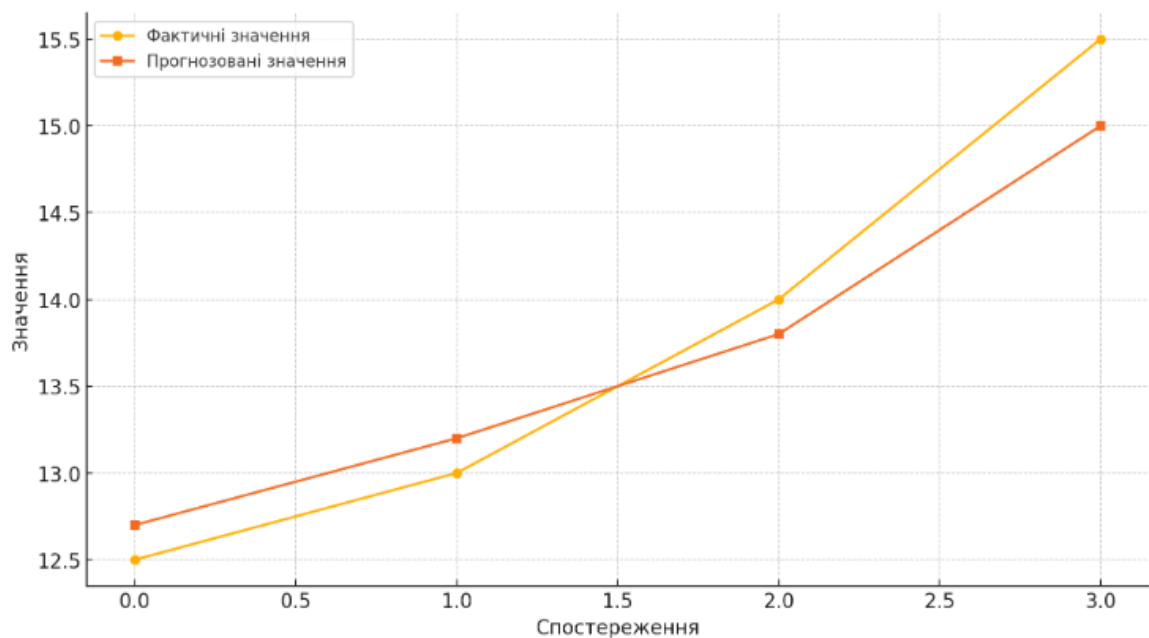


Рисунок 3.20 – Зіставлення прогнозних і фактичних значень з оцінками точності

Для регресійних моделей контингенту одержувачів пенсій ($R^2 = 0,98$; $n = 5$; $p = 3$) скоригований коефіцієнт детермінації за формулою (3.39) дорівнює $R^2_{adj} = 1 - (1 - 0,98) \cdot 4/1 = 0,92$, тобто помітно нижчий за нескоригований. Це кількісно підтверджує зауваження щодо ризику перенавчання за малого обсягу вибірки.

Для класифікаційної моделі оцінювання ймовірності виходу на пенсію на тестовій вибірці з 20 спостережень отримано $TP = 14$, $FN = 4$, $FP = 1$, $TN = 1$. За формулами (3.43)–(3.47): $Accuracy = 15/20 = 0,750$; $Precision = 14/15 = 0,933$; $Recall = 14/18 = 0,778$; $F1 = 2 \cdot 0,933 \cdot 0,778 / (0,933 + 0,778) = 0,849$. Зведення метрик наведено у табл. 3.15.

Таблиця 3.15 – Зведені результати оцінювання якості моделей

Тип моделі	Метрика	Значення	Інтерпретація
Кількісний прогноз	MAE	0,275 в. п.	типове відхилення від факту
Кількісний прогноз	RMSE	0,304 в. п.	з урахуванням великих похибок
Кількісний прогноз	MAPE	1,95 %	відносна похибка нижче 2 %
Регресійна	R^2 / R^2_{adj}	0,98 / 0,92	розрив вказує на перенавчання
Класифікаційна	Accuracy	0,750	частка правильних рішень
Класифікаційна	Precision	0,933	низька частка хибних спрацювань
Класифікаційна	Recall	0,778	пропущено 22 % подій
Класифікаційна	F1	0,849	збалансована оцінка

Для задач планування бюджету співвідношення між Precision і Recall має асиметричну управлінську вагу: пропуск події (низький Recall) призводить до недооцінки потреби у фінансуванні, тоді як хибне спрацювання спричиняє лише надлишкове резервування. Тому, за налаштування порогу класифікації доцільно зміщувати його у бік підвищення повноти навіть ціною зниження прецизійності.

В роботі було запропоновано власну систему критеріїв оцінювання прогнозів шляхом поєднання метрик точності кількісних прогнозів, показники статистичної адекватності регресійних та ймовірнісних моделей, критерії стійкості оцінок за неповних і спотворених даних, критерії верифікації прогнозів за фактичними даними і сценаріями, що забезпечує побудову адекватних моделей досліджуваних процесів та отримання високоякісних оцінок прогнозів.

Запропонований профіль якості прогнозової моделі – це за своєю суттю вектор характеристик (профіль) моделі M , що описується в загальному вигляді як

$$Profile(M) = \{A, Q, P, R, I, S\} \quad (3.48)$$

або

$$Profile(M) = \begin{pmatrix} A(M) \\ Q(M) \\ P(M) \\ R(M) \\ I(M) \\ S(M) \end{pmatrix} \in [0, 1]^6 \quad (3.49)$$

Де A – точність прогнозу (accuracy); Q – статистична адекватність (statistical adequacy); P – ймовірнісна адекватність (probabilistic adequacy); R – стійкість до неповних та спотворених даних (robustness); I – якість прогнозних інтервалів (interval quality); S – сценарна адекватність (scenario validity).

Кожна компонента профілю в свою чергу сама є агрегованою характеристикою групи критеріїв.

$$A(M) = f_A(MAE, RMSE, MAPE, sMAPE, R^2) \quad (3.50)$$

$$Q(M) = f_Q(F, p, AIC, BIC, DW, BP, LjungBox) \quad (3.51)$$

$$P(M) = f_P(Brier, LogLoss, Calibration) \quad (3.52)$$

$$R(M) = f_R(RI_{missing}, RI_{noise}, RI_{outlier}) \quad (3.53)$$

$$I(M) = f_I(PICP, MPIW) \quad (3.54)$$

$$S(M) = f_S(S_1, \dots, S_k) \quad (3.55)$$

Для того щоб отримати оцінку точності прогнозу $A(M)$ необхідно мати показники оцінювання адекватності моделі: MAE – середня абсолютна похибка прогнозування, що характеризує середню величину абсолютного відхилення прогнозованих значень від фактичних; RMSE – корінь із середньоквадратичної похибки, що характеризує величину прогнозних помилок із підвищеною чутливістю до великих відхилень; MAPE – середня абсолютна відносна похибка

у відсотковому вираженні, що характеризує середню відносну точність прогнозу; sMAPE – симетрична середня абсолютна відносна похибка, яка характеризує відносне відхилення прогнозних значень від фактичних із симетризацією відносно прогнозу та фактичного значення; R^2 – коефіцієнт детермінації, що характеризує частку варіації залежної змінної, пояснену моделлю.

Для того щоб отримати статистичну адекватність $Q(M)$ необхідно розрахувати показники моделі: F – F -критерій Фішера, що характеризує статистичну значущість регресійної моделі в цілому та дозволяє оцінити, чи пояснює модель статистично значущу частину варіації залежної змінної; p – рівень статистичної значущості (p -value), який характеризує статистичну значущість оцінених параметрів моделі або відповідного статистичного тесту; менше значення p свідчить про сильніші підстави для відхилення нульової гіпотези; AIC – інформаційний критерій Акаїке, що використовується для порівняння моделей з урахуванням компромісу між якістю їх апроксимації та складністю моделі; менше значення AIC відповідає кращій моделі серед порівнюваних; BIC – байєсівський інформаційний критерій, який оцінює якість моделі з урахуванням її складності та обсягу вибірки; менше значення BIC свідчить про перевагу моделі серед альтернативних моделей; DW – критерій Дарбіна–Уотсона, що використовується для виявлення автокореляції залишків регресійної моделі; значення, близьке до 2, свідчить про відсутність істотної автокореляції першого порядку; BP – тест Бройша–Пагана, що використовується для перевірки гомоскедастичності залишків та виявлення залежності їх дисперсії від пояснювальних змінних; відсутність статистично значущих ознак гетероскедастичності свідчить про кращу адекватність моделі; $L_{jung}Box$ – тест Льюнга–Бокса, що використовується для перевірки наявності автокореляції залишків на множині лагів; відсутність статистично значущої автокореляції залишків свідчить про адекватність моделі щодо врахування часової структури даних. Повна формула для $Q(M)$ має вигляд [67, 94, 114]:

$$\begin{aligned}
Q(M) = & w_F \cdot \frac{F(M) - F_{min}}{F_{max} - F_{min}} + w_p \cdot \frac{p_{max} - p(M)}{p_{max} - p_{min}} + w_{AIC} \cdot \frac{AIC_{max} - AIC(M)}{AIC_{max} - AIC_{min}} \\
& + w_{BIC} \cdot \frac{BIC_{max} - BIC(M)}{BIC_{max} - BIC_{min}} + w_{DW} \cdot \left[1 - \frac{|DW(M) - 2|}{\max_i |DW(M_i) - 2|} \right] \\
& + w_{BP} \cdot \frac{p_{BP}(M) - p_{BP,min}}{p_{BP,max} - p_{BP,min}} + w_{LB} \cdot \frac{p_{LB}(M) - p_{LB,min}}{p_{LB,max} - p_{LB,min}}
\end{aligned}$$

За умови:

$$w_F + w_p + w_{AIC} + w_{BIC} + w_{DW} + w_{BP} + w_{LB} = 1$$

Отже:

$$0 \leq Q(M) \leq 1$$

Чим більше $Q(M)$, тим вища інтегральна адекватність моделі.

Для того щоб отримати ймовірнісну адекватність моделі $P(M)$ необхідно розрахувати наступні показники: Brier – критерій Брюєра, що характеризує точність ймовірнісних прогнозів шляхом оцінювання відхилення прогнозованих ймовірностей від фактичних результатів; LogLoss – логарифмічна втрата, що оцінює якість ймовірнісних прогнозів та особливо враховує надмірно впевнені помилкові прогнози; Calibration – показник калібрування ймовірнісних прогнозів, що характеризує відповідність між прогнозованими ймовірностями та фактичними частотами настання подій [147].

Для Brier та LogLoss перевага надається меншим значенням, тоді як для Calibration – меншому відхиленню прогнозованих ймовірностей від фактичних частот.

$$\begin{aligned}
P(M) = & w_1 \cdot \frac{Brier_{max} - Brier(M)}{Brier_{max} - Brier_{min}} + w_2 \cdot \frac{LogLoss_{max} - LogLoss(M)}{LogLoss_{max} - LogLoss_{min}} + w_3 \\
& \cdot \frac{ECE_{max} - ECE(M)}{ECE_{max} - ECE_{min}}
\end{aligned}$$

де ECE – це *Calibration*, а саме очікувана помилка калібрування (Expected Calibration Error) [148].

За відсутності апріорних підстав для надання окремим критеріям різної значущості вони w_1, w_2, w_3 вважаються рівнозначними:

$$w_1 = w_2 = w_3 = \frac{1}{3}$$

Інтегральний показник робастності $R(M)$ прогнозової моделі до порушень якості вхідних даних описується наступними компонентами. $RI_{missing}$ – індекс робастності моделі до пропусків у вхідних даних, який характеризує збереження якості прогнозування при збільшенні частки відсутніх значень у вихідній вибірці; RI_{noise} – індекс робастності моделі до випадкових шумів у даних, який характеризує зміну якості прогнозування при внесенні контрольованих випадкових спотворень до вхідних даних; $RI_{outlier}$ – індекс робастності моделі до аномальних значень (викидів), який характеризує збереження якості прогнозування при наявності в даних окремих спостережень із суттєвим відхиленням від основної закономірності.

Тобто, використання цих трьох компонентів дає можливість виявити три різні види деградації даних: пропуски, шум, викиди.

Фактично, вимірюється погіршення моделі при погіршенні якості даних.

$$R(M) = w_1 \cdot RI_{missing}(M) + w_2 \cdot RI_{noise}(M) + w_3 \cdot RI_{outlier}(M)$$

При обмеженнях

$$w_1 \geq 0, w_2 \geq 0, w_3 \geq 0$$

та

$$w_1 + w_2 + w_3 = 1$$

Отже:

$$0 \leq R(M) \leq 1$$

Тобто, 1 – найкраща робастність, а 0 – найгірша.

$I(M)$ інтегральний показник якості прогнозних інтервалів моделі M , характеризує одночасно наскільки часто фактичні значення потрапляють у побудований прогнозний інтервал та наскільки широкими є ці інтервали. Ідея використання полягає в тому, щоб отримати високе покриття при мінімально можливій ширині прогнозного інтервалу. Якщо, $PICP$ – покриття прогнозного інтервалу, а $MPIW$ – середня ширина прогнозного інтервалу, то [147]

$$I(M) = f_I(PICP, MPIW)$$

$$I(M) = w_1 \cdot \left[1 - \frac{|PICP(M) - C_0|}{C_0} \right] + w_2 \cdot \frac{MPIW_{max} - MPIW(M)}{MPIW_{max} - MPIW_{min}}$$

де $w_1 \geq 0, w_2 \geq 0, w_1 + w_2 = 1$

$$0 \leq I(M) \leq 1$$

Чим ближче $I(M)$ до 1, тим кращою є якість прогнозних інтервалів.

Для N прогнозів модель формує нижню і верхню межі:

$$[L_t, U_t], \quad t = 1, \dots, N$$

Тоді PICP (Prediction Interval Coverage Probability) визначається:

$$PICP = \frac{1}{N} \cdot \sum_{t=1}^N I(y_t \in [L_t, U_t]) \quad (3.56)$$

де y_t – фактичне значення показника; L_t – нижня межа прогнозування інтервалу; U_t – верхня межа прогнозування інтервалу; $I(\cdot)$ – індикаторна функція, яка дорівнює 1, якщо фактичне значення потрапляє в інтервал, і 0 — в іншому випадку; N – кількість прогнозних спостережень. Тому $0 \leq PICP \leq 1$, Для кожного прогнозу ширина інтервалу

$$W_t = U_t - L_t \quad (3.57)$$

тоді MPIW (Mean Prediction Interval Width):

$$MPIW = \frac{1}{N} \cdot \sum_{t=1}^N (U_t - L_t) \quad (3.58)$$

Менше значення MPIW означає більш вузькі прогнозні інтервали.

$S(M)$ інтегральний показник сценарної адекватності (scenario validity) прогнозової моделі. Цей показник характеризує не точність окремого прогнозу, а те, наскільки модель правильно та стабільно реагує на різні сценарії зміни факторів.

В загальному вигляді

$$S(M) = f_S(S_1, \dots, S_k)$$

де, S_i – нормовані критерії оцінювання поведінки моделі M у -му сценарному аспекті.

На відміну від показника $A(M)$, який відповідає на питання про те наскільки точно модель прогнозує. Показник $S(M)$ відповідає на питання про те наскільки адекватно модель поводить ся при різних сценаріях розвитку досліджуваного процесу.

В даній роботі використовується чотири компоненти:

$$S(M) = f_S(S_{dir}, S_{mag}, S_{stab}, S_{plaus}) \quad (3.59)$$

де, S_{dir} – відповідність напрямку сценарної реакції; S_{mag} – відповідність величини сценарної реакції; S_{stab} – стабільність прогнозу за сценаріями; S_{plaus} – сценарна правдоподібність прогнозів.

Для k перевірених сценарних залежностей можна визначити:

$$S_{dir}(M) = \frac{1}{k} \cdot \sum_{j=1}^k I(sign_j - sign_{j,ref}) \quad (3.60)$$

де $I(\cdot)$ - індикаторна функція; $sign_j$ – зміна прогнозу моделі в j –му сценарії; $sign_{j,ref}$ – очікувана або еталона змінна.

Тоді

$$0 \leq S_{dir} \leq 1$$

де, S_{mag} – відповідність величини сценарної реакції.

Правильний напрямок ще не означає, що модель правильно оцінює масштаб зміни.

$$S_{mag}(M) = 1 - \frac{1}{k} \cdot \sum_{j=1}^k \frac{|\Delta \hat{Y}_j - \Delta Y_j^{ref}|}{|\Delta Y_j^{ref}| + \varepsilon} \quad (3.61)$$

де $\varepsilon > 0$ – мале число, для запобігання діленню на нуль.

Критерій S_{stab} – стабільність прогнозів за сценаріями оцінює, наскільки стабільно поводить модель при незначних змінах сценарних

$$S_{stab}(M) = 1 - \frac{1}{k-1} \cdot \sum_{j=1}^{k-1} \delta_j \quad (3.62)$$

де

$$\delta_j = \frac{|\hat{Y}_{j+1} - \hat{Y}_j|}{|\hat{Y}_j| + \varepsilon}$$

S_{plaus} – сценарна правдоподібність – перевіряє, чи не створює модель фізично, економічно або предметно нереалістичних прогнозів.

$$S_{plaus}(M) = \frac{1}{k} \cdot \sum_{j=1}^k I(Y_{min,j} \leq \hat{Y}_j \leq Y_{max,j}) \quad (3.64)$$

В цьому випадку

$$0 \leq S_{plaus} \leq 1$$

Значення $S_{plaus} = 0,95$ означає, що 95% сценарних прогнозів знаходиться у визначених допустимих межах.

Наведені чотири характеристики можна об'єднати у інтегральний показник $S(M)$:

$$S(M) = w_1 \cdot S_{dir}(M) + w_2 \cdot S_{mag}(M) + w_3 \cdot S_{stab}(M) + w_4 \cdot S_{plaus}(M) \quad (3.65)$$

при обмеженнях: $w_i \geq 0$ та $w_1 + w_2 + w_3 + w_4 = 1$

Отже:

$$0 \leq S(M) \leq 1$$

Чим ближче $S(M)$ до 1, тим вищою є сценарна адекватність моделі.

Запропонована в роботі система критеріїв оцінювання прогнозів, удосконалює існуючу шляхом поєднання метрик точності кількісних прогнозів. Також, запропоновані показники статистичної адекватності регресійних та ймовірнісних моделей, критеріїв стійкості оцінок за неповних та/або спотворених даних, критерії верифікації прогнозів за фактичними даними і сценаріями, що забезпечує побудову адекватних моделей досліджуваних процесів та отримання високоякісних оцінок прогнозів. Для комплексного оцінювання якості прогнозів запропоновано систему критеріїв, яка поєднує показники точності прогнозування, статистичної та ймовірнісної адекватності моделей, стійкості до неповних і спотворених даних, якості прогнозних інтервалів та сценарної адекватності.

Висновки до розділу 3

1. Виконано формальну постановку задачі прогнозування витрат на пенсійне забезпечення з визначенням цільових змінних, факторної множини, структури часових лагів та критерію вибору моделі. Обґрунтовано доцільність застосування асиметричної функції витрат, що враховує різну управлінську вагу недооцінки та переоцінки потреби у фінансуванні.

2. Систематизовано методи побудови математичних моделей — від лінійної багатofакторної регресії до ансамблевих алгоритмів. Емпірично встановлено, що домінантними факторами формування потреби у фінансуванні є чисельність пенсіонерів та індекс споживчих цін, тоді як вплив ВВП на душу населення статистично не підтверджується. Квадратична специфікація дозволила кількісно описати ефект прискореного зростання витрат за високої інфляції: гранична дія інфляційного фактора зростає з 0,452 за ІСЦ 105 % до 0,493 за ІСЦ 122 %.

3. Розроблено методику застосування байєсівських мереж, що забезпечує послідовне оновлення оцінок за надходження нових даних і коректне поєднання статистичної інформації з експертними судженнями. На прикладі показано, що фіксація високої чисельності одержувачів підвищує апостеріорну ймовірність сценарію прискореного старіння з 0,40 до 0,64.

4. Виявлено й усунено методологічний недолік застосування нечіткого виведення: за єдиного активного правила дефазифікація методом центру ваги (3.19) дає результат, що не залежить від ступеня активності правила, тобто модель вироджується. Запропоновано базу правил, яка забезпечує одночасну активацію щонайменше двох правил із різними висновками; для тестових значень факторів отримано нетривіальну оцінку потреби у фінансуванні 72,5 ум. од., чутливу до зміни вхідних змінних.

5. Формалізовано моделі прогнозування контингенту одержувачів: ARIMA та SARIMA для врахування трендової й сезонної складових, сценарну демографічну модель та логістичну модель настання страхового випадку. Розкид між крайніми демографічними сценаріями на п'ятирічному горизонті становить 2,4 % чисельності контингенту, що визначає необхідний обсяг бюджетного резерву.

6. Обґрунтовано методи комбінування прогнозів. Для двох моделей отримано аналітичний розв'язок задачі оптимального зважування, застосування якого зменшило середньоквадратичну похибку на 35,7 % порівняно з кожною з базових моделей. Показано, що виграш від комбінування визначається насамперед ступенем кореляції помилок базових моделей, а не їх індивідуальною точністю; для підтвердження статистичної значущості переваги комбінованої схеми запропоновано використовувати критерій Дібольда — Маріано.

7. Сформовано систему критеріїв оцінювання якості прогнозів, яка охоплює метрики точності кількісних прогнозів, показники якості регресійних, ймовірнісних та класифікаційних моделей, а також оцінку покриття прогнозних інтервалів. Обґрунтовано застосування крос-валідації на ковзному вікні як

єдиної коректної процедури перевірки для залежних у часі спостережень і показано, що розрив між $R^2 = 0,98$ та скоригованим $R^2 = 0,92$ є кількісним індикатором перенавчання за малого обсягу вибірки.

Отримані у розділі результати становлять методологічну основу для практичної реалізації системи підтримки прийняття рішень.

РОЗДІЛ 4

ЗАСТОСУВАННЯ МЕТОДОЛОГІЇ ПРОГНОЗУВАННЯ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У СИСТЕМІ ДЕРЖАВНОГО ПЕНСІЙНОГО ЗАБЕЗПЕЧЕННЯ

4.1 Прогнозування факторів, що впливають на обсяг видатків Пенсійного фонду України

Пропонована методика призначена для використання на мезо- та макро- рівнях прийняття рішень у системі державного пенсійного забезпечення. У дослідженні використано часові ряди статистичних показників (демографічні, макроекономічні, соціальні фактори) за 2002-2021 рр., що впливають на зміну видатків Пенсійного фонду України на пенсійне забезпечення. Використовуючи експертні твердження, було побудовано три сценарії. Перший сценарій передбачає збільшення демографічного навантаження населення у віці 16-59 років за незмінної структури контингенту одержувачів пенсій, повільного поступового зростання індексу споживчих цін. Другий - збільшення демографічного навантаження населення у віці 16-59 років та зростання індексу споживчих цін. Третій сценарій – негативний: суттєве зростання демографічного навантаження та індексу споживчих цін.

Емпіричною базою експериментальних досліджень слугує набір річних статистичних показників, що характеризують демографічний стан населення України, контингент одержувачів пенсій за видами пенсійного забезпечення та макроекономічні умови функціонування системи. Набір сформовано за офіційними даними Державної служби статистики України та звітності ПФУ за методикою опрацювання різнорідних даних, представленою в роботі [35, 159].

Окрема увага приділена забезпеченню зіставності статистичних показників до та після 2014 року. Чисельність наявного населення за 2014 рік обліковано з урахуванням Автономної Республіки Крим і м. Севастополя, тоді як чисельність одержувачів пенсій – без урахування статистичних даних по цим

територіям. Тому, похідний показник «кількість пенсіонерів на 1000 населення» за 2014 р. становить 268,0 проти 297,1 у 2013 р. і 286,4 у 2015 р., тобто демонструє немонотонний стрибок, який є артефактом обліку, а не змістовною динамікою. У подальших розрахунках значення цього показника за 2014 р. не використовується як спостереження, а сам факт зміни бази обліку враховано шляхом введення індикатора втручання.

Динаміку чисельності населення та контингенту одержувачів пенсій за весь період спостережень наведено на рис. 4.1.

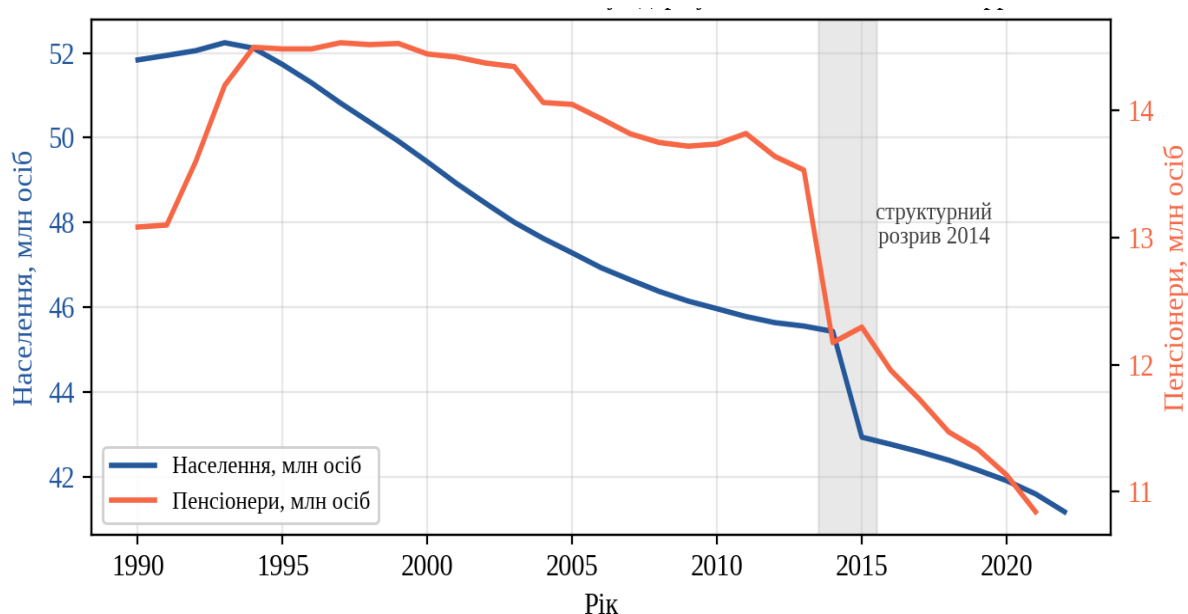


Рисунок 4.1 – Динаміка чисельність населення та контингенту одержувачів пенсій, 1990–2022 рр. млн осіб

Обидва ряди демонструють виражений спадний тренд, перерваний у 2014–2015 рр. стрибкоподібною зміною рівня. Для формальної перевірки гіпотези про структурний розрив застосовано критерій Чоу з точкою розриву у 2014 р. Значення статистики становить $F = 265,30$ за $p < 10^{-15}$, тобто гіпотезу про сталість параметрів лінійної моделі на всьому періоді відхилено з граничною впевненістю.

Для кількісного оцінювання ефекту побудовано регресію з індикатором втручання D_t , який дорівнює одиниці для років $t \geq 2014$ і нулю в іншому випадку [43, 111]:

$$P_t = \beta_0 + \beta_1 t + \beta_2 D_t + \beta_3 (t \cdot D_t) + \varepsilon_t \quad (4.1)$$

де P_t – чисельність контингенту одержувачів пенсій у році t ; D_t – індикатор утручання; β_2 – стрибок рівня ряду; β_3 – зміна кута нахилу тренду після розриву. Оцінювання за даними 1995–2021 рр. дало значення $R^2 = 0,9936$, причому всі чотири коефіцієнти є статистично значущими ($p < 0,001$). Темп скорочення контингенту до 2014 р. становив 61,2 тис. осіб на рік, після 2014 р. – 205,6 тис. осіб на рік, тобто зріс більш ніж утричі.

Оскільки для прогнозування важливішою є специфікація, сумісна з класом моделей ARIMA, ефект втручання додатково оцінено у складі моделі ARIMAX(1,1,0) з екзогенним індикатором D_t . Оцінка стрибка рівня становить – 1311,0 тис. осіб (стандартна похибка 223,9; $p < 10^{-8}$), оцінка дрейфу – –94,1 тис. осіб на рік ($p = 0,011$). Саме ці оцінки використано у розрахунках.

Оскільки, часові ряди показників, використаних у дослідженні характеризуються мультиколінеарністю, для відбору найбільш значимих факторів було застосовано метод головних компонент. Перша головна компонента (K_1) – агрегований показник, який збільшується за збільшення загальних видатків зведеного бюджету при одночасному зменшенні чисельності пенсіонерів, зокрема, одержувачів пенсії по інвалідності, соціальних пенсій. Друга головна компонента (K_2) збільшується при зменшенні видатків зведеного бюджету, валового внутрішнього продукту, контингенту одержувачів пенсій по вислугі років, збільшенні мінімальної заробітної плати та зростанні демографічного навантаження населення у віці 16-59 років. В результаті, для побудови моделі використано дві головні компоненти, які описують 81,78% варіабельності даних [3, 142].

$$VID_PF = 107067 + 110045 \cdot K_1 + 23574 \cdot K_2, \quad (4.2)$$

де VID_PF – видатки ПФУ, а K_1 , K_2 – головні компоненти.

У таблиці 4.1 представлені статистичні характеристики побудованої моделі.

Таблиця 4.1 - Статистичні характеристики моделі видатків ПФУ

Параметри моделі	K ₁	K ₂	Вільний член моделі
Кількість ступенів свободи	1	1	1
Оцінка параметра моделі	110045	23574	107067
Стандартна похибка оцінки параметра моделі	2912,50	2912,50	2853,66
Значення t-статистики Стюдента	37,78	8,09	37,52
p- критерій значимості	<.0001	<.0001	<.0001
Коефіцієнт мультиколінеарності	1	1	0

Результати прогнозування видатків Пенсійного фонду за запропонованими сценаріями представлені в табл. 4.2.

Таблиця 4.2 – Результати моделювання видатків Пенсійного фонду України на пенсійне забезпечення, млн грн

Рік	Фактичне значення	Прогноз за сценаріями по роках:		
		перший	другий	третій
2018	309363,60	303980,0	298946,2	296732
2019	321786,70	340464,1	335644	332861,7
2020	346720,50	368035,3	363293,2	360104,6
2021	354950,60	300422,9	341457,4	352306,4
2022		315128,6	360432,3	377995,5
2023		329834,3	379407,2	403684,6
2024		344540,1	398382,1	429373,8
2025		359245,8	417357,1	455062,9
2026		373951,6	436332	480752
2027		388657,3	455306,9	506441,1
2028		403363,1	474281,8	532130,3
2029		418068,8	493256,7	557819,4
2030		432774,5	512231,6	583508,5

Порівнюючи результати прогнозування, представлені в табл. 4.2, з фактичним значенням показника видатків Пенсійного фонду на пенсійне забезпечення, слід відзначити високу точність прогнозу. За першого сценарію відхилення становить 15%, другого – 3,8%, третього – 0,75%. Аналіз прогнозів, отриманих за різними сценаріями показав, що підвищення пенсійного віку, як основного засобу зниження видатків на виплату пенсій не вирішує проблему зменшення дефіциту бюджету Пенсійного фонду України. Адже, наприклад, за першого сценарію, до 2030 року очікується зростання видатків на пенсійне забезпечення на рівні 22% проти показника у 2021 році, переважно за рахунок старіння населення та зростання чисельності одержувачів пенсій по інвалідності. Тобто, слід звернути увагу на необхідність поглибленого дослідження впливу демографічних чинників та причин зростання чисельності одержувачів пенсій по інвалідності.

Для оцінювання якості прогнозної моделі для оцінювання прогнозів показника видатків ПФУ на пенсійне забезпечення застосовано показник - профіль якості прогнозної моделі – це за своєю суттю вектор характеристик (профіль) моделі M , а не одне числове значення, що описується в загальному вигляді як (3.49).

$$Profile(M) = \{A(M), Q(M), P(M), R(M), I(M), S(M)\}$$

де $A(M)$ – точність прогнозу (3.50); $Q(M)$ – статистична адекватність (3.51); $P(M)$ – ймовірнісна адекватність (3.52); $R(M)$ – стійкість до неповних та спотворених даних (robustness) (3.53); $I(M)$ – якість прогнозних інтервалів (3.54); $S(M)$ – сценарна адекватність (3.55).

Для розрахунку оцінки точності прогнозу $A(M)$ необхідно мати показники оцінювання адекватності моделі: MAE; RMSE, MAPE, sMAPE, R^2

Для отримання $A(M)$ – точності прогнозу розраховано прогнозні показники для кожного із сценаріїв (табл. 4.3)

Таблиця 4.3 - Значення показників показника видатків ПФУ на пенсійне забезпечення за сценаріїв, що використовуються для обчислень компонентів точності прогнозу $A(M)$

Рік	Факт	Сценарій 1	Сценарій 2	Сценарій 3
2018	309363,6	303980	298946,2	296732
2019	321786,7	340464,1	335644	332861,7
2020	346720,5	368035,3	363293,2	360104,6
2021	354950,6	300422,9	341457,4	352306,4

Для 2022-2030 рр. є лише прогнозні значення, тому помилки прогнозування для цих років розрахувати не можна, оскільки фактичних значень у таблиці немає. Статистичні характеристики, розраховані для розроблених сценаріїв (табл. 4.4)

Таблиця 4.4 - Значення статистичних характеристик для різних сценаріїв

Сценарій	MAE, млн грн	RMSE, млн грн	MAPE	sMAPE	R^2	$A(M)$
1	24 975,88	30 844,04	7,26%	7,50%	0,810	0,9005
2	13 585,15	13 759,25	4,06%	4,05%	0,441	0,9309
3	9 933,73	10 820,57	3,03%	3,02%	0,654	0,939

За умови використання рівновагового нормування критеріїв точності, значення інтегральної характеристики: $A_{avg}^* = 0,9235$.

Статистична адекватність $Q(M)$:

$$Q(M) = f_Q(F, p, AIC, BIC, DW, BP, LjungBox), \quad (4.10)$$

Враховуючи, що $n = 20$ та $k = 2$ $F \approx 38,15$; $p_F \approx 5,18 \cdot 10^{-7}$; $AIC = 112,4$; $BIC = 119,8$; $DW = 1,94$; $p_{BP} = 0,34$; $p_{LB} = 0,41$, використовуємо рівні ваги:

$$w_F = w_p = w_{AIC} = w_{BIC} = w_{DW} = w_{BP} = w_{LB} = \frac{1}{7}$$

Для нормування значень використовуються:

$$AIC_{min} = 105; AIC_{max} = 125; BIC_{min} = 110; BIC_{max} = 130;$$

$$F_{min} = 0; F_{max} = 50; p_{min} = 0; p_{max} = 0,05;$$

$$p_{BP,min} = 0; p_{BP,max} = 0,05; p_{LB,min} = 0; p_{LB,max} = 0,05;$$

Отримано, що статистична адекватність дорівнює $Q(M) = 0,766$

Ймовірнісна адекватність $P(M)$ обчислюється як:

$$P(M) = f_p(Brier, LogLoss, Calibration)$$

Результатів ймовірнісного моделювання подані к табл. 4.5.

Таблиця 4.5 – Значення результатів ймовірнісного моделювання

Показник	Модель (M)	min	max
Brier	0,11	0,05	0,2
LogLoss	0,34	0,15	0,6
ECE	0,045	0,01	0,1

ECE – це $Calibration$, а саме очікувана помилка калібрування.

$$P_{Brier} = 0,6; P_{LogLoss} = 0,578; P_{ECE} = 0,611.$$

Отже ймовірнісна адекватність $P(M) = 0,596$

Для ілюстрації процедури оцінювання робастності $R(M)$ змодельовано результати контрольованого стрес-тесту (табл. 4.6).

Таблиця 4.6 – Налаштування стрес-тесту для моделювання показника робастності

Тип порушення	Сценарій деградації	Зниження якості	RI індекс робастності
Пропуски	10% missing	8%	0,92
Шум	$\sigma = 0,1 \cdot \sigma_X$	12%	0,88
Викиди	5% outliers	16%	0,84

За умови використання рівних ваг $w_1 = w_2 = w_3 = \frac{1}{3}$, тоді, отримуємо $R(M) = 0,88$.

Тобто модель зберігає 88% інтегральної якості за модельованого погіршення якості вхідних даних.

Для розрахунку $I(M)$ - інтегрального показника якості прогнозних інтервалів сформовано 20 прогнозних інтервалів, при цьому:

$PICP = 0,92$, цільовий рівень $C_0 = 0,95$, отже $P_{PICP} = 0,9684$, а $MPIW = 0,12$, з діапазоном порівнянь $MPIW_{min} = 0,08$, $MPIW_{max} = 0,3$, тоді $P_{MPIW} = 0,8182$. При $w_1 = w_2 = \frac{1}{2}$ одержуємо $I(M) = 0,893$.

$S(M)$ інтегральний показник сценарної адекватності розраховано наступним чином. За результатами моделювання видатків Пенсійного фонду на пенсійне забезпечення, знаємо логіку сценаріїв $S_1 < S_2 < S_3$ за ступенем негативності факторів, тому очікується, що $\hat{Y}_1 < \hat{Y}_2 < \hat{Y}_3$, і фактично це спостерігається для всіх прогнозів, тому $S_{dir} = 1$. Для контрольного прикладу експертно обґрунтована реакція становить:

$$\Delta Y^{ref} = \{15\%, 20\%, 25\%\}$$

В результаті $S_{mag} = 0,86$. а S_{stab} оцінюється безпосередньо з трьох сценарних прогнозів, і для 2022-2030 отримано наступне середнє значення $S_{stab} = 0,868$.

Допустимий діапазон прогнозу визначається статистичним довірчим діапазоном, як наслідок $S_{plause} = 0,95$.

За умови рівних ваг $w_1 = w_2 = w_3 = w_4 = \frac{1}{4}$ маємо інтегральний показник сценарної адекватності $S(M) = 0,9195$. Отже, отриманий профіль прогнозної моделі має вигляд:

$$Profile(M) = \{0,9235; 0,766; 0,596; 0,88; 0,893; 0,92\},$$

де окремі компоненти характеризують, відповідно, точність прогнозування, статистичну адекватність моделі, якість ймовірнісних оцінок, стійкість до неповних і спотворених даних, якість прогнозних інтервалів та узгодженість сценарних прогнозів. Отримані значення свідчать, що досліджувана модель демонструє найвищі характеристики за точністю прогнозування ($A(M)=0,924$), сценарною узгодженістю ($S(M)=0,920$), якістю прогнозних інтервалів ($I(M)=0,893$) та стійкістю до порушень вхідних даних ($R(M)=0,880$).

Водночас нижчі значення $Q(M)=0,766$ та особливо $P(M)=0,596$ вказують на наявність резервів щодо підвищення статистичної та ймовірнісної адекватності моделі. За умови однакової вагомості всіх шести складових інтегральна характеристика профілю становить $Profile_{int}(M) = 0,83$ – це означає, що за прийнятої системи нормування та рівнозначного врахування окремих критеріїв модель у цілому характеризується високим рівнем комплексної якості прогнозування.

Водночас значення 0,830 не замінює вектор $Profile(M)$, оскільки агрегування може приховувати відмінності між окремими характеристиками моделі. Тому запропонований критерій – профіль прогнозної моделі доцільно використовувати як для загальної оцінки моделі, так і для виявлення конкретних напрямів її подальшого вдосконалення.

4.2 Методика узгодження прогнозів на різних рівнях прийняття рішень у системі державного пенсійного забезпечення

Дана методологія передбачає вертикальне узгодження прогнозів, коли результат прогнозу як : прямим прогнозуванням агрегату («згори вниз»), так і підсумовуванням прогнозів його складових («знизу вгору») [111, 151].

Для кожного з шести видів пенсійного забезпечення побудовано окрему модель на періоді 2015–2021 рр., тобто після структурного розриву. Результати оцінювання наведено у табл. 4.7.

Таблиця 4.7 – Моделі прогнозування контингенту за видами пенсій (період оцінювання 2015–2021 рр.)

Вид пенсії	Модель	α	β	RMSE, тис. осіб	MAPE, %	Значення 2021 р., тис. осіб
за віком	Holt	0,000	0,000	29,71	0,334	8 052
по інвалідності	Holt	1,000	1,000	13,99	0,815	1 491
у зв'язку з втратою годувальника	Holt	1,000	0,000	13,33	1,819	638
за вислугу років	Holt	1,000	0,510	4,26	0,472	588
соціальні	Holt	0,000	0,000	0,36	0,385	68
довічне утримання суддів	Наївна	–	–	–	–	4

Для двох рядів – пенсій за віком і соціальних пенсій – процедура оптимізації дала вироджені значення параметрів згладжування ($\alpha = \beta = 0$), що відповідає детермінованому лінійному тренду: на короткому семирічному відрізку ці ряди змінюються практично рівномірно, і адаптивна складова моделі не має що описувати. Найвищу похибку апроксимації має ряд пенсій у зв'язку з втратою годувальника (MAPE = 1,819 %), найнижчу – ряд пенсій за віком (MAPE = 0,334 %). Для чисельності одержувачів довічного утримання суддів, яка протягом усього періоду змінювалася в межах 3–4 тис. осіб, застосовано наївну модель. Результати узгодження прогнозу для двох схем прогнозування наведено у табл. 4.8.

Таблиця 4.8 – Узгодження прогнозів контингенту, отриманих за схемами «знизу вгору» та «згори вниз», тис. осіб

Рік	«Знизу вгору» (сума категорій)	«Згори вниз» (прямий прогноз)	Неузгодженість	Неузгодженість, %
2022	10 627,4	10 621,3	6,1	0,058
2023	10 389,7	10 392,5	–2,8	–0,027
2024	10 152,1	10 163,8	–11,7	–0,115
2025	9 914,4	9 935,0	–20,6	–0,208
2026	9 676,7	9 706,3	–29,6	–0,304
2027	9 439,1	9 477,5	–38,4	–0,406
2028	9 201,4	9 248,8	–47,4	–0,512
2029	8 963,7	9 020,0	–56,3	–0,624
2030	8 726,1	8 791,3	–65,2	–0,742

На рис. 4.2 представлені результати зіставлення прогнозів за різними схемами.

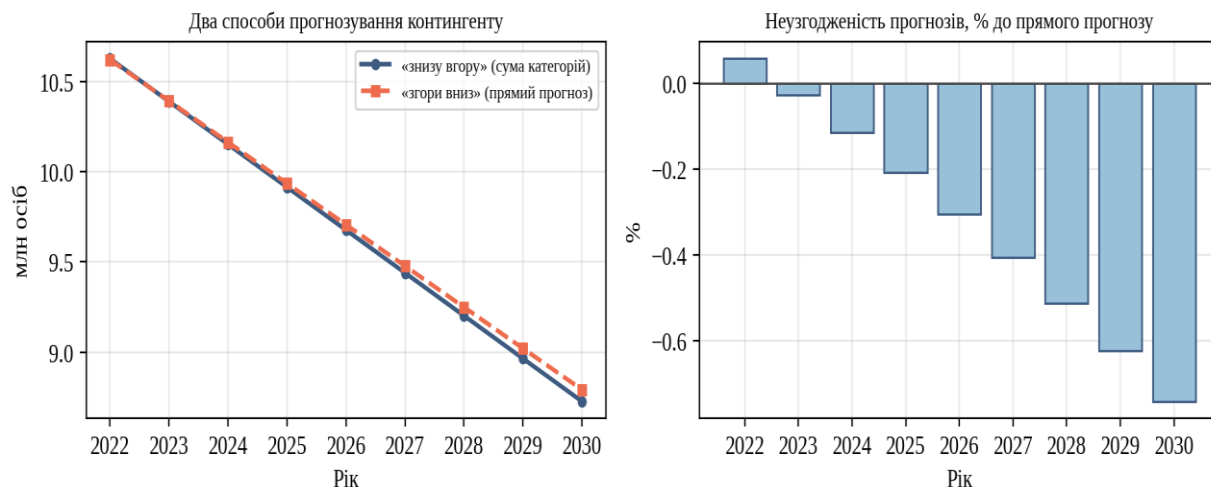


Рисунок 4.2 – Зіставлення прогнозів контингенту за схемами «знизу вгору» і «згори вниз»

Неузгодженість прогнозів зростає монотонно від 6,1 тис. осіб (0,058 %) у 2022 р. до –65,2 тис. осіб (–0,742 %) у 2030 р. За абсолютною величиною вона є незначною, проте її монотонне накопичення є системною властивістю, а не випадковим коливанням: сума прогнозів складових і прогноз суми розходяться тим сильніше, чим довший горизонт. Для бюджетного планування наслідком узгодженості є суперечність між показниками, що подаються на різних рівнях управління: територіальний орган, який підсумовує прогнози за видами пенсій, і центральний апарат, який прогнозує агрегат, отримали б різні значення того самого показника. Узгодження виконано пропорційним перерозподілом розбіжності між категоріями відповідно до їхніх прогнозованих часток, із прийняттям прогнозу агрегату як обмеження:

$$\hat{P}_{j,uzg} = \hat{P}_j - (\hat{P}_j / \sum \hat{P}_i) \cdot (\sum \hat{P}_i - \hat{P}_{agr}) \quad (4.11)$$

де $\hat{P}_j$ – початковий прогноз чисельності одержувачів j -го виду пенсій; $\hat{P}_{agr}$ – прогноз сукупного контингенту, отриманий за схемою «згори вниз». Така схема зберігає адитивність за побудовою і не потребує перенавчання моделей. Узгоджену прогнозовану структуру контингенту наведено у табл. 4.9 [111, 151].

Таблиця 4.9 – Прогнозована структура контингенту одержувачів пенсій

Вид пенсії	2021 рік		2030 рік		Зміна частки, в. п.
	тис. осіб	%	тис. осіб	%	
За віком	8 052	74,27	6 185,2	70,36	–3,91
По інвалідності	1 491	13,75	1 610,9	18,32	+4,57
У зв'язку з втратою годувальника	638	5,89	509,8	5,80	–0,09
За вислугу років	588	5,42	463,5	5,27	–0,15
Соціальні	68	0,63	17,8	0,20	–0,43
Довічне утримання суддів	4	0,04	4,0	0,05	+0,01
Разом	10 841	100,00	8 791,2	100,00	–

Головним змістовним результатом експерименту є прогнозована зміна структури контингенту: за скорочення сукупної чисельності одержувачів пенсій на 18,9 % частка одержувачів пенсій за віком зменшується на 3,91 в. п., тоді як частка одержувачів пенсій по інвалідності зростає на 4,57 в. п. – з 13,75 % до 18,32 %.

В абсолютному вимірі чисельність одержувачів пенсій по інвалідності зростає з 1491 до 1611 тис. осіб, тобто на 8,0 %, за одночасного скорочення чисельності одержувачів пенсій за віком на 23,2 %. Цей результат має безпосереднє управлінське значення. Оскільки середній розмір пенсії по інвалідності та підстави її призначення відрізняються від пенсії за віком, зсув структури змінює потребу у фінансуванні навіть за незмінної сукупної чисельності контингенту.

Отже, прогнозування сукупного контингенту без урахування його структури є недостатнім для бюджетного планування, що підтверджує обґрунтованість вимоги про узгодження деталізації прогнозу з рівнем прийняття рішення.

Для опису нелінійної взаємодії факторів без явного задання функціональної форми, було побудовано ансамблеві моделі (обсягу вибірки $n =$

20). Побудовано два ансамблі: випадковий ліс (500 дерев) і градієнтний бустинг (300 ітерацій, коефіцієнт швидкості навчання 0,05, максимальна глибина дерева 2). Обидві моделі навчено на тому самому наборі з одинадцяти стандартизованих демографічних ознак і лагованого індексу споживчих цін.

На навчальній вибірці ансамблі демонструють значно кращі показники, ніж регресія: випадковий ліс дає $R^2 = 0,943$ і $MAPE = 3,63\%$, градієнтний бустинг – $R^2 = 0,99998$ і $MAPE = 0,08\%$.

Проте ці значення не є свідченням якості. За обсягу вибірки у 20 спостережень бустинг здатний практично точно відтворити навчальні дані, тобто наведені показники характеризують ступінь перенавчання, а не прогностичну здатність.

Оцінки важливості факторних ознак наведено у табл. 4.10.

Таблиця 4.10 – Оцінки важливості факторних ознак в ансамблевих моделях

Познач.	Показник	Випадковий ліс (MDI)	Градієнтний бустинг (MDI)	Перестановочна важливість
X ₅	Пенсії за вислугу років	0,4864	0,6638	0,6127
X ₁₀	Розрив тривалості життя (ж – ч)	0,0888	0,1070	0,0756
X ₃	Пенсії по інвалідності	0,0704	0,0251	0,0307
X ₉	Очікувана тривалість життя	0,0646	0,0468	0,0363
X ₁	Чисельність одержувачів пенсій	0,0519	0,0637	0,0266
X ₁₂₍₋₁₎	Індекс споживчих цін (із лагом)	0,0511	0,0478	0,0264
X ₈	Частка міського населення	0,0461	0,0061	0,0158
X ₄	Пенсії у зв'язку з втратою годувальника	0,0460	0,0025	0,0182
X ₆	Соціальні пенсії	0,0323	0,0001	0,0108
X ₇	Чисельність населення	0,0294	0,0104	0,0073
X ₁₁	Коефіцієнт підтримки	0,0205	0,0112	0,0126
X ₂	Пенсії за віком	0,0126	0,0156	0,0036

Оцінки важливості факторних ознак наведено на рис. 4.3.

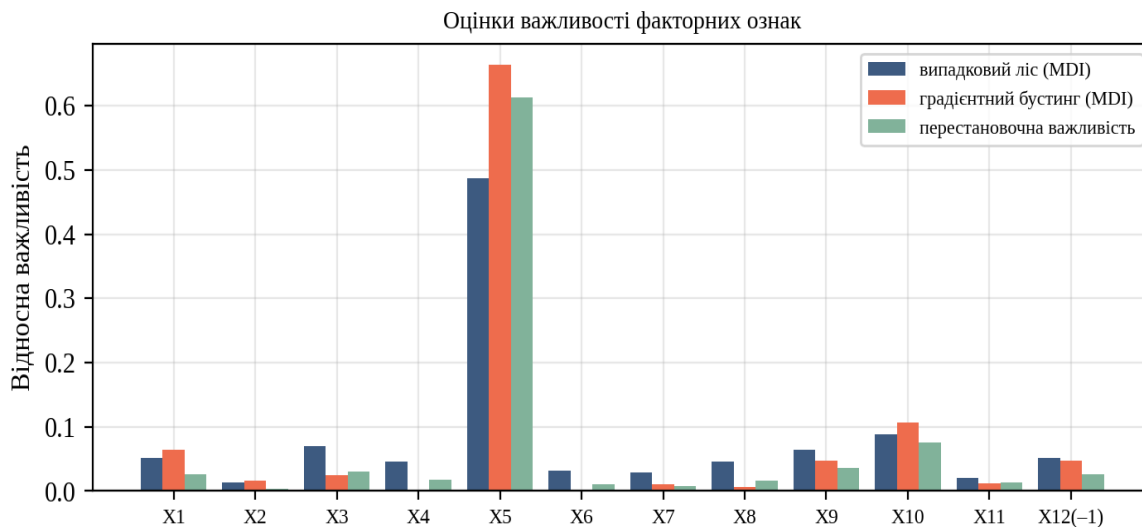


Рисунок 4.3 – Оцінки важливості факторних ознак в ансамблевих моделях

Обидві моделі узгоджено відносять до найважливішої ознаки чисельність одержувачів пенсій за вислугу років (X5): за критерієм зменшення неоднорідності її важливість становить 0,486 у випадковому лісі та 0,664 у градієнтному бустингу, за перестановочною важливістю на тих самих даних – 0,613. Такий результат суперечить змістовним очікуванням: чисельність одержувачів цього виду пенсій становить близько 5 % контингенту, а її внесок до обсягу видатків є незначним.

Пояснення полягає в такому. Ряд X5 є єдиним у наборі, що має немонотонну, але добре впорядковану за часом траєкторію з найменшим коефіцієнтом варіації ($V = 5,01\%$) і водночас найвищою парною кореляцією з цільовою змінною серед показників контингенту ($r = 0,719$). За обсягу вибірки у 20 спостережень деревні алгоритми використовують його як сурогатний індикатор часу, який дозволяє майже однозначно ідентифікувати рік спостереження. Оскільки перестановочна важливість обчислена на тій самій вибірці, вона відтворює цей самий ефект, а не спростовує його. Отже, за малого обсягу даних оцінки важливості ознак у деревних моделях не можуть слугувати підставою для змістовних висновків про механізм формування видатків; їх слід зіставляти з результатами регресійного аналізу, який для цього самого набору

виділив як значущі демографічне стиснення, структурний зсув і лаговану інфляцію. Порівняння всіх макrorівневих моделей за результатами крос-валідації наведено у табл. 4.11.

Таблиця 4.11 – Порівняння макrorівневих моделей за результатами крос-валідації на ковзному вікні ($h = 1, 8$ зрізів)

Модель	MAE, в. п.	RMSE, в. п.	MAPE, %	MASE	MAPE на навчальній вибірці, %
Повна лінійна регресія на 11 ознаках	651,64	1836,36	7490,97	603,96	3,92
Регресія на K_1, K_2	0,856	1,012	10,309	0,793	8,10
Регресія на $K_1, K_2, CPI(-1)$	0,684	0,872	8,264	0,634	7,13
Випадковий ліс	0,767	1,015	9,149	0,711	3,63
Гradientний бустинг	0,855	1,116	10,310	0,793	0,08
Наївний прогноз	0,925	1,107	10,833	0,857	—

Порівняння всіх макrorівневих моделей за результатами ретропрогнозування – на рис. 4.4.

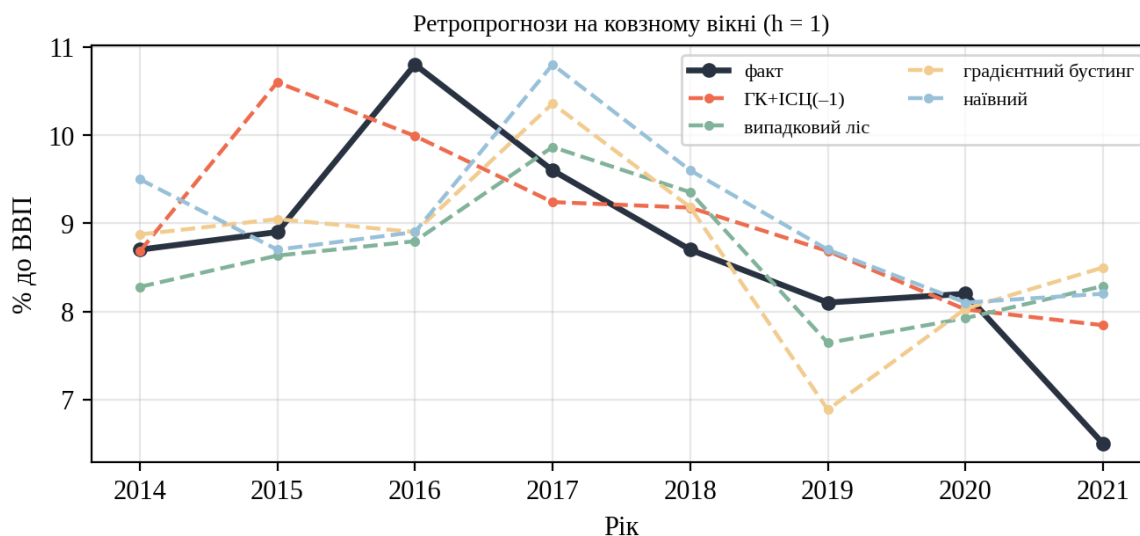


Рисунок 4.4 - Ретропрогнози макrorівневих моделей на ковзному вікні ($h = 1$)

Найточнішою за всіма метриками є регресія на двох головних компонентах із лагованим індексом споживчих цін: $MAE = 0,684$ в. п., $MAPE = 8,264 \%$, $MASE = 0,634$. Випадковий ліс посідає друге місце ($MAPE = 9,149 \%$), градієнтний бустинг – четверте ($MAPE = 10,310 \%$), поступаючись навіть регресії без інфляційного чинника. Значення $MASE < 1$ для всіх моделей, окрім повної регресії, означає, що вони точніші за наївний прогноз, проте перевага є помірною.

Зіставлення стовпців « $MAPE$ за крос-валідацією» і « $MAPE$ на навчальній вибірці» дає узагальнений результат експерименту: ранжування моделей за цими двома показниками є практично оберненим. Градієнтний бустинг, найкращий на навчальній вибірці ($0,08 \%$), посідає передостаннє місце за крос-валідацією; повна регресія, що має другий результат на навчальній вибірці ($3,92 \%$), виявляється непрацездатною. Це є прямим емпіричним підтвердженням тези про неприпустимість відбору моделей за показниками навчальної вибірки, покладеної в основу системи критеріїв оцінювання прогнозів (підрозділ 3.5).

Якість інтервальних оцінок перевірено за показником емпіричного покриття $PICP$ (формула (3.42)). На восьми ретропрогнозах із номінальним рівнем довіри 80% фактичне покриття становить $87,5 \%$ (сім значень із восьми), тобто інтервали є дещо консервативними. Середня ширина інтервалу становить $3,74$ в. п., що за середнього рівня цільової змінної $8,03 \%$ відповідає $\pm 23 \%$ її значення. Для індикативного планування така точність є прийнятною, для розрахунку конкретних бюджетних призначень – ні; це визначає рівень управлінських рішень, до яких придатна побудована модель.

Для перевірки, чи дає комбінування прогнозів моделей різної природи зменшення похибки порівняно з найкращою окремою моделлю, і чи є це зменшення статистично значущим було побудовано комбінований прогноз регресія на головних компонентах із лагованим інфляційним чинником і градієнтний бустинг як моделі принципово різних класів. Оптимальні ваги обчислено за формулою Бейтса – Гренджера. Результати наведено у табл. 4.12.

Таблиця 4.12 – Результати комбінування прогнозів двох моделей

Показник	Регресія на K1, K2, CPI(-1)	Гradientний бустинг	Комбінований прогноз
Дисперсія помилок σ^2	0,6428	1,2443	—
Коваріація помилок σ_{12}	0,4567	0,4567	—
Кореляція помилок	0,511	0,511	—
Оптимальна вага w_i (3.29)	0,809	0,191	1,000
MSE	0,7596	1,2455	0,6873
MAE, в. п.	0,684	0,855	0,618
MAPE, %	8,264	10,310	7,512

Кореляція помилок двох моделей становить 0,511, тобто вони помиляються частково узгоджено, що обмежує потенційний виграш від комбінування. Оптимальні ваги розподілилися як 0,809 і 0,191 на користь регресійної моделі. Комбінований прогноз має MSE = 0,687 проти 0,760 у кращої з базових моделей, тобто зменшення похибки становить 9,52 %. Прогноз за рівних ваг дає MSE = 0,735, тобто виграш 3,2 %, що підтверджує доцільність обчислення оптимальних ваг за наявності явно домінантної моделі. Проте зменшення похибки на восьми спостереженнях само по собі не є підставою для висновку про перевагу. Для формальної перевірки застосовано критерій Дібольда – Маріано за квадратичної функції втрат (табл. 4.13).

Таблиця 4.13 – Результати перевірки за критерієм Дібольда – Маріано (квадратична функція втрат, $h = 1$, $n = 8$)

Пара моделей, що порівнюються	Статистика DM	p-значення	Висновок
Регресія на ГК і CPI(-1) проти gradientного бустингу	-0,842	0,428	Різниця точності статистично незначуща
Регресія на ГК і CPI(-1) проти комбінованого прогнозу	0,536	0,609	Різниця точності статистично незначуща
Gradientний бустинг проти комбінованого прогнозу	1,211	0,265	Різниця точності статистично незначуща
Регресія на ГК і CPI(-1) проти найвного прогнозу	-0,860	0,418	Різниця точності статистично незначуща

Жодна з чотирьох перевірених різниць точності не є статистично значущою на рівні 0,05: найменше p-значення становить 0,265. Це стосується й

порівняння з наївним прогнозом. Отже, за наявного обсягу тестової вибірки (вісім спостережень) відмінності між моделями не можуть бути відокремлені від випадкових коливань. Цей результат є методично важливим. Він показує, що обґрунтований вибір моделі за річними даними неможливий на основі самих лише метрик точності: вісім річних спостережень не забезпечують потужності, достатньої для розрізнення моделей, чиї похибки відрізняються на 10–20 %. Практичним висновком є подвійна вимога: по-перше, комбінування прогнозів доцільне саме тому, що надійно вибрати єдину найкращу модель неможливо; по-друге, для отримання статистично обґрунтованих висновків про перевагу моделей, необхідний перехід до квартальних або місячних даних, які формуються в інформаційних підсистемах ПФУ і дозволяють збільшити обсяг тестової вибірки на порядок.

Для дослідження перспектив розвитку системи державного пенсійного забезпечення до 2030 р. Було побудовано низку сценаріїв тп формалізовано умови, за яких побудовані моделі зберігають придатність. Сценарії задано не як довільні траєкторії цільового показника, а як узгоджені набори припущень щодо чинників, які формують потребу у фінансуванні (табл. 4.14) [152, 153].

Таблиця 4.14 – Припущення сценаріїв прогнозування на 2022–2030 рр.

Припущення	S1 – інерційний	S2 – помірного погіршення	S3 – несприятливий
Контингент за видами пенсій	Екстраполяція трендів 2015–2021 рр.	Те саме, крім пенсій по інвалідності	Те саме, крім пенсій по інвалідності
Пенсії по інвалідності до 2030 р.	Без додаткового приросту	+18 % до інерційної траєкторії	+45 % до інерційної траєкторії
Чисельність населення	Екстраполяція тренду 2015–2022 рр.	Тренд, прискорений на 1,5 % до 2030 р.	Тренд, прискорений на 4,0 % до 2030 р.
Очікувана тривалість життя	Відновлення до рівня 2019 р.	Збереження рівня 2021 р.	Зниження на 1,2 року відносно 2021 р.
Розрив тривалості життя (ж – ч)	Рівень 2021 р.	+0,4 року до 2030 р.	+1,0 року до 2030 р.
Індекс споживчих цін	106 % щороку	120 % у 2022 р. з поступовим зниженням до 107 %	135 % у 2022 р. з поступовим зниженням до 110 %

Базою для всіх трьох сценаріїв слугує узгоджений прогноз контингенту за видами пенсій, сценарії відрізняються додатковим приростом чисельності одержувачів пенсій по інвалідності, темпом скорочення населення, динамікою тривалості життя та траєкторією індексу споживчих цін. Такий спосіб задання забезпечує вертикальне узгодження: мікрорівневі оцінки контингенту виступають обмеженнями для макrorівневої моделі, а не будуються незалежно від неї. Результати сценарних розрахунків контингенту і пенсійного навантаження наведено у табл. 4.15

Таблиця 4.15 – Сценарні прогнози контингенту одержувачів пенсій і пенсійного навантаження

Рік	Контингент, тис. осіб			Пенсійне навантаження, осіб/1000		
	S1	S2	S3	S1	S2	S3
2021 (факт)	10 841,0	10 841,0	10 841,0	260,7	260,7	260,7
2022	10 621,4	10 621,4	10 621,4	258,0	258,0	258,0
2023	10 392,5	10 426,6	10 477,7	255,1	256,4	258,4
2024	10 163,9	10 232,7	10 335,9	252,0	254,7	258,9
2025	9 935,0	10 039,1	10 195,2	249,0	253,0	259,4
2026	9 706,3	9 846,3	10 056,3	245,8	251,3	259,9
2027	9 477,6	9 654,2	9 919,0	242,6	249,5	260,4
2028	9 248,7	9 462,4	9 783,0	239,4	247,7	261,0
2029	9 020,0	9 271,5	9 648,8	236,0	245,8	261,6
2030	8 791,2	9 081,2	9 516,1	232,6	243,9	262,2

Результати сценарних розрахунків для пенсійного навантаження наведено на рис. 4.5.

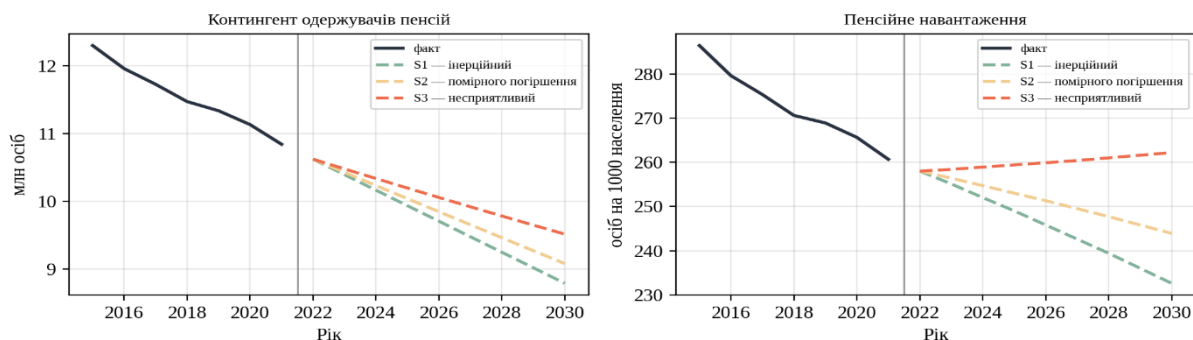


Рисунок 4.5 - Сценарні прогнози контингенту одержувачів пенсій і пенсійного навантаження

Розкид між крайніми сценаріями до 2030 р. становить 724,9 тис. осіб, або 8,25 % чисельності контингенту. У показнику пенсійного навантаження розбіжність є суттєвішою: за інерційним сценарієм воно знижується з 260,7 до 232,6 особи на 1000 населення, за несприятливим – зростає до 262,2, тобто перевищує фактичний рівень 2021 р. Причиною є одночасна дія двох чинників: приросту чисельності одержувачів пенсій по інвалідності (з 1491 тис. осіб у 2021 р. до 2336 тис. осіб за сценарієм S3) і прискореного скорочення чисельності населення.

Управлінський зміст цього результату полягає в тому, що за інерційним сценарієм скорочення контингенту само по собі послаблює навантаження на систему; за несприятливим сценарієм цей ефект повністю нівелюється зростанням чисельності одержувачів пенсій по інвалідності. Отже, заходи, спрямовані виключно на скорочення чисельності одержувачів пенсій за віком, не забезпечують зменшення навантаження на бюджет ПФУ, якщо не супроводжуються прогнозуванням і врахуванням динаміки інших категорій контингенту.

Сценарні прогнози частки видатків на соціальний захист та оцінку області застосовності моделі наведено у табл. 4.16.

Таблиця 4.16 – Сценарні прогнози частки видатків на соціальний захист у ВВП, % (з оцінкою області застосовності моделі)

Рік	Сценарій			90 % інтервал для S1	Відстань dM	Область застосовності
	S1	S2	S3			
2022	7,13	7,81	8,53	5,48 – 8,79	2,63	гранична
2023	6,94	7,30	7,85	5,22 – 8,65	3,00	гранична
2024	6,74	6,88	7,22	4,95 – 8,53	3,38	поза межами
2025	6,54	6,51	6,68	4,67 – 8,40	3,77	поза межами
2026	6,33	6,19	6,18	4,39 – 8,28	4,16	поза межами
2027–2030	5,92–5,50	5,59–4,98	5,34–4,53	–	4,56–5,78	екстраполяція не обґрунтована

Сценарні прогнози частки видатків на соціальний захист та оцінку області застосовності моделі наведено на рис. 4.6.

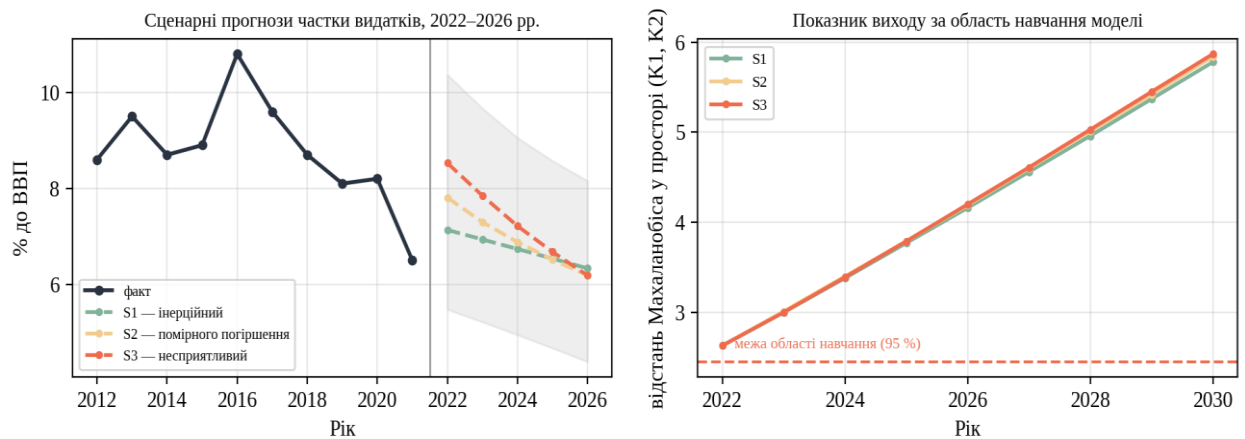


Рисунок 4.6 – Сценарні прогнози частки видатків і показник виходу за область навчання моделі

Для формального оцінювання придатності макрорівневої моделі на прогнозованому горизонті введено показник віддаленості сценарної точки від області, на якій модель навчалася. В якості міри використано відстань Махаланобіса у просторі головних компонент:

$$d_M = \sqrt{(k_t - \bar{k})^T S^{-1} (k_t - \bar{k})} \quad (4.12)$$

де $k_t = (K_{1t}, K_{2t})^T$ – вектор-стовпець значень двох головних компонент у році t ; $\bar{k}$ – вектор середніх значень цих компонент на навчальній вибірці; S^{-1} – їх обернена коваріаційна матриця. За двовимірного нормального наближення критичне значення рівня 0,95 становить 2,45.

Розрахунки показують, що вже у 2022 р. сценарні значення компонент лежать на межі області навчання ($d_M = 2,63$), а починаючи з 2024 р. – суттєво поза нею ($d_M \geq 3,38$), досягаючи 5,78 у 2030 р. Причина є структурною: головні компоненти побудовані на показниках із вираженим монотонним трендом, тому будь-яка їх екстраполяція неминуче виводить систему за межі спостереженої області. Для порівняння, діапазон значень першої компоненти на навчальній вибірці становить від $-4,91$ до $3,91$, тоді як сценарні значення на 2030 р. перевищують 7.

Практичним підтвердженням цього діагнозу є втрата впорядкованості сценаріїв. Якщо у 2022–2024 рр. прогнози розташовані у змістовно очікуваному порядку ($S1 < S2 < S3$), то починаючи з 2025 р. цей порядок порушується, а у 2026 р. несприятливий сценарій дає нижчу оцінку, ніж інерційний. Інверсія не має економічного тлумачення й повністю пояснюється лінійною екстраполяцією за межі області навчання у поєднанні з властивістю головних компонент, зафіксованою у розроблених сценаріях, що передбачає зміну лише чисельності одержувачів пенсій по інвалідності, не може бути подана у просторі компонент ізольовано, оскільки ця ознака входить до K_2 разом з іншими з протилежним знаком.

Отже, показник d_M дає кількісний критерій, за яким визначається максимальний горизонт довіри до макrorівневої моделі. За наявних даних цей горизонт становить два роки, після чого модель підлягає обов'язковому переоцінюванню за оновленою вибіркою. Це формалізує вимогу адаптивного оновлення прогнозів, сформульовану у даному дослідженні, і водночас, визначає регламент роботи аналітичної підсистеми: показник d_M обчислюється автоматично для кожного прогнозного зрізу й слугує сигналом про необхідність перенавчання моделі. Зведені результати прогнозування, найкращих моделей на рівнях управлінських рішень, до яких придатні отримані прогнози, наведено у табл. 4.17. Зіставлення рядків таблиці виявляє закономірність, що має безпосереднє значення для побудови системи підтримки прийняття рішень: точність прогнозу визначається не класом застосованої моделі, а інформаційною достатністю вихідного набору даних щодо конкретної задачі. Задача прогнозування контингенту розв'язується з похибкою менше 1 %, оскільки набір містить повний і внутрішньо узгоджений опис відповідного процесу. Задача прогнозування частки видатків розв'язується з похибкою 7–8 %, оскільки набір не містить показників, що безпосередньо формують обсяг видатків: середнього розміру пенсії, чисельності платників єдиного внеску, фонду оплати праці та обсягу ВВП. Наявні демографічні показники є лише опосередкованими індикаторами цих чинників.

Таблиця 4.17 – Зведені результати експериментальних досліджень

Задача	Найкраща модель	Показник точності	Придатність до рівня рішення
Прогноз сукупного контингенту, $h = 1$	Holt (лінійний тренд)	MAPE = 0,690 %; MASE = 0,493	Оперативні та тактичні рішення
Прогноз сукупного контингенту, $h = 3$	Holt (лінійний тренд)	MAPE = 1,584 %; MASE = 1,108	Тактичні рішення
Прогноз контингенту за видами пенсій	Holt за категоріями з узгодженням	MAPE = 0,334– 1,819 %	Тактичні рішення, планування структури видатків
Прогноз частки видатків у ВВП, $h = 1$	Регресія на K_1 , K_2 , $CPI(-1)$	MAPE = 8,264 %; MASE = 0,634	Індикативне планування, стратегічні рішення
Те саме, комбінований прогноз	Комбінація регресії та бустингу	MAPE = 7,512 %; MSE = 0,687	Індикативне планування
Сценарні розрахунки, $h \geq 3$	Не рекомендовано без переоцінювання	$dM > 2,45$	Лише як орієнтир діапазону

Цей результат обґрунтовує вимогу до інформаційного забезпечення аналітичної підсистеми: набір даних для макрорівневих задач має формуватися шляхом інтеграції демографічної статистики з показниками персоніфікованого обліку та бюджетної звітності ПФУ, що технічно реалізується засобами сховища даних Інтегрованої комплексної інформаційної системи. Побудовані моделі придатні для індикативного планування й оцінювання діапазонів, тоді як розрахунок конкретних бюджетних призначень потребує розширення факторного простору.

4.3 Інформаційна технологія прогнозування контингенту одержувачів пенсій за віком в умовах невизначеності

Оскільки зростання кількості одержувачів пенсій є одним із ключових факторів, що впливають на збільшення видатків на державне пенсійне забезпечення, питання прогнозування контингенту одержувачів пенсій є однією з ключових задач підтримки прийняття рішень в системі державного пенсійного

забезпечення. В роботі запропонована методика прогнозування контингенту одержувачів пенсій на основі даних персоніфікованого обліку. Оскільки дана інформація конфіденційна, були згенеровані набори даних, що відповідають наявним в базі даних персоніфікованого обліку застрахованих осіб. Було враховано, що часові ряди даних застрахованих осіб містять пропуски даних про періоди роботи, які не можна заповнювати без наявності підтвердженої документально інформації про стаж. Досліджувалась динаміка найбільш чисельної категорії одержувачів пенсій за віком.

Для опрацювання пропусків даних про стаж виконано частотний аналіз (рис.4.7), значення розподілу страхового стажу за весь період трудової діяльності застрахованої особи, є комбінацією розподілів значення страхового стажу до та після 2004 року.

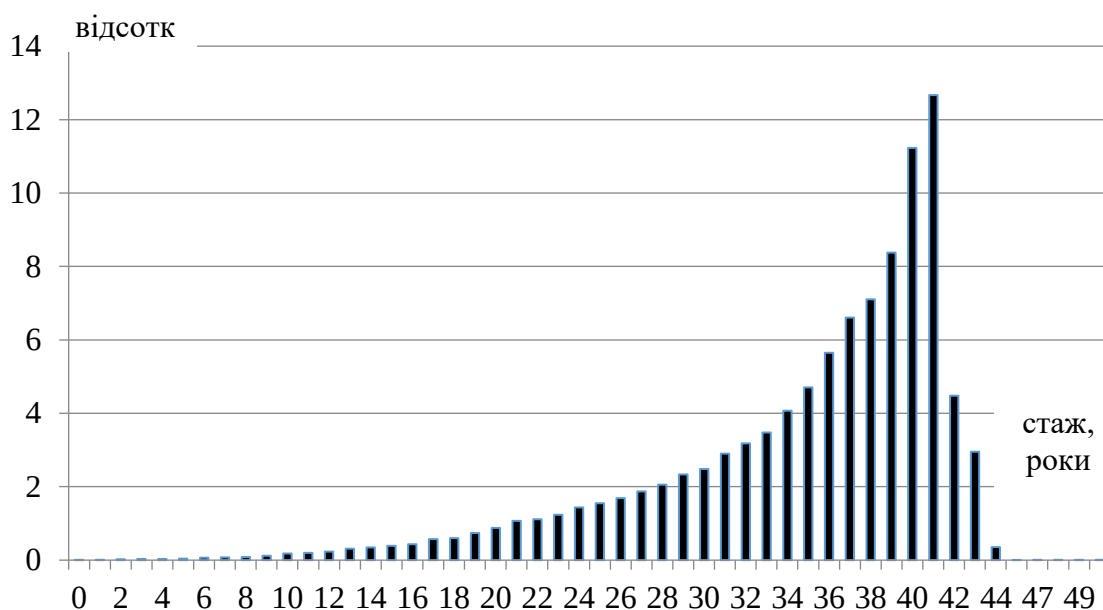


Рисунок 4.7 - Розподіл значень страхового стажу застрахованих осіб, що народилися у 1964 році, за весь період їх трудової діяльності

Якщо, наприклад, роком настання права на пенсію є 2024 рік, наявні пропуски даних опрацьовано виходячи з припущення про те, що 90-98% працездатного населення працювали у період до 1990 року, а за період з 2004 по 2024 рр. можуть «докупити» не більше п'яти років страхового стажу до необхідного для призначення їм пенсії за віком у 2024 році.

На рис. 4.14 представлено припущення щодо розподілу страхового стажу застрахованих осіб, яким у 2024 році виповнюється 60 років (рік народження – 1964).

Розподіл значень страхового стажу за весь період трудової діяльності застрахованої особи, описується статистичними характеристиками: математичне очікування $\mu = 34,81$, стандартне відхилення $\sigma = 7,08$, до 2004 року - математичне очікування $\mu = 18,67$, стандартне відхилення $\sigma = 4,09$, після 2004 року - математичне очікування $\mu = 15,66$, стандартне відхилення $\sigma = 4,67$

Для моделювання було використано математичний апарат гібридних мереж Байєса (рис. 4.8, табл. 4.18).



Рисунок 4.8 - Топологія побудованої мережі Байєса

Вершина «Ймовірність отримання пенсії» характеризується випадковою змінною, умовний розподіл якої залежить від значень батьківських вершин. В простішому випадку цей розподіл нормальний, а в складнішому випадку, аналогічно вершині "Ймовірність отримати пенсію", може використовуватись змішаний нормальний розподіл. Тобто, для кожного з розподілів має бути заданий список його параметрів:

- математичні очікування та дисперсії для розподілів, що описують ступінь впливу дискретних батьків;
- вагові коефіцієнти, що враховують ступінь впливу на математичне очікування неперервних батьків.

Таблиця 4.18 - Налаштування вершин побудованої гібридної мережі Байєса

Назва вершини	Значення, що описують вершину
Страховий стаж за весь період трудової діяльності	$34,33 \pm 7,08$.
Страховий стаж до 2004 року	$18,67 \pm 4,09$
Страховий стаж після 2004 року	$15,66 \pm 4,67$
Ймовірність придбання 1-5 років страхового стажу	$12,85 \pm 2,1$
Вік особи, роки	45-65
Пільговий стаж (шкідливі умови)	«так» - наявний пільговий стаж 6,5 років та більше: ймовірність – 2%, «ні» - якщо менше 6,5 років – 98%
Ознака права на дострокову пенсію або зменшення пенсійного віку	«так» - ймовірність 2,5%, «ні» - ймовірність – 97,5%
Ймовірність отримання пенсії	Так/ні

Таблиця 4.19 - Результати моделювання кількості осіб 1964-1968 рр. народження, що вийдуть на пенсію за віком за наявності різного страхового стажу, враховано, що стаж, якого не вистачає (але не більше 5 років), можна докупити

Рік народження	Кількість застрахованих осіб (наявна інформація про страховий стаж)	в тому числі, від загальної кількості застрахованих осіб, %:			
		пенсія за віком буде призначена з настанням пенсійного віку	пенсія буде призначена з настанням пенсійного віку, за умови придбання стажу	пенсія буде призначена пізніше (у 63 років)	пенсія буде призначена пізніше (65 років)
1964	480964	77,77	10,44	6,41	3,21
1965	462280	75,67	11	6,75	3,95
1966	485298	72,01	12,26	7,78	4,7
1967	489583	68,79	13,48	8,63	5,25
1968	500903	65,06	15,44	9,42	5,87

Отже, як показали розрахунки (табл. 4.19), якщо особі, що планує виходити на пенсію при досягненні передбаченого законом віку (60, 63, 65 років), не

вистачає 5 і менше років стажу, до необхідного страхового стажу для виходу на пенсію після досягнення передбаченого законом віку, вона може укласти договір з Пенсійним фондом України про добровільну участь у системі пенсійного страхування і «докупити» стаж. Ймовірність такої події, якщо «докупляти»: 1 рік - 91,98%, 2 роки - 90,59%, 3 роки - 55,19%, 4 роки - 23,35%, 5 років - 12,85%. Тоді, найбільш ймовірним сценарієм для 2024 року є наступний: особи 1964 року народження набудуть права на пенсію за віком у віці 60 років і за наявного страхового стажу 31 рік. Таких осіб може бути 374046 (77,7%). Крім того, одержувати пенсію у 60 років зможуть ще 50213 осіб (10,44%), якщо зможуть (будуть мати бажання) «докупити» страховий стаж. Тобто, всього 424259 (88,14%) осіб набудуть права на пенсію за віком у 60 років. У 63 роки, за наявності страхового стажу від 21 до 31 року, з числа осіб 1964 року народження набудуть право на пенсію 30829 (6,41%) осіб. Право на пенсію у 65 років, маючи страховий стаж від 15 років набуде 15438 (3,21%) осіб. Співставлення результатів прогнозування із наявними статистичними даними підтвердило обґрунтованість запропонованої методики.

Висновки до розділу 4

У роботі проведено експериментальне дослідження запропонованої методології прогнозування процесів у системі державного пенсійного забезпечення на макро- та мікрорівнях. Встановлено наявність структурної зміни у 2014-2015 рр., статистично підтвердженої критерієм Чоу ($F=265,30$, $p < 10^{-15}$). Для моделювання видатків Пенсійного фонду застосовано метод головних компонент; дві компоненти пояснюють 81,78% варіабельності даних.

Комплексне оцінювання якості моделі за профілем Profile(M) за однакової вагомості складових дало значення 0,83, при цьому підтверджено необхідність аналізу як інтегральної оцінки, так і окремих її характеристик. Для прогнозування структури контингенту встановлено накопичення розбіжності між схемами “знизу вгору” та “згори вниз” від 6,1 тис. осіб у 2022 р. до 65,2 тис.

осіб у 2030 р., що обґрунтовує використання процедури вертикального узгодження.

Сценарне прогнозування показало, що до 2030 р. розкид між крайніми сценаріями становить 724,9 тис. осіб (8,25 %), а пенсійне навантаження може змінюватися від 232,6 до 262,2 особи на 1000 населення. За результатами ретропрогнозування найкращою за використаною системою метрик виявилась регресійна модель на двох головних компонентах із лагованим ІСЦ ($MAE=0,684$, $MAPE=8,264\%$, $MASE=0,634$); комбінування моделей забезпечило $MAE=0,618$ та $MAPE=7,512\%$.

Контроль області застосовності за відстанню Махаланобіса показав вихід сценарних значень за межі області навчання починаючи з 2024 р. ($d_M > 3,38$ при критичному значенні 2,45), що визначає необхідність обережного використання довгострокової екстраполяції та періодичного оновлення моделей.

На мікрорівні із застосуванням гібридної мережі Байєса здійснено прогнозування набуття права на пенсію за індивідуальними характеристиками. Для когорти 1964 року народження 77,7 % осіб можуть набути право на пенсію у 60 років за наявного стажу, а ще 10,44 % – за умови набуття додаткового стажу. Отримані результати підтверджують працездатність запропонованої методології для прогнозування та підтримки прийняття рішень на різних рівнях системи державного пенсійного забезпечення.

ВИСНОВКИ

На основі виконаних теоретичних та експериментальних досліджень вирішено важливу науково-прикладну проблему в галузі інформаційних технологій – підвищення ефективності прийняття рішень в управлінні розвитком процесів різної природи, що мають місце у системі державного пенсійного забезпечення методами та засобами сучасних інформаційних технологій. При цьому отримано такі нові результати:

1. Виконано аналіз існуючої системи державного пенсійного забезпечення, визначено її місце та роль у соціальній сфері України, виявлено, що дана система є складною ієрархічною, а процеси, що характеризують її розвиток, переважно нелінійні, їх перебіг відбувається в умовах невизначеності під впливом соціально-економічних, демографічних, екологічних та суспільно-політичних чинників, що забезпечило адекватність побудованих моделей.

2. Розвинуто системну методологію оброблення різнорідних та неповних даних, отриманих з різних джерел, врахувавши особливості опрацювання відкритої інформації та інформації з обмеженим доступом, методику їх використання для прогнозування процесів різної природи, характерних для системи державного пенсійного забезпечення в умовах невизначеності, що забезпечило якість побудованих прогнозів.

3. Розроблено метод прогнозування контингенту одержувачів пенсій, який враховуючи ієрархічність процесу прийняття рішень у системі державного пенсійного забезпечення, забезпечує формування наборів даних для аналізу та деталізувати прогнози залежно від рівня управління, використовуючи поєднання ймовірнісно-статистичних методів та інтелектуального аналізу даних, дозволяє опрацьовувати часові ряди із значною кількістю пропусків, є універсальним для розв'язання задач опису та прогнозування розвитку системи державного пенсійного забезпечення та її складових в умовах інформаційної та статистичної невизначеностей.

4. Розроблено системну методологію побудови прогнозів, основу якої

складають розроблені моделі зокрема, математичні та їх ансамблі, алгоритми машинного навчання, сценарний аналіз, методи опрацювання різномірної інформації, що забезпечило якість, комплексність, стійкість, узгодженість результатів прогнозування на різних рівнях управління та етапах прийняття рішень, контроль виконання та оцінювання результативності рішень, забезпечуючи зворотній зв'язок у системі обслуговування населення, що дозволяє підвищити ефективність функціонування системи державного пенсійного забезпечення в умовах невизначеності, зокрема, спричиненої військовим конфліктом.

5. Розвинуто методику адаптивного прогнозування розвитку системи державного пенсійного забезпечення України, за рахунок розширення набору моделей та методів аналізу даних, додавання методики оновлення оцінок за надходження нових даних і коректного поєднання статистичної інформації з експертними судженнями, призначену для використання на різних рівнях управління, що забезпечило узгодженість та якість рішень, що приймаються в системі ПФУ як складової соціальної сфери України.

6. Удосконалено систему критеріїв оцінювання прогнозів та виконано апробацію запропонованих підходів, встановлено, що їх використання забезпечує адекватний опис причинно-наслідкових зв'язків соціально-економічних, демографічних, екологічних, суспільно-політичних груп чинників та визначення можливих варіантів розвитку досліджуваних процесів, що дозволило підвищити точність прогнозування на 10-15%, а використання ймовірнісних моделей, штучного інтелекту та комбінованих прогнозів - зменшити похибки при прогнозуванні до майже 20% у порівнянні з існуючими підходами.

7. Виконано аналіз адекватності побудованих моделей та коректності формування висновків на їх основі шляхом проведення емпіричних досліджень, підтверджено можливість їх застосування для розв'язання задач оцінювання і прогнозування розвитку системи державного пенсійного забезпечення як складової соціальної сфери України.

8. Удосконалено інформаційну технологію прогнозування динаміки контингенту одержувачів пенсій та показників бюджету ПФУ, яка удосконалює наявне рішення шляхом додавання розроблених прогнозних моделей, технологій збору, оброблення, аналізу різнорідних даних, у якій особливу увагу приділено опрацюванню пропусків та аномалій, візуалізації результатів, розмежування доступу до інформаційних ресурсів яка відрізняється гнучкістю,, здатністю до масштабування, що забезпечує її подальше використання як у Пенсійному фонді України, так і у системах підтримки прийняття рішень органів державного управління та місцевого самоврядування.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Зарудний О. Б., Коваль Р. Г. Методика прогнозування часових рядів для визначення потреби у видатках на соціальний захист та соціальне забезпечення», *Міжнародний науково-технічний журнал «Проблеми керування та інформатики»*. 2024. Vol. 69, №о. 5. С. 64–77. <https://doi.org/10.34229/1028-0979-2024-5-5>
2. Zarudnyi, O., Koval, R. Application of probabilistic and stochastic models and data mining for forecasting the contingent of old age pension recipients in the context of systemic uncertainty. *Technology Audit and Production Reserves*, 2024. Vol. 5. №. 2(79), P. 56–62. <https://doi.org/10.15587/2706-5448.2024.313960>
3. Зарудний О. Застосування інтелектуального аналізу еколого-економічних даних у системі пенсійного страхування та для прогнозування витрат на соціальний захист і соціальне забезпечення. *Екологічна безпека та природокористування*, 2024. Т. 51(3), С. 161–176. UR: <https://es-journal.in.ua/article/view/314099>. <https://doi.org/10.32347/2411-4049.2024.3.161-176>
4. Трофимчук О., Коваль Р., Зарудний О. Методика прогнозування кількості інвалідів з числа санітарних втрат. *Екологічна безпека та природокористування*, 2025. Т. 54 (2). С. 121–135. <https://doi.org/10.32347/2411-4049.2025.2.121-135>
5. Zarudnyi, O., Koval, R. Methodology for Constructing an Analytical Subsystem of the Unified Information System of the Social Sphere of Ukraine. *Cybernetics and Systems Analysis*. 2025. Vol. 61. №. 4. P. 487–494. ISSN 1573-8337. <https://doi.org/10.1007/s10559-025-00785-9>
6. Korbicz J., Sholokhov O., Koval R., Zarudnyi O. Application of SAS Text Miner for the analysis of citizens' appeals in the system of social protection and social security. 8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society. (AISTDS 2024). (Kyiv, 1 Oct. 2024). *Proceedings of the 8th International Scientific and Practical Conference "Applied Information Systems and Technologies in the Digital Society*. 2024. Vol. 3942, P. 113 – 124. ISSN 1613-0073 URL: https://ceur-ws.org/Vol-3942/S_09_Korbicz.pdf,

7. Корбич Ю., Шолохов О., Зарудний О., Коваль Р. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб населення у соціальному захисті та соціальному забезпеченні. *Прикладні інформаційні системи та технології в цифровому суспільстві*. 2024. Т. 1 № 1. С. 172-184. <https://doi.org/10.17721/3041-2323.2024.172-184>
8. Зарудний О.Б., Коваль Р. Г. Проблеми прогнозування факторів, що впливають на обсяг видатків Пенсійного фонду України Інформаційно-комунікаційні технології для перемоги та відновлення // Колективна монографія за матеріалами XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток» (Київ, 14-15 листопада 2023 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2023. С. 76-79. URL: https://itgip.org/wp-content/uploads/2023/11/1_zbirka_08_11_23-1-1.pdf
9. Зарудний О.Б., Коваль Р. Г. Застосування методу сингулярного розкладання у задачі аналізу електронних звернень громадян до Пенсійного фонду України. Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення // Колективна монографія за матеріалами XXIII Міжнародної науково-практичної конференції «Математичне моделювання та інформаційно-комунікаційні технології для зміцнення та відновлення» (Київ, 12-13 листопада 2024 р.) / За заг. ред. С.О. Довгого. К.: ТОВ «Видавництво «Юстон», 2024. С. 156-159. URL: https://itgip.org/wp-content/uploads/2024/11/2024-11-24_zbirka_all_07_11_2024_148x210.pdf
10. Зарудний О. Б., Коваль Р. Г. Шолохов О. В. Методика застосування кластеризації текстів на основі лінгвістичних правил для дослідження потреб населення у соціальному захисті та соціальному забезпеченні Прикладні інформаційні системи та технології в цифровому суспільстві: зб. Тез доповідей і наук. повідомлень учасників VIII Міжнародної науково-практичної конференції «Прикладні інформаційні системи та технології в цифровому суспільстві» (Київ, 01 жовтня 2024 р.) // за заг. ред. В. Плєскач, О. Фендьо. К.: Київський нац.

ун-т ім. Тараса Шевченка, 2024. С. 167-178. URL: https://aistis.knu.ua/wp-content/uploads/2024/10/Text_Book_confernce_01.10.2024.pdf

11. Зарудний О. Б., Колбун В. А., Малецький О. М., Машкін В. Г., Рябцева Т. Б., Матвєєв С. В., Хандрос Ю. О., Коваль Р. Г., Бондарчук О. М., Педан Ю. А., Павлюков І. О. Комп'ютерна програма «Інтегрована комплексна система Пенсійного фонду України «ІКІС ПФУ», версія 2» №11056; заявл.31.10.2011; опубл. 11.11.2011.

12. Закон України «Про загальнообов'язкове державне пенсійне страхування»: Науково-практичний коментар / Б.О. Зайчук, О.Б. Зарудний, С.Б. Березіна та ін. Київ.: НВП «АВТ», 2005. 1024 с. (Інтегрований навчально-атестаційний комплекс).URL: <https://essuir.sumdu.edu.ua/items/75829b1c-c646-4d79-8e96-5171c0e86c28>

13. Зайчук Б. О., Зарудний О.Б., Березіна С. Б. Пенсійний фонд України. Організаційно-правові та соціально-економічні засади функціонування. . Київ: НВП "АВП", 2006. 298 с.

14. Зайчук Б. О., Зарудний О.Б., Березіна С. Б. Загальнообов'язкове державне пенсійне страхування : навч. посіб. (Інтегрований навчально-атестаційний комплекс). Київ : НВП "АВП", 2005. 256 с.

15. Про пенсійне забезпечення громадян в СРСР. Закон СРСР від 17.04.91. № 2121-І (v2121400-91). URL: <https://zakon.rada.gov.ua/laws/show/v1480400-90#Text>

16. Офіційний сайт. Пенсійний фонд України. URL: <https://www.pfu.gov.ua/pro-pfu/istoriya-stanovlennya/>

17. Про створення Пенсійного фонду України. Постанова Кабінету Міністрів України від 28.01.1992 № 39. URL: <https://zakon.rada.gov.ua/laws/show/39-92-п>

18. Питання Пенсійного фонду України. Постанова Кабінету Міністрів України №о 534 від 16.09.92 <https://zakon.rada.gov.ua/laws/show/534-92-п#Text>

19. Про загальнообов'язкове державне пенсійне страхування. Закон України від від 24.06.2004. № 1878-IV. URL: <https://zakon.rada.gov.ua/laws/show/1058-15#Text>

20. Про недержавне пенсійне забезпечення. Закон України від 15.12.2005, № 3201-IV. <https://zakon.rada.gov.ua/laws/show/1057-15#Text>

21. Про збір та облік єдиного внеску на загальнообов'язкове державне соціальне страхування. Закон України від 08.07.2010 № 2464-VI. URL: <https://zakon.rada.gov.ua/laws/show/2464-17>
22. Про затвердження Положення про Пенсійний фонд України. Постанова Кабінету Міністрів України від 23.07.2014 № 280. URL: <https://zakon.rada.gov.ua/laws/show/280-2014-п>
23. Про затвердження регламенту роботи інформаційно-телекомунікаційної системи Пенсійного фонду України. Постанова Правління Пенсійного фонду України від 11.10.2021 28-1. URL: <https://www.pfu.gov.ua/content/uploads/2021/10/Postanova-28-1-vid-11-zhovtnya-2021-.pdf>
24. Трембіта. Система електронної взаємодії державних електронних інформаційних ресурсів. URL: <https://trembita.gov.ua/ua>
25. Про затвердження Положення про Єдину інформаційну систему соціальної сфери. Постанова Кабінету Міністрів України від 14.04.2021 № 404. URL: <https://zakon.rada.gov.ua/laws/show/404-2021-п>
26. Про внесення змін до Положення про Єдину інформаційну систему соціальної сфери. Постанова Кабінету Міністрів України від 27.10.2023 р. № 1130. <https://zakon.rada.gov.ua/laws/show/1130-2023-п#Text>
27. Про затвердження Порядку проведення експериментального проекту з реалізації функціоналів першої черги Єдиної інформаційної системи соціальної сфери. Постанова Правління Пенсійного фонду України від 03.03.2021 № 8-1. URL: <https://zakon.rada.gov.ua/laws/show/za400-21#Text>
28. Смуш-Кулеша М. Федорова А., Мойса Б. Соціальні права в Україні під час війни. Звіт про оцінку потреб. Рада Європи. 2022, 64 с. URL: <https://rm.coe.int/needs-assessment-ua-2/1680a9b408>
29. Шаповалова Т. Поняття і зміст соціального захисту та соціального забезпечення населення в сучасній Україні. *Економічний аналіз*. 2022. Т. 32. № 3. С. 123-130. <https://doi.org/10.35774/econa2022.03.123>
30. Predictive Modeling Applications in Actuarial Science: Volume 1, Predictive Modeling Techniques (International Series on Actuarial Science/Ed. Edward W. Frees,

- G. Meyers. Cambridge: University Press., 2014. 1567 p.
<https://doi.org/10.1017/CBO9781139342674>
31. Simandl M., Lesek M., Straka O. Pension fund model design and state estimation. *IFAC world congress*. Preprint of the The 16th IFAC Symposium on Analysis, Design and Evaluation of Human – Machine Systems. (Prague. 3 Jul. - 8 Jul 2005). Prague: IFAC 2005. P. 1-6, URL: https://kky-sw.zcu.cz/en/publications/SimandlM_2005_Pensionfundmodel
 32. Iyer S. Actuarial mathematics of social security pensions. Geneva: International Labour Organization. 1999. 142 p. URL: https://www.issa.int/sites/default/files/documents/publications/2Actuarial_mathematics_of_ss_pensions_en-29172.pdf
 33. Statistical models and techniques used in actuarial analysis. URL: <https://fastercapital.com/topics/statistical-models-and-techniques-used-in-actuarial-analysis.html>
 34. Frees E. W. *Regression Modeling with Actuarial and Financial Applications*. New York: Cambridge University Press. 2010. 584 p. URL: <https://users.ssc.wisc.edu/~ewfrees/ActuarialRegression2009/>
 35. Gupta R. Y., Mudigonda S. S., Baruah P. K., Kandala P. K. Implementation of Correlation and Regression Models for Health Insurance Fraud in Covid-19 Environment using Actuarial and Data Science Techniques. *International Journal of Recent Technology and Engineering (IJRTE)*. 2020. Vol. 9, Issue 3. P. 699-706. <https://doi.org/10.35940/ijrte.C4686.099320>
 36. Oduwole O. A., Jenyo I.A., Onamade A .A., Adegbite, O. Adegoke B.O An improved design and implementation of a pension contribution monitoring system. *Journal of Computing and Informatics*. 2021. Vol. 2, Issue 1, P. 89-98
 37. Black E., Lattyak C. G., Chairperson V, Stone.L. K. Senior Pension Fellow Modeling—for Pension Actuaries. Washington: The American Academy of Actuaries. 2023. URL: https://www.actuary.org/sites/default/files/2023-01/Modeling_Practice_Note.pdf
 38. Social Protection Systems /Ed. E. Schuring, M. Loewe. Elgar Publishing. 2021. 776 p. <https://doi.org/10.4337/9781839109119>

39. World Bank. 2010. Modeling Pension Reform: The World Bank's Pension Reform Options Simulation Toolkit. World Bank Pension Reform Primer Series, World Bank PROST Model. URL: <http://hdl.handle.net/10986/11074>
40. Новітні тренди розвитку страхового ринку: синергія інновацій та фінансової безпеки. Матеріали VII Міжнародної науково-практичної інтернет-конференції, м. Київ, 15 травня 2025 р. Київ, КНЕУ, 2025. 280 с. URL: https://kneu.edu.ua/userfiles/Credit_Economics_Department/2024kbss/ZBD086D0A0NIK_mater__VII__MD196zhnar__konferen__15_05_25_.pdf
41. Бреднюк В. І., Джоші О. І. Економіко-математичне моделювання динаміки пенсійного забезпечення в Україні. *Наукові записки Національного університету "Острозька академія". Серія "Економіка"*. 2013. Вип. 22. С. 133–136. URL: <https://eprints.oa.edu.ua/id/eprint/2146>
42. Kolawole F. L., Festus A. F. Management information systems and effective business decisions: an evaluation of pension fund administrators in Nigeri. *Global Research Journal of Business Management*. 2022. Vol. 2, No. 2: P. 22 – 37.
43. Pradana F. R., Pertiwi A. Analysis and Design Management Information System Pension Benefit Feature of Core System Dana Pensiun Lembaga Keuangan. *International Research Journal of Advanced Engineering and Science*. 2022. Vol. 7, Issue 1, P. 239-243.
44. Antari P., Bufra F. S., Andesti C. L. Analysis of pension fund payment data processing information system design. *Journal of International Accounting, Taxation and Information Systems*, 2024. №1(3), P. 144-149
45. D’Arcangelis A. M., Levantesi S., Rotundo G. A complex networks approach to pension funds. *Journal of Business Research*. 2021, Vol. 129, P. 687-702 <https://doi.org/10.1016/j.jbusres.2019.10.071>
46. Iyer S. Stochastic Actuarial Modelling of a Defined-Benefit Social Security Pension Scheme: An Analytical Approach. *Annals of Actuarial Science*. 2008. Vol. 3(1-2), P. 127-185. <https://doi.org/10.1017/S174849950000049X>
47. González Parra G., Arenas A. J. A mathematical model for social security systems with dynamical systems. *Ingeniería Y Ciencia*, 2014. Vol. 10(19), P. 33–53. <https://doi.org/10.17230/ingciencia.10.19.2>

48. Lähderanta T., Salonen J., Möttönen J., Sillanpää M.J. Modelling old-age retirement: An adaptive multi-outcome LAD-lasso regression approach. *International Social Security Review*. 2022. Vol. 75, Issue 1. 3-29. <https://doi.org/10.1111/issr.12287>
49. Larochelle J.-P., Carlson P., Cote V. C., Lu Y., Shapiro N., Tam A., Thusu V., Zhang A. *Predictive Analytics and Machine Learning. Practical Applications for Actuarial Modeling (Nested Stochastic)*. Schaumburg: Society of Actuaries. Research Institute. URL: <https://www.soa.org/49ae74/globalassets/assets/files/resources/research-report/2023/predictive-analytics-and-machine-learning.pdf>
50. Tomar D. J. Information technology assessment of the pension funds. Technical Assistance for Policy Reform II BearingPoint. Egypt: Dokki, Giza, 2006. 59 c. URL: https://pdf.usaid.gov/pdf_docs/PA00WC4R.pdf
51. Anusiuba O. I. A., Ezuruka E. O., Ekwealor O. U., Karim U. Development of an Enhanced Online Pension Management System *Journal of Scientific Engineering and Applied Science*. 2021. Vol.7, Issue 3. P 62-91
52. Pradana F. R., Pertiwi A. Analysis and design management information system pension benefit feature of core system Dana Pensiun Lembaga Keuangan *International Research Journal of Advanced Engineering and Science*,, 2022. Vol. 7, Issue 1, P. 239–243. URL: <https://irjaes.com/wpcontent/uploads/2022/03/IRJAES-V7N1P281Y22.pdf>.
53. Математичні методи і моделі в управлінні економічними процесами : монографія / Л. М. Малярець, Є. Ю. Місюра, В. В. Койбічук та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2016. 420 с. URL: <https://essuir.sumdu.edu.ua/server/api/core/bitstreams/d333ff2a-1afd-4da0-8ea9-da4fe59c346a/content>
54. Державне підприємство «Інформаційно-обчислювальний центр Міністерства соціальної політики, сім'ї та єдності України». URL: <https://www.ioc.gov.ua>
55. ТОВ "НВП "Медирент". Офіційний сайт. URL: <https://medirent.ua>

56. Prado R., West M. AR and TVAR Time Series Modelling and Decompositions Stationary and Nonstationary/Time-Frequency Analysis URL: <https://www2.stat.duke.edu/~mw/mwsoftware/TVAR/index.html>
57. Huerta G., West M. Priors and component structures in autoregressive time series models. *Journal of the Royal Statistical Society: Series B*. 1999. Vol. 61, no. 4. P. 881–899. <https://doi.org/10.1111/1467-9868.00208>
58. JMulTi URL: <http://www.jmulti.org>
59. Orocos Bayesian Filtering Library. URL: <https://orocos.org/bfl.html>
60. Hoeting J. A., Madigan D., Raftery A. E., Volinsky C. T. Bayesian Model Averaging: A Tutorial. *Statistical Science*. 1999. Vol. 14, No. 4. P. 382–417. <https://doi.org/10.1214/ss/1009212519>
61. OpenBUGS. URL: <https://www.medschl.cam.ac.uk/institutions/mrc-bsu>
62. Lunn D., Spiegelhalter D., Thomas A., Best N. The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*. 2009. Vol. 28, № 25. P 3049–3067. <https://doi.org/10.1002/sim.3680>
63. Team R Core. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. URL: <https://www.r-project.org/>
64. Python, URL: <https://www.python.org>
65. The pandas development team. pandas documentation. URL: <https://pandas.pydata.org/docs/>
66. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011. Vol. 12. P. 2825–2830. URL: <https://www.jmlr.org/papers/v12/pedregosa11a.html>
67. Seabold S., Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*. 2010. P. 92–96. <https://doi.org/10.25080/Majora-92bf1922-011>
68. Find the information that matters using natural language processing (NLP). URL: https://www.sas.com/ru_ua/software/visual-text-analytics.html
69. Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by M. W. Berry. Springer, 2003. 261 p.
70. SAS Institute. URL: <https://www.sas.com>

71. Згуровський М. З., Панкратова Н. Д. Основи системного аналізу. Київ : Видавнича група BHV, 2007. 544 с.
72. Інститут економіки та прогнозування НАН України. Офіційний сайт. URL: <https://ief.org.ua/publication/monohrafii>
73. Trofymchuk O. M, Bidiuk P. I., Prosiiankina-Zharova T. I., Terentiev O. M. Decision support systems for modelling, forecasting and risk estimation: monography. Riga : LAP LAMBERT Academic Publishing. 2019. 176 p.
74. Lupenko S.A., Stoliar M.V., Terentiev O.M., Savastiyanov V.V. Automated semantic ontology construction for foresight studies using large language models. *System Research and Information Technologies*. 2026. Vol. 2. P. 134-152. <https://doi.org/10.20535/SRIT.2308-8893.2026.2.09>
75. Довгий С. О., Бідюк П. І., Трофимчук О. М., Савенков О. І. Методи прогнозування в системах підтримки прийняття рішень. Київ: Азимут-Україна. 2011. 608 с.
76. Trofymchuk O., Bidiuk P., Terentiev O., Klymenko V. The methodology for adaptive modeling and forecasting nonlinear and nonstationary processes. *International Scientific Technical Journal «Problems of control and informatics»*. 2024. Vol. 69 No. 1. С. 63-79. <https://doi.org/10.34229/1028-0979-2024-1-6>
77. Варенко В. М., Братусь І. В., Дорошенко В. С., Смольников Ю. Б., Юрченко В. О. Системний аналіз інформаційних процесів: Навч. посіб. Київ: Університет «Україна», 2013. 203 с. URL: https://duikt.edu.ua/uploads/l_18_98931046.pdf
78. Glass K. Real-time Macroeconomic Data and Uncertainty. DEP (Socioeconomics). Discussion Papers. *Macroeconomics and Finance Series*. 2014. Vol. 6. P. 1–29. URL: <https://www.econstor.eu/bitstream/10419/120854/1/832072516.pdf>
79. Trukhan S. V., Bidiuk P. I. Methods of building mathematical models of actuarial processes. *Eastern-European Journal of Enterprise Technologies*, 2015. Vol. 1. № (4 (73)), P. 27–35. <https://doi.org/10.15587/1729-4061.2015.36486>
80. Gembarski, P. C., Plappert, S., Lachmayer, R. Making design decisions under uncertainties: probabilistic reasoning and robust product design. *Journal of Intelligent*

Information Systems, 2021. 57 (3), 563–581. <https://doi.org/10.1007/s10844-021-00665-6>

81. Пенсійний фонд України. Офіційний вебпортал: статистика та звітність. URL: <https://www.pfu.gov.ua/>

82. Державна служба статистики України, Офіційний сайт. <https://www.ukrstat.gov.ua>

83. Національний банк України. Статистика. URL: <https://bank.gov.ua/ua/statistic>

84. Пенсійний фонд України. Офіційний вебпортал: електронних послуг. URL: <https://www.pfu.gov.ua/>

85. Портал відкритих даних. URL: <https://data.gov.ua>

86. Гуськова В. Г., Бідюк П. І., Гасанов А. С. Ймовірісно-статистичні методи моделювання і прогнозування. Київ: Вид-во НПУ імені М. П. Драгоманова, 2022. 429 с. URL: <http://enpuir.npu.edu.ua/handle/123456789/41164>

87. Sichko T., Neskorodieva T., Rymar P., Methods and models of decision-making in uncertain conditions. *Computer systems and information technologies*, 2020. Vol. 3, P. 47-51. <https://doi.org/10.31891/CSIT-2021-5-6>

88. Терентьєв О. М. Моделі, методи та інформаційні технології прогнозування нелінійних нестационарних процесів в умовах невизначеності дис. ... д-ра тех. наук: 05.13.06/ Інститут телекомунікацій і глобального інформаційного простору НАН України. 2021. 469 с.

89. Згуровський М. З, Бідюк П. І., Терентьєв О. М., Просянкін-Жарова Т. І. Байєсівські мережі у системах підтримки прийняття рішень. Київ: Едельвейс, 2015. 300 с.

90. Koller D., Friedman N. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009. URL: <https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/>

91. Bates J. M., Granger C. W. J. The Combination of Forecasts. *Operational Research Quarterly*. 1969. Vol. 20, no. 4. P. 451–468. DOI: <https://doi.org/10.1057/jors.1969.103>

92. Wang X., Hyndman R. J., Li F., Kang Y. Forecast combinations: An over 50-year review. *International Journal of Forecasting*. 2023. Vol. 39, №. 4. P. 1518–1547. <https://doi.org/10.1016/j.ijforecast.2022.11.005>
93. Hyndman R. J., Athanasopoulos G. *Forecasting: Principles and Practice*. 3rd ed. Melbourne: OTexts, 2021. URL: <https://otexts.com/fpp3/>
94. Коваль Р. Г. Інформаційні технології збору та оброблення даних для прогнозування витрат на пенсії і соціальні допомоги : дис. ... канд. техн. наук : 05.13.06. Інститут телекомунікацій і глобального інформаційного простору НАН України. Київ, 2025.
95. Про захист персональних даних. Закон України від 01.06.2010 № 2297-VI. URL: <https://zakon.rada.gov.ua/laws/show/2297-17>
96. Kimball R., Caserta J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. Wiley, 2004. URL: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-etl-toolkit/>
97. NIST/SEMATECH. e-Handbook of Statistical Methods. URL: <https://www.itl.nist.gov/div898/handbook/>
98. Little R. J. A., Rubin D. B. *Statistical Analysis with Missing Data*. Wiley, 2019. <https://doi.org/10.1002/9781119482260>
99. Jolliffe I. T., Cadima J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*. 2016. Vol. 374. Article 20150202. DOI: <https://doi.org/10.1098/rsta.2015.0202>
100. Міністерство соціальної політики, сім'ї та єдності України. Офіційний вебсайт. URL: <https://www.msp.gov.ua/>
101. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. <https://doi.org/10.1007/978-0-387-84858-7>
102. Czernicki D. How cloud computing transforms actuarial modeling infrastructure. Retrieved from: https://www.ey.com/en_us/insights/insurance/cloud-computing-implications-for-actuarial-modeling

103. Постанова Кабінету Міністрів України від 16 грудня 2004 р. № 1677 «Методика проведення актуарних розрахунків у системі загальнообов’язкового державного пенсійного страхування». URL: <https://www.kmu.gov.ua/npas/10301286>
104. Pensions Schemes and Projection Models in EU-25 Member States Occasional Papers, 2007. № 35, P. 377. URL: https://ec.europa.eu/economy_finance/publications/pages/publication10173_en.pdf
105. Sze M. (1993). The Process of Pension Forecasting. *Journal of actuarial practice*, 1993. Vol. 1 (1), P. 31–50. URL: <https://core.ac.uk/download/pdf/127444889.pdf>
106. Jia T., Wang C., Tian Z., Wang B., Tian F. Design of Digital and Intelligent Financial Decision Support System Based on Artificial Intelligence. *Computational Intelligence and Neuroscience*. 2022, ID 1962937, P. 1–7. <https://doi.org/10.1155/2022/1962937>
107. Kimball R., Caserta J. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley, 2004. URL: <https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/books/data-warehouse-dw-etl-toolkit/>
108. O’Brien R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*. 2007. Vol. 41. P. 673–690. <https://doi.org/10.1007/s11135-006-9018-6>
109. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey. *ACM Computing Surveys*. 2009. Vol. 41, No. 3. Article 15. <https://doi.org/10.1145/1541880.1541882>
110. Dempster A. P., Laird N. M., Rubin D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B*. 1977. Vol. 39, no. 1. P. 1–38. DOI: <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>
111. Diebold F. X., Mariano R. S. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*. 1995. Vol. 13, No. 3. P. 253–263. <https://doi.org/10.1080/07350015.1995.10524599>

112. Harvey D., Leybourne S., Newbold P. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*. 1997. Vol. 13, No. 2. P. 281–291. [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
113. Tsay R.S. Analysis of financial time series. Hoboken : Wiley & Sons, Inc., 2010. 707 p. <https://doi.org/10.1002/9780470644560>
114. Athanasopoulos G., Hyndman R. J., Kourentzes N., Panagiotelis A. Forecast reconciliation: A review. *International Journal of Forecasting*. 2024. Vol. 40, no. 2. P. 430–456. DOI: <https://doi.org/10.1016/j.ijforecast.2023.10.010>
115. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*. 2003. Vol. 3. P. 1157–1182. URL: <https://www.jmlr.org/papers/v3/guyon03a.html>
116. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B*. 1996. Vol. 58, No. 1. P. 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
117. McCrea R., King R., Graham L., Börger L. Realising the promise of large data and complex models. *Methods in Ecology and Evolution*. 2023. Vol.14, Issue 1, P. 4–11. <https://doi.org/10.1111/2041-210X.14050>
118. Breiman L. Random Forests. *Machine Learning*. 2001. Vol. 45. P. 5–32. <https://doi.org/10.1023/A:1010933404324>
119. Rossi, P.E., G. M. Allenby, R. McCulloch, Bayesian Statistics and Marketing. Hoboken: John Wiley & Sons, Ltd, 2005. 368 p. <http://doi.org/10.1002/0470863692>
120. Бідюк П. І., Терентьєв О. М., Коновалюк М. М. Байєсівські мережі в технологіях інтелектуального аналізу даних. Штучний інтелект. 2010. № 2. С. 104–113. URL: <https://nasplib.isoftware.kiev.ua/server/api/core/bitstreams/3f5f21ee-7bb6-4165-b0bf-84fd2858d4fe/content>
121. Hautaniemi S. K., Korpisaari P. T., Saarinen J. P. P. Target identification with dynamic hybrid Bayesian networks. *Target identification with dynamic hybrid Bayesian networks*, Processing for Remote Sensing VI conference Image and Signal, Barcelona, Spain, January 2001. Vol. 4170. <https://doi.org/10.1117/12.413885>
122. Press, S.J., Subjective and Objective Bayesian Statistics: Principles, Models, Applications. Hoboken: John Wiley & Sons, Inc., 2003.

123. Gustafsson O. Some contributions to heteroscedastic time series analysis and computational aspects of Bayesian VARs. dissertation for the Degree of Doctor of Philosophy in Statistics. Stockholm, 2020. 48 p. URL: <https://www.diva-portal.org/smash/get/diva2:1497177/FULLTEXT01.pdf>
124. Rabbi M. F. A Dynamic Systems Approach to Integrated Sustainability: Synthesizing Theory and Modeling Through the Synergistic Resilience Framework. *Sustainability* 2025, Vol. 17, Issue 11. 4878. <https://doi.org/10.3390/su17114878>
125. Takagi T., Sugeno M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*. 1985. Vol. SMC-15, No. 1. P. 116–132. <https://doi.org/10.1109/TSMC.1985.6313399>
126. Wang, J., Yu, C., Zhang, J. Constructing the regional intelligent economic decision support system based on fuzzy C-mean clustering algorithm. *Soft Computing*, 2019. Vol. 24 (11), P. 7989–7997. <https://doi.org/10.1007/s00500-019-04091-3>
127. Бідюк, П. І., Терентьєв О. М., Просянкін-Жарова Т. І. Прикладна статистика. Вінниця : ПП "ТД"Едельвейс і К", 2013. 304 с
128. Методи машинного навчання у задачах системного аналізу і прийняття рішень : монографія / В. Є. Стрілець, С. І. Шматков, М. Л. Угрюмов та ін. Харків : Харківський національний університет імені В. Н. Каразіна, 2020. 160 с. URL: https://old.karazin.ua/images/redactor/news/2020-12-28/Shmatkov_.pdf
129. Salman H. A., Kalakech A., Steiti A., Trans. Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, 2024, P. 69-79. <https://doi.org/10.58496/BJML/2024/007>
130. Friedman J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. 2001. Vol. 29, No. 5. P. 1189–1232. <https://doi.org/10.1214/aos/1013203451>
131. Gradient Boosting: теорія та пакети. URL: <https://www.kaggle.com/code/nadiia789/gradient-boosting>
132. Oracle Business Intelligence (Oracle BI). URL: <https://www.oracle.com/ua/business-analytics/business-intelligence/technologies/bi.html>
133. Genie. URL: <https://download.bayesfusion.com/files.html?category=Academia>

134. Mamdani E. H., Assilian S. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man-Machine Studies*. 1975. Vol. 7, No. 1. P. 1–13. [https://doi.org/10.1016/S0020-7373\(75\)80002-2](https://doi.org/10.1016/S0020-7373(75)80002-2)
135. Zgurovsky M. Z., Zaychenko Y. P. The Fundamentals of Computational Intelligence: System Approach. Switzerland: Springer International Publishing, 2016. 375 p
136. Грімнак О. В., Катренко А. В. Система підтримання прийняття рішень для багатокрокових процесів з використанням ланцюгів Маркова. *Вісник Національного університету «Львівська політехніка» «Інформаційні системи та мережі»*. 2008. № 631 (3). С. 148-155. URL: <https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-631-2008/systema-pidtrymannya-pryynyattya-rishen-dlya-bagatokrokovykh>
137. Gallagher M., Sturgeon S., Finch B., Villongco F.. Probabilistic Analysis of Complex Combat Scenarios. *Military Operations Research*. 2022. Vol. 27, No. 1. P. 87–106. URL: <https://www.jstor.org/stable/27116757>.
138. Кучеренко С. К., Колосов А. М., Барилов А. І. Методологія сценарного планування післявоєнного відновлення України. *European Journal of Management Issues*, 2025. № 33(1). С. 57-65. <https://doi.org/10.15421/192506>.
139. Phoong S.W.; Phoong S.Y.; Phoong, K.H. Analysis of structural changes in financial datasets using the breakpoint test and the Markov Switching Model. *Symmetry* 2020, Vol. 12. Issue 3. P. 401. <https://doi.org/10.3390/sym12030401>
140. Pratiwi H. , Susanti Y., Sulistijowati H. Sri A Robust Regression by Using Huber Estimator and Tukey Bisquare Estimator for Predicting Availability of Corn in Karanganyar Regency, *Indonesia Indonesian Journal of Applied Statistics* 2018. Vol. I No.1 URL: <https://jurnal.uns.ac.id/ijas/article/view/24090/17210>
141. Christodoulou-Volos, C., Tserkezos, D. Sensitivity of the Ramsey’s Regression Specification Error Term Test on the Degree of Nonlinearity of the Functional Form. *Journal of Applied Economic Sciences*, 2023. Vol. XVIII, 1(79): P. 5 – 10. [https://doi.org/10.57017/jaes.v18.1\(79\).01](https://doi.org/10.57017/jaes.v18.1(79).01)
142. Bai J., Perron P. Estimating and testing linear models with multiple structural changes. 1998 Vol. 66, No. 1. P. 47-78. <https://doi.org/10.2307/2998540>

143. Liu J., Wu S., Zidek J. V. On segmented multivariate regression. *Statistica Sinica*. 1997. Vol. 7, No. 2. P. 497-525. <https://www.jstor.org/stable/24306090>
144. Huang L, Zhou X, Shi L, Gong L. Time Series Feature Selection Method Based on Mutual Information. *Applied Sciences*. 2024; Vol.14. Issue 5: P. 1-14. <https://doi.org/10.3390/app14051960>
145. Draper N., Smith G. *Applied regression analysis*. New York: John Wiley & Sons, Inc., 1986. 366 p.
146. Gneiting T., Raftery A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*. 2007. Vol. 102, no. 477. P. 359–378. <https://doi.org/10.1198/016214506000001437>
147. Guo C., Pleiss G., Sun Y., Weinberger K. Q. On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning*. 2017. Vol. 70. P. 1321–1330. URL: <https://proceedings.mlr.press/v70/guo17a.html>
148. Shapovalenko, N. A Suite of Models for CPI Forecasting. *Visnyk of the National Bank of Ukraine*, 2021. Vol. 252, P. 4-36. <https://doi.org/10.26531/vnbu2021.252.01>
149. De Gooijer J.G. *Elements of nonlinear time series analysis and forecasting*. Cham : Springer, 2017. 626 p.
150. Castle J.L., Doornik J.A., Hendry D.F. Modelling non-stationary «Big Data». *International Journal of Forecasting*. 2021. Vol. 37, N. 4. P. 1556–1575. DOI: <https://doi.org/10.1016/j.ijforecast.2020.08.002>
151. Wang X., Hyndman R. J., Li F., Kang Y. Forecast combinations: An over 50-year review *International Journal of Forecasting*. 2023. Vol. 39, No. 4. P. 1518–1547. <https://doi.org/10.1016/j.ijforecast.2022.11.005>
152. Borisenko O.A. *System theory. The Information approach*. Sumy: Sumy State University, 2010, 210 p.

УКРАЇНА



ДЕРЖАВНА СЛУЖБА ІНТЕЛЕКТУАЛЬНОЇ ВЛАСНОСТІ УКРАЇНИ

РІШЕННЯ

про реєстрацію договору, який стосується права автора на твір

Державна служба інтелектуальної власності розглянула заяву
Товариство з обмеженою відповідальністю "Науково-виробниче підприємство
"Медирент", вул. Щекавицька, 9 А, м. Київ, 04071
(повне ім'я фізичної або повне офіційне найменування юридичної особи, адреса)

про реєстрацію авторського договору від 31.10.2011, № 11056 про передачу (відчуження)
майнових прав і прийняла рішення зареєструвати авторський договір, відповідно до якого
майнові права на твір

Комп'ютерна програма "Інтегрована комплексна інформаційна система
Пенсійного фонду України "ІКІС ПФУ", версія 2"; Зарудний Олексій Борисович,
Колбуи Віктор Андрійович, Малецький Олександр Миколайович,
Машкін Владислав Геннадійович, Рябцева Тетяна Борисівна, Матвєєв Сергій
Валентинович, Хандрос Юрій Олександрович, Коваль Роман Григорович,
Бондарчук Олексій Миколайович, Педан Юрій Анатолійович, Павлюков Іван
Олексійович

(вид, повна, скорочена (за наявності) назва твору, повне ім'я, псевдонім (за наявності) автора(ів))

передаються(відчужуються)

Приватне акціонерне товариство "АТ.ЛАС", вул. Щорса, 31, м. Київ, 01133
(повне ім'я фізичної(их) або повне офіційне найменування юридичної(их) особи(осіб), яка(ї) передає(ють)(відчужує(ють)) право
на твір, адреса)

Товариство з обмеженою відповідальністю "Науково-виробниче підприємство
"Медирент", вул. Щекавицька, 9 А, м. Київ, 04071
(повне ім'я фізичної або повне офіційне найменування юридичної особи, якій передається (відчужується) право на твір, адреса)

Повністю

Реєстраційний номер 1672
Дата реєстрації 11.11.2011

Голова Державної служби інтелектуальної власності


М.В.Паладій





МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДЕРЖАВНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

вул. Кіото, 19, м. Київ, 02156, тел. +380 (44) 531 47 41, e-mail: knute@knute.edu.ua, код ЄДРПОУ 44470624

11.08.2026 № 2493/01

На № _____

Директору інституту телекомунікацій
та глобального інформаційного
простору
Олександр ТРОФИМЧУКУ

Шановний пане Олександр!

Результати, отримані в ході виконання дисертаційного дослідження Зарудного О.Б. на тему «Моделі та методи прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення» використовуються у навчальному та науковому процесі у Державному торговельно-економічному університеті.

З повагою

Ректор



Анатолій МАЗАРАКІ

ТОВ "Науково-виробниче підприємство "МЕДИРЕНТ"
вул.Казимира Малевича буд. 86Б, м. Київ, 03150, Україна
+38 (044) 200 54 10, +38 (044) 200 54 11
office@medirent.com.ua
www.medirent.com.ua



Research and development enterprise "MEDIRENT", LLC
86B, Kazimir Malevich st, 03150, Kyiv, Ukraine
+38 (044) 200 54 10, +38 (044) 200 54 11
office@medirent.com.ua
www.medirent.com.ua

Вих. №140 від 17 червня 2025 року

Директору інституту
телекомунікацій та глобального
інформаційного простору
Трофимчуку О.М.

Шановний Олександрє Миколайовичу!

Результати, отримані в ході виконання дисертаційного дослідження на тему «Моделі та методи прогнозування для підтримки прийняття рішень у системі державного пенсійного забезпечення» впроваджені в процесі створення та модернізації Інтегрованої комплексної інформаційної системи Пенсійного фонду (ІКІС ПФ), яка розробляється групою компаній «Медирент» на замовлення Пенсійного фонду України, починаючи із 2002 року по цей час.

Директор

ТОВ „НВП <МЕДИРЕНТ>”



Олександр Колесник

ДОДАТОК В

В.1. Модуль конфігурації

Призначення. Задає шляхи до каталогів, періоди оцінювання, склад факторного простору та позначення ознак. Модуль не виконує обчислень і підключається всіма іншими модулями.

Лістинг В.1 – config.py

1) Заголовок модуля, опис і підключення бібліотек

```
# -*- coding: utf-8 -*-
"""
```

Спільні налаштування для експериментів розділу 4 дисертації.
Усі шляхи задаються відносно каталогу, у якому лежать скрипти, тому пакет можна переносити цілком.

```
"""
```

```
from pathlib import Path
```

2) Каталоги проекту

```
ROOT = Path(__file__).resolve().parent
DATA = ROOT / 'data'           # сюди покласти 1.xlsx
OUT = ROOT / 'out'             # проміжні результати (csv, json)
FIG = ROOT / 'fig'             # рисунки у форматі png
for p in (DATA, OUT, FIG):
    p.mkdir(exist_ok=True)
XLSX = DATA / '1.xlsx'
```

3) Періоди оцінювання (табл. 4.6 «Протокол обчислювальних експериментів»)

```
YEAR_CONTINGENT = (1995, 2021) # моделі сукупного контингенту
YEAR_MACRO = (2002, 2021)      # макрорівнева модель видатків
YEAR_STRUCT = (2015, 2021)     # моделі структури контингенту
(після структурного розриву)
BREAK_YEAR = 2014              # рік структурного розриву
HORIZON = 9                    # 2022-2030 pp.
FORECAST_YEARS = list(range(2022, 2022 + HORIZON))
```

```
RANDOM_STATE = 42
```

4) Факторний простір

демографічні та структурні ознаки, що подаються методу головних компонент

```
DEM = ['pens_tot', 'pens_age', 'pens_dis', 'pens_surv',
       'pens_serv', 'pens_soc',
       'pop_tot', 'share_urb', 'le_all', 'le_gap',
       'support_ratio']
```

позначення ознак у тексті розділу

```
LABELS = {
    'pens_tot': 'X1 чисельність пенсіонерів',
    'pens_age': 'X2 пенсіонери за віком',
    'pens_dis': 'X3 пенсіонери по інвалідності',
    'pens_surv': 'X4 пенсії у зв'язку з втратою годувальника',
    'pens_serv': 'X5 пенсії за вислугу років',
    'pens_soc': 'X6 соціальні пенсії',
    'pop_tot': 'X7 чисельність населення',
    'share_urb': 'X8 частка міського населення',
    'le_all': 'X9 очікувана тривалість життя',
    'le_gap': 'X10 розрив тривалості життя (ж - ч)',
    'support_ratio': 'X11 коефіцієнт підтримки',
    'cpi': 'X12 індекс споживчих цін',
```

```

        'cpi_l1': 'X12(-1) індекс споживчих цін із лагом 1 рік',
    }
CAT = ['pens_age', 'pens_dis', 'pens_surv', 'pens_serv',
       'pens_soc', 'pens_judge']
CAT_LABELS = {
    'pens_age': 'за віком',
    'pens_dis': 'по інвалідності',
    'pens_surv': 'у зв'язку з втратою годувальника',
    'pens_serv': 'за вислугу років',
    'pens_soc': 'соціальні',
    'pens_judge': 'довічне утримання суддів',
}

```

В.2. Крок 0. Підготовки вихідного набору даних

Вхідні дані: data/1.xlsx. Результат: out/clean.csv.

Лістинг В.2 – 00_prep.py

1) Заголовок модуля, опис і підключення бібліотек

```

# -*- coding: utf-8 -*-
"""

```

Крок 0. Підготовки вихідного набору даних (підрозділ 4.1.1–4.1.4).

Виконує:

- * зчитування 1.xlsx і присвоєння машинозчитуваних імен стовпцям;
- * усунення шести типів дефектів (табл. 4.2);
- * обчислення похідних показників (підрозділ 4.1.4);
- * контроль узгодженості показника «пенсіонерів на 1000

населення».

Результат: out/clean.csv

```

"""

```

```

import numpy as np
import pandas as pd
from config import XLSX, OUT

```

2) Опис структури вихідного файлу

```

COLS = ['year', 'pop_tot', 'pop_urb', 'pop_rur', 'le_all', 'le_m',
        'le_f',
        'soc_gdp', 'cpi', 'pens_tot', 'pens_age', 'pens_dis',
        'pens_surv',
        'pens_serv', 'pens_soc', 'pens_judge', 'pens_per1000',
        '_empty',
        'infl_times']

```

3) Числове приведення значень

```

def to_number(value):

```

```

    """Числове приведення з урахуванням коми як десяткового
    роздільника.

```

Дефект 2 (табл. 4.2): очікувану тривалість життя чоловіків і жінок

за 2018 р. збережено як текст «66,69» і «76,72».

```

    """

```

```

    if pd.isna(value):
        return np.nan
    if isinstance(value, str):
        value = value.replace(',', '.').replace(' ',
        '').replace('\xa0', '')
    try:

```

```

        return float(value)
    except ValueError:
        return np.nan
4) Основна процедура підготовки даних
def main():
5) Зчитування файлу та присвоєння імен стовпцям
    raw = pd.read_excel(XLSX, header=None)
    df = raw.iloc[1:].copy()          # перший рядок - заголовки
    українською
    df.columns = COLS
    # Дефект 5: стовпець «інфляція, разів» заповнено для 4 років
із 33
    # і дублює індекс споживчих цін в іншому масштабі - вилючаємо.
    df = df.drop(columns=['_empty', 'infl_times'])
    for c in df.columns:
        df[c] = df[c].apply(to_number)
    # Дефект 1: помилка кодування ключа - рік 2002 записано як
20021.
    df.loc[df.year > 3000, 'year'] = 2002
    df['year'] = df['year'].astype(int)
    df = df.sort_values('year').reset_index(drop=True)
6) Похідні показники (підрозділ 4.1.4)
    df['pens_per1000_calc'] = df.pens_tot / df.pop_tot * 1000
    df['share_urb'] = df.pop_urb / df.pop_tot * 100
    df['support_ratio'] = df.pop_tot / df.pens_tot          # осіб
населення на 1 пенсіонера
    df['pens_age_share'] = df.pens_age / df.pens_tot * 100
    df['pens_dis_share'] = df.pens_dis / df.pens_tot * 100
    df['le_gap'] = df.le_f - df.le_m
    df['cpi_ll'] = df.cpi.shift(1)                          # лаг
інфляційного чинника
7) Звіт про дефекти
    print('Рядків: %d, період %d-%d' % (len(df), df.year.min(),
df.year.max()))
    print('\nПропуски за показниками (дефекти 3 і 4):')
    gaps = df.isna().sum()
    print(gaps[gaps > 0].to_string())
    chk = df[['year', 'pens_per1000',
'pens_per1000_calc']].dropna()
    print('\nКонтроль показника «на 1000 населення»: max
|відхилення| = %.2e'
        % (chk.pens_per1000 -
chk.pens_per1000_calc).abs().max())
    # Дефект 6: порушення зіставності бази обліку за 2014 р.
    b = df.loc[df.year.isin([2013, 2014, 2015]),
'pens_per1000_calc'].tolist()
    print('\nПоказник В (осіб на 1000) за 2013-2015 pp.: %.1f,
%.1f, %.1f' % tuple(b))
    print('    немонотонний стрибок 2014 р. - артефакт обліку, а не
динаміка')
    df.to_csv(OUT / 'clean.csv', index=False)
    print('\nЗбережено:', OUT / 'clean.csv')
if __name__ == '__main__':

```

```

    main()
В.3. Крок 1. Моделі прогнозування контингенту
Вхідні дані: out/clean.csv. Результат: out/e1.json, out/e1_cv.csv, out/e1_forecast.csv.
Лістинг В.3 – 01_contingent.py
1) Заголовок модуля, опис і підключення бібліотек
# -*- coding: utf-8 -*-
"""
Крок 1. Експеримент 1 – моделі часових рядів для сукупного
контингенту
одержувачів пенсій (підрозділи 4.1.3 і 4.2.2).
Виконує:
    * перевірку структурного розриву 2014 р. (критерій Чоу, регресія
з утрюванням);
    * побудову восьми моделей-кандидатів;
    * крос-валідацію на ковзному вікні за двома наборами початків
прогнозу
    (2010-2020 рр. і 2015-2020 рр.) – табл. 4.7;
    * оцінювання адекватності на навчальній вибірці – табл. 4.8;
    * прогноз на 2022-2030 рр. з інтервальними оцінками – табл. 4.9.
Результати: out/e1.json, out/e1_cv.csv, out/e1_forecast.csv
"""
import json
import warnings
import numpy as np
import pandas as pd
import statsmodels.api as sm
from scipy import stats
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.holtwinters import Holt
from config import OUT, YEAR_CONTINGENT, BREAK_YEAR, HORIZON
warnings.filterwarnings('ignore')
MODELS = ['naive', 'drift', 'drift3', 'trend', 'holt',
          'arima110', 'arima011', 'arimax']
MODEL_LABELS = {
    'naive': 'Наївний («без змін）」,
    'drift': 'Наївний із дрейфом',
    'drift3': 'Дрейф за останні 3 роки',
    'trend': 'Лінійний тренд',
    'holt': 'Holt (лінійний тренд)',
    'arima110': 'ARIMA(1,1,0) з дрейфом',
    'arima011': 'ARIMA(0,1,1) з дрейфом',
    'arimax': 'ARIMAX(1,1,0) з утрюванням',
}
}
MIN_TRAIN = 16                # перший зріз крос-валідації: навчання на
1995-2010 рр.
H_MAX = 3                    # горизонти оцінювання точності
2) Моделі-кандидати
def fit_predict(y_train, d_train, h, kind):
    """Навчання моделі на y_train і прогноз на h кроків уперед.
    d_train – значення індикатора утрювання D на навчальній
    вибірці
    (потрібне лише для моделі arimax).
    """

```

```

n = len(y_train)
t = np.arange(n)
if kind == 'naive':
    return np.repeat(y_train[-1], h)
if kind == 'drift':
    slope = (y_train[-1] - y_train[0]) / (n - 1)
    return y_train[-1] + slope * np.arange(1, h + 1)
if kind == 'drift3':
    slope = (y_train[-1] - y_train[-4]) / 3.0
    return y_train[-1] + slope * np.arange(1, h + 1)
if kind == 'trend':
    res = sm.OLS(y_train, sm.add_constant(t)).fit()
    future = sm.add_constant(np.arange(n, n + h),
has_constant='add')
    return res.predict(future)
if kind == 'holt':
    return Holt(y_train,
initialization_method='estimated').fit().forecast(h)
if kind == 'arima110':
    return ARIMA(y_train, order=(1, 1, 0),
trend='t').fit().forecast(h)
if kind == 'arima011':
    return ARIMA(y_train, order=(0, 1, 1),
trend='t').fit().forecast(h)
if kind == 'arimax':
    res = ARIMA(y_train, order=(1, 1, 0),
                exog=d_train.reshape(-1, 1), trend='t').fit()
    return res.forecast(h, exog=np.ones((h, 1)) * d_train[-1])
raise ValueError('невідома модель: %s' % kind)

```

3) Структурний розрив

```

def chow_test(y, k):
    """Критерій Чоу для точки розриву з індексом k."""
    n = len(y)
    t = np.arange(n)
    pooled = sm.OLS(y, sm.add_constant(t)).fit().ssr
    left = sm.OLS(y[:k], sm.add_constant(t[:k])).fit().ssr
    right = sm.OLS(y[k:], sm.add_constant(t[k:])).fit().ssr
    p_par = 2 # константа й кут
нахилю
    f = ((pooled - (left + right)) / p_par) / ((left + right) / (n
- 2 * p_par))
    return f, 1 - stats.f.cdf(f, p_par, n - 2 * p_par)

```

4) Метрики точності (формули (3.34)–(3.37))

```

def accuracy(actual, forecast, y_full):
    actual, forecast = np.asarray(actual), np.asarray(forecast)
    err = actual - forecast
    naive_mae = np.mean(np.abs(np.diff(y_full))) # знаменник
MASE
    return {
        'MAE': float(np.mean(np.abs(err))),
        'RMSE': float(np.sqrt(np.mean(err ** 2))),
        'MAPE': float(np.mean(np.abs(err / actual)) * 100),
        'MASE': float(np.mean(np.abs(err)) / naive_mae),
    }

```

```

    }
def rolling_cv(y, d, years, first_origin_year):
    """Крос-валідація на ковзному вікні з розширюваною навчальною
    вибіркою.
    Випадкове розбиття для часових рядів неприйнятне, оскільки
    допускає
    використання майбутньої інформації (підрозділ 3.5.4).
    """
    store = {m: {h: [] for h in range(1, H_MAX + 1)} for m in
    MODELS}
    for n in range(MIN_TRAIN, len(y)):
        if years[n - 1] < first_origin_year:
            continue
        h = min(H_MAX, len(y) - n)
        for m in MODELS:
            f = np.asarray(fit_predict(y[:n], d[:n], h, m),
dtype=float)
            for step in range(1, h + 1):
                store[m][step].append((y[n + step - 1], f[step -
1]))
    out = {}
    for m in MODELS:
        out[m] = {}
        for step in range(1, H_MAX + 1):
            a = [p[0] for p in store[m][step]]
            f = [p[1] for p in store[m][step]]
            out[m][step] = accuracy(a, f, y)
    return out
def main():
    df = pd.read_csv(OUT / 'clean.csv')
    lo, hi = YEAR_CONTINGENT
    d = df[(df.year >= lo) & (df.year <=
hi)].reset_index(drop=True)
    y = d.pens_tot.values.astype(float)
    years = d.year.values
    D = (years >= BREAK_YEAR).astype(float)      # індикатор
утручання
    print('Період: %d-%d, спостережень: %d' % (lo, hi, len(y)))
5) 1. Структурний розрив (підрозділ 4.1.3)
    k = int(np.where(years == BREAK_YEAR)[0][0])
    f_chow, p_chow = chow_test(y, k)
    print('\nКритерій Чоу, точка розриву %d: F = %.2f, p = %.3g'
        % (BREAK_YEAR, f_chow, p_chow))
    t = np.arange(len(y))
    X = sm.add_constant(np.column_stack([t, D, t * D]))
    inter = sm.OLS(y, X).fit()
    print('Регресія з утручанням (4.1): R2 = %.4f' %
inter.rsquared)
    print('    темп до розриву:    %.1f тис. осіб/рік' %
inter.params[1])
    print('    темп після:         %.1f тис. осіб/рік' %
(inter.params[1] + inter.params[3]))
    print('    p-значення:', np.round(inter.pvalues, 5))

```

```

    arimax = ARIMA(y, order=(1, 1, 0), exog=D.reshape(-1, 1),
trend='t').fit()
    print('\nARIMAX(1,1,0) з утрочанням:')
    print(arimax.summary().tables[1])
6) 2. Крос-валідація за двома наборами початків (табл. 4.7)
    cv_all = rolling_cv(y, D, years, 2010)
    cv_post = rolling_cv(y, D, years, 2015)
    rows = []
    for m in MODELS:
        rows.append([
            MODEL_LABELS[m],
            cv_all[m][1]['MAPE'], cv_all[m][1]['MASE'],
            cv_all[m][2]['MAPE'], cv_all[m][3]['MAPE'],
            cv_post[m][1]['MAPE'], cv_post[m][1]['MASE'],
            cv_post[m][2]['MAPE'], cv_post[m][3]['MAPE'],
        ])
    # у табл. 4.7 розділу подано всі стовпці, крім «MAPE h=2*»,
    # який використовується лише для побудови рис. 4.10
    tab47 = pd.DataFrame(rows, columns=[
        'Модель', 'MAPE h=1', 'MASE h=1', 'MAPE h=2', 'MAPE h=3',
        'MAPE h=1*', 'MASE h=1*', 'MAPE h=2*', 'MAPE h=3*'])
    print('\nТабл. 4.7 (* - початки прогнозу після структурного
розриву):')
    print(tab47.round(3).to_string(index=False))
    tab47.to_csv(OUT / 'el_cv.csv', index=False)
7) 3. Адекватність на навчальній вибірці (табл. 4.8)
    insample = {}
    n = len(y)
    fits = {
        'trend': sm.OLS(y, sm.add_constant(np.arange(n))).fit(),
        'holt': Holt(y, initialization_method='estimated').fit(),
        'arima110': ARIMA(y, order=(1, 1, 0), trend='t').fit(),
        'arimax': arimax,
    }
    for name, res in fits.items():
        fitted = np.asarray(res.fittedvalues)
        err = y - fitted
        insample[name] = {
            'AIC': float(res.aic), 'BIC': float(res.bic),
            # перше спостереження вилучаємо: для моделей у
різницях воно неінформативне
            'RMSE': float(np.sqrt(np.mean(err[1:] ** 2))),
            'MAPE': float(np.mean(np.abs(err[1:] / y[1:])) * 100),
        }
    print('\nТабл. 4.8 - адекватність на навчальній вибірці:')
    print(pd.DataFrame(insample).T.round(3).to_string())
8) 4. Прогноз 2022–2030 рр. (табл. 4.9)
    fc = arimax.get_forecast(HORIZON, exog=np.ones((HORIZON, 1)))
    ci80 = fc.conf_int(alpha=0.20)
    ci95 = fc.conf_int(alpha=0.05)
    out = pd.DataFrame({
        'year': np.arange(2022, 2022 + HORIZON),
        'fc': fc.predicted_mean,

```

```

        'lo80': ci80[:, 0], 'hi80': ci80[:, 1],
        'lo95': ci95[:, 0], 'hi95': ci95[:, 1],
        'holt': Holt(y,
initialization_method='estimated').fit().forecast(HORIZON),
    })
    out['diff_pct'] = (out.holt / out.fc - 1) * 100
    print('\nТабл. 4.9 – прогноз контингенту, тис. осіб:')
    print(out.round(1).to_string(index=False))
    out.to_csv(OUT / 'el_forecast.csv', index=False)
    json.dump({
        'chow': {'F': float(f_chow), 'p': float(p_chow)},
        'intervention': {'params': inter.params.tolist(),
                        'pvalues': inter.pvalues.tolist(),
                        'R2': float(inter.rsquared)},
        'arimax_params': dict(zip(arimax.param_names, map(float,
arimax.params))),
        'arimax_pvalues': dict(zip(arimax.param_names, map(float,
arimax.pvalues))),
        'cv_all': cv_all, 'cv_post': cv_post, 'insample':
insample,
    }, open(OUT / 'el.json', 'w'), ensure_ascii=False, indent=1,
default=float)
    print('\nЗбережено: el.json, el_cv.csv, el_forecast.csv')
if __name__ == '__main__':
    main()

```

В.4. Крок 2. Факторний простір, макрорівнева модель і комбінування прогнозів

Вхідні дані: out/clean.csv. Результат: out/main.json.

Лістинг В.4 – 02_models.py

1) Заголовок модуля, опис і підключення бібліотек

```

# -*- coding: utf-8 -*-
"""

```

Крок 2. Експерименти 2-5 (підрозділи 4.1.5, 4.1.6, 4.2.3, 4.3.1-4.3.3).

Виконує:

- * кореляційний аналіз і діагностику мультиколінеарності – табл. 4.4;
- * метод головних компонент – табл. 4.5;
- * макрорівневу регресійну модель на головних компонентах – табл. 4.13, 4.14;
- * ансамблеві моделі машинного навчання та оцінки важливості – табл. 4.16;
- * крос-валідацію макрорівневих моделей – табл. 4.15;
- * комбінування прогнозів і критерій Дібольда – Маріано – табл. 4.17, 4.18;
- * прогнозування структури контингенту та вертикальне узгодження – табл. 4.10-4.12;
- * емпіричне покриття прогнозних інтервалів PICP.

Результат: out/main.json

```

"""

```

```

import json
import warnings

```

```

import numpy as np

```

```

import pandas as pd
import statsmodels.api as sm
from scipy import stats
from sklearn.decomposition import PCA
from sklearn.ensemble import GradientBoostingRegressor,
RandomForestRegressor
from sklearn.inspection import permutation_importance
from sklearn.preprocessing import StandardScaler
from statsmodels.stats.diagnostic import het_breuschpagan
from statsmodels.stats.stattools import durbin_watson, jarque_bera
from statsmodels.tsa.holtwinters import Holt
from config import (OUT, DEM, CATS, CAT_LABELS, YEAR_MACRO,
YEAR_STRUCT,
                    HORIZON, RANDOM_STATE)
warnings.filterwarnings('ignore')
MIN_TRAIN_MACRO = 12          # перший зріз крос-валідації
макромоделі
R = {}                        # накопичувач результатів
def make_rf():
    return RandomForestRegressor(n_estimators=500,
random_state=RANDOM_STATE)
def make_gb():
    return GradientBoostingRegressor(n_estimators=300,
learning_rate=0.05,
                                max_depth=2,
random_state=RANDOM_STATE)
def diebold_mariano(e1, e2, h=1):
    """Критерій Дібольда - Маріано за квадратичної функції втрат
(формула (3.33))."""
    d = e1 ** 2 - e2 ** 2
    n = len(d)
    dbar = d.mean()
    var = (np.sum((d - dbar) ** 2) / n) / n
    stat = dbar / np.sqrt(var)
    return float(stat), float(2 * (1 - stats.t.cdf(abs(stat), n -
1)))
def main():
    df = pd.read_csv(OUT / 'clean.csv')
    lo, hi = YEAR_MACRO
    d = df[(df.year >= lo) & (df.year <=
hi)].reset_index(drop=True)
    y = d.soc_gdp.values.astype(float)
    years = d.year.values
    scaler = StandardScaler().fit(d[DEM])
    Z = scaler.transform(d[DEM])
    pca = PCA().fit(Z)
    K = pca.transform(Z)
2) 4.1.4. Описова статистика (табл. 4.3)
    desc = d[DEM + ['cpi', 'soc_gdp']].describe().T[['mean',
'std', 'min', 'max']]
    desc['CV'] = desc['std'] / desc['mean'] * 100          #
коефіцієнт варіації, %
    R['desc'] = desc.round(3).to_dict()

```

```

print('Табл. 4.3 - описова статистика за %d-%d pp.: ' % (lo,
hi))
print(desc.round(3).to_string())
3) 4.1.5. Кореляції та мультиколінеарність (табл. 4.4)
cor = {}
for c in DEM + ['cpi', 'cpi_l1']:
    r, p = stats.pearsonr(d[c], y)
    cor[c] = [round(float(r), 4), round(float(p), 4)]
R['cor_y'] = cor
# коефіцієнт інфляції дисперсії (формула (3.3))
vif = []
for j in range(Z.shape[1]):
    others = np.delete(Z, j, axis=1)
    r2 = sm.OLS(Z[:, j],
sm.add_constant(others)).fit().rsquared
    vif.append(1 / (1 - r2) if r2 < 1 else np.inf)
R['vif'] = dict(zip(DEM, np.round(vif, 1).tolist()))
R['cond'] =
float(np.linalg.cond(np.column_stack([np.ones(len(Z)), Z])))
print('\nТабл. 4.4 - кореляції з Y та VIF:')
print(pd.DataFrame({'r': [cor[c][0] for c in DEM],
                    'p': [cor[c][1] for c in DEM],
                    'VIF': np.round(vif, 1)}),
index=DEM).to_string())
print('CPI поточний:', cor['cpi'], ' CPI з лагом:',
cor['cpi_l1'])
print('Число обумовленості матриці плану: %.1f' % R['cond'])
# повна регресія - для демонстрації нестійкості оцінок
full = sm.OLS(y, sm.add_constant(Z)).fit()
err_full = y - full.fittedvalues
R['full_ols'] = {'R2': float(full.rsquared), 'R2adj':
float(full.rsquared_adj),
                'Fp': float(full.f_pvalue),
                'MAPE_train': float(np.mean(np.abs(err_full /
y)) * 100)}
print('\nПовна регресія на 11 ознаках: R2 = %.4f, MAPE на
навчанні = %.2f %%'
      % (full.rsquared, R['full_ols']['MAPE_train']))
4) 4.1.6. Головні компоненти (табл. 4.5)
R['ev'] = pca.explained_variance_ratio_.tolist()
R['eig'] = pca.explained_variance_.tolist()
R['load'] = pd.DataFrame(pca.components_[:2].T, index=DEM,
                        columns=['K1',
'K2']).round(3).to_dict()
print('\nТабл. 4.5 - пояснена дисперсія, %:',
      np.round(np.array(R['ev']) * 100, 2)[:4])
print('накопичено двома компонентами: %.2f %%' %
(sum(R['ev'][:2]) * 100))
5) 4.3.1. Макрорівнева модель (формула (4.3), табл. 4.13, 4.14)
X = np.column_stack([K[:, :2], d.cpi_l1.values])
macro = sm.OLS(y, sm.add_constant(X)).fit()
resid = y - macro.fittedvalues
R['macro'] = {

```

```

        'params': macro.params.round(4).tolist(), 'bse':
macro.bse.round(4).tolist(),
        't': macro.tvalues.round(3).tolist(), 'p':
macro.pvalues.round(5).tolist(),
        'R2': float(macro.rsquared), 'R2adj':
float(macro.rsquared_adj),
        'AIC': float(macro.aic), 'BIC': float(macro.bic),
        'F': float(macro.fvalue), 'Fp': float(macro.f_pvalue),
        'RMSE': float(np.sqrt(np.mean(resid ** 2))),
        'MAE': float(np.mean(np.abs(resid))),
        'MAPE': float(np.mean(np.abs(resid / y)) * 100),
        'DW': float(durbin_watson(resid)),
        'JB': [float(v) for v in jarque_bera(resid)[:2]],
        'BP': [float(v) for v in het_breuschpagan(resid,
sm.add_constant(X))[:2]],
    }
    R['macro_fit'] = {'year': years.tolist(), 'actual':
y.tolist(),
                    'fit': np.round(macro.fittedvalues,
3).tolist(),
                    'resid': np.round(resid, 3).tolist()}
    print('\nМодель (4.3):', macro.params.round(4).tolist())
    print('R2 = %.4f, скориг. R2 = %.4f, MAPE = %.2f %%, DW =
%.3f'
          % (macro.rsquared, macro.rsquared_adj,
R['macro']['MAPE'], R['macro']['DW']))
    # та сама модель без інфляційного чинника - для оцінювання
Його внеску
    macro0 = sm.OLS(y, sm.add_constant(K[:, :2])).fit()
    e0 = y - macro0.fittedvalues
    R['macro_noCPI'] = {'R2': float(macro0.rsquared), 'R2adj':
float(macro0.rsquared_adj),
                      'AIC': float(macro0.aic),
                      'MAPE': float(np.mean(np.abs(e0 / y)) *
100)}

```

6) 4.3.2. Ансамблеві моделі (табл. 4.16)

```

names = DEM + ['cpi_l1']
X_ml = np.column_stack([Z, d.cpi_l1.values])
rf, gb = make_rf().fit(X_ml, y), make_gb().fit(X_ml, y)
R['ml_imp'] = {
    'rf': dict(zip(names, np.round(rf.feature_importances_,
4).tolist())),
    'gb': dict(zip(names, np.round(gb.feature_importances_,
4).tolist()))
}
perm = permutation_importance(rf, X_ml, y, n_repeats=50,
random_state=0)
R['ml_perm'] = dict(zip(names, np.round(perm.importances_mean,
4).tolist()))
R['ml_insample'] = {}
for tag, model in (('rf', rf), ('gb', gb)):
    e = y - model.predict(X_ml)
    R['ml_insample'][tag] = {

```

```

        'RMSE': float(np.sqrt(np.mean(e ** 2))),
        'MAPE': float(np.mean(np.abs(e / y)) * 100),
        'R2': float(1 - np.sum(e ** 2) / np.sum((y - y.mean())
** 2))),
    }
    print('\nВажливість ознак (RF / GB / перестановочна) для X5 -
пенсії за вислугу років:',
        R['ml_imp']['rf']['pens_serv'],
R['ml_imp']['gb']['pens_serv'],
        R['ml_perm']['pens_serv'])
7) Крос-валідація макрорівневих моделей (табл. 4.15)
    # Стандартизація і PCA перенавчаються на кожному зрізі, щоб
уникнути
    # витоку інформації з тестової частини.
    preds = {k: [] for k in ['ols_full', 'pca2', 'pca2cpi', 'rf',
'gb', 'naive']}
    test_years = []
    for n in range(MIN_TRAIN_MACRO, len(y)):
        sc = StandardScaler().fit(d[DEM].iloc[:n])
        Ztr, Zte = sc.transform(d[DEM].iloc[:n]),
sc.transform(d[DEM].iloc[n:n + 1])
        pc = PCA().fit(Ztr)
        Ktr, Kte = pc.transform(Ztr), pc.transform(Zte)
        ytr = y[:n]
        c_tr, c_te = d.cpi_11.values[:n], d.cpi_11.values[n:n + 1]
        preds['ols_full'].append(sm.OLS(ytr,
sm.add_constant(Ztr)).fit()
                                .predict(sm.add_constant(Zte,
has_constant='add'))[0])
        preds['pca2'].append(sm.OLS(ytr, sm.add_constant(Ktr[:,
:2])).fit()
                                .predict(sm.add_constant(Kte[:, :2],
has_constant='add'))[0])
        Xtr = np.column_stack([Ktr[:, :2], c_tr])
        Xte = np.column_stack([Kte[:, :2], c_te])
        preds['pca2cpi'].append(sm.OLS(ytr,
sm.add_constant(Xtr)).fit()
                                .predict(sm.add_constant(Xte,
has_constant='add'))[0])
        Mtr = np.column_stack([Ztr, c_tr])
        Mte = np.column_stack([Zte, c_te])
        preds['rf'].append(make_rf().fit(Mtr,
ytr).predict(Mte)[0])
        preds['gb'].append(make_gb().fit(Mtr,
ytr).predict(Mte)[0])
        preds['naive'].append(ytr[-1])
        test_years.append(int(years[n]))
    actual = np.array([y[list(years).index(t)] for t in
test_years])
    naive_mae = np.mean(np.abs(np.diff(y)))
    R['macro_cv'] = {}
    for tag, v in preds.items():
        e = actual - np.array(v)

```

```

R['macro_cv'][tag] = {
    'MAE': float(np.mean(np.abs(e))),
    'RMSE': float(np.sqrt(np.mean(e ** 2))),
    'MAPE': float(np.mean(np.abs(e / actual)) * 100),
    'MASE': float(np.mean(np.abs(e)) / naive_mae),
}
R['macro_cv_years'] = test_years
R['macro_cv_actual'] = np.round(actual, 3).tolist()
R['macro_cv_pred'] = {k: np.round(np.array(v), 3).tolist() for
k, v in preds.items()}
print('\nТабл. 4.15 - крос-валідація макрорівневих моделей (h
= 1):')
print(pd.DataFrame(R['macro_cv']).T.round(4).to_string())
8) 4.3.3. Комбінування прогнозів (табл. 4.17, 4.18)
e1 = actual - np.array(preds['pca2cpi'])
e2 = actual - np.array(preds['gb'])
s1, s2 = np.var(e1), np.var(e2)
s12 = np.mean((e1 - e1.mean()) * (e2 - e2.mean()))
w1 = float(np.clip((s2 - s12) / (s1 + s2 - 2 * s12), 0, 1))
# формула (3.29)
comb = w1 * np.array(preds['pca2cpi']) + (1 - w1) *
np.array(preds['gb'])
ec = actual - comb
R['comb'] = {
    'w1': w1, 'w2': 1 - w1,
    'var1': float(s1), 'var2': float(s2), 'cov': float(s12),
    'corr': float(np.corrcoef(e1, e2)[0, 1]),
    'MSE1': float(np.mean(e1 ** 2)), 'MSE2': float(np.mean(e2
** 2)),
    'MSEc': float(np.mean(ec ** 2)),
    'MSE_eq': float(np.mean((actual - 0.5 *
(np.array(preds['pca2cpi'])
+
np.array(preds['gb']))) ** 2)),
    'MAEc': float(np.mean(np.abs(ec))),
    'MAPEc': float(np.mean(np.abs(ec / actual)) * 100),
    'gain': float((1 - np.mean(ec ** 2) / min(np.mean(e1 **
2),
np.mean(e2 **
2))) * 100),
    'comb_pred': np.round(comb, 3).tolist(),
}
R['dm'] = {
    'pca_vs_gb': diebold_mariano(e1, e2),
    'pca_vs_comb': diebold_mariano(e1, ec),
    'gb_vs_comb': diebold_mariano(e2, ec),
    'pca_vs_naive': diebold_mariano(e1, actual -
np.array(preds['naive'])),
}
print('\nКомбінування: w1 = %.3f, виграш MSE = %.2f %%' % (w1,
R['comb']['gain']))
print('Дібольд - Маріано:', {k: (round(v[0], 3), round(v[1],
3))

```

```
for k, v in R['dm'].items()})
```

9) 4.2.3. Структура контингенту та вертикальне узгодження (табл. 4.10–4.12)

```

slo, shi = YEAR_STRUCT
ds = df[(df.year >= slo) & (df.year <=
shi)].reset_index(drop=True)
struct, fc_cat = {}, {}
for c in CATS:
    v = ds[c].values.astype(float)
    if c == 'pens_judge':
        # ряд змінюється в межах 3-4 тис. осіб - наївна модель
        f = np.repeat(v[-1], HORIZON)
        struct[c] = {'model': 'наївна'}
    else:
        m = Holt(v, initialization_method='estimated').fit()
        f = np.asarray(m.forecast(HORIZON))
        e = v - np.asarray(m.fittedvalues)
        struct[c] = {
            'model': 'Holt',
            'alpha': float(m.params['smoothing_level']),
            'beta': float(m.params['smoothing_trend']),
            'RMSE': float(np.sqrt(np.mean(e[1:] ** 2))),
            'MAPE': float(np.mean(np.abs(e[1:] / v[1:]))) *
100),
        }
    fc_cat[c] = f
    struct[c]['last'] = float(v[-1])
    struct[c]['fc'] = np.round(f, 1).tolist()
bottom_up = np.sum([fc_cat[c] for c in CATS], axis=0)
top_down = np.asarray(Holt(ds.pens_tot.values.astype(float),
initialization_method='estimated')
.fit().forecast(HORIZON))
discrepancy = bottom_up - top_down
# узгодження за формулою (4.2): пропорційний перерозподіл
розбіжності
lam = np.array([fc_cat[c] for c in CATS])
weights = lam / lam.sum(axis=0)
reconciled = lam - weights * discrepancy
R['struct'] = {
    'cats': struct, 'labels': CAT_LABELS,
    'years': list(range(2022, 2022 + HORIZON)),
    'bu': np.round(bottom_up, 1).tolist(),
    'td': np.round(top_down, 1).tolist(),
    'disc': np.round(discrepancy, 1).tolist(),
    'disc_pct': np.round(discrepancy / top_down * 100,
3).tolist(),
    'rec': {c: np.round(reconciled[i], 1).tolist() for i, c in
enumerate(CATS)},
    'share2021': {c: round(float(ds[c].values[-1] /
ds.pens_tot.values[-1] * 100), 2)
for c in CATS},
    'share2030': {c: round(float(reconciled[i][-1] /
top_down[-1] * 100), 2)
for i, c in enumerate(CATS)},

```

```

    }
    print('\nТабл. 4.11 - неузгодженість «знизу вгору» vs «згори
вниз», %:',
        R['struct']['disc_pct'])
    print('Частки 2021 p.:', R['struct']['share2021'])
    print('Частки 2030 p.:', R['struct']['share2030'])
10) Емпіричне покриття прогнозних інтервалів PICP (формула (3.41))
    hits, widths = 0, []
    for n in range(MIN_TRAIN_MACRO, len(y)):
        sc = StandardScaler().fit(d[DEM].iloc[:n])
        Ztr, Zte = sc.transform(d[DEM].iloc[:n]),
sc.transform(d[DEM].iloc[n:n + 1])
        pc = PCA().fit(Ztr)
        Xtr = np.column_stack([pc.transform(Ztr)[: , :2],
d.cpi_ll.values[:n]])
        Xte = np.column_stack([np.ones(1), pc.transform(Zte)[: ,
:2],
                                d.cpi_ll.values[n:n + 1]])
        res = sm.OLS(y[:n], sm.add_constant(Xtr)).fit()
        point = float((Xte @ res.params)[0])
        var = float((Xte @ res.cov_params() @ Xte.T)[0, 0]) +
res.mse_resid
        tc = stats.t.ppf(0.9, res.df_resid) # 80-
відсотковий інтервал
        low, high = point - tc * np.sqrt(var), point + tc *
np.sqrt(var)
        widths.append(high - low)
        hits += int(low <= y[n] <= high)
    R['picp'] = {'PICP': hits / len(widths), 'n': len(widths),
                'nominal': 0.80, 'mean_width':
float(np.mean(widths))}
    print('\nPICP = %.3f за номінального рівня 0,80 (n = %d),
середня ширина %.2f в. п.'
        % (R['picp']['PICP'], R['picp']['n'],
R['picp']['mean_width']))
    json.dump(R, open(OUT / 'main.json', 'w'), ensure_ascii=False,
                indent=1, default=float)
    print('\nЗбережено:', OUT / 'main.json')
if __name__ == '__main__':
    main()

```

В.5. Крок 3. Сценарне прогнозування та оцінювання області застосовності

Вхідні дані: out/clean.csv, out/main.json. Результат: out/scen.json.

Лістинг В.5 – 03_scenarios.py

1) Заголовок модуля, опис і підключення бібліотек

```

# -*- coding: utf-8 -*-
"""

```

Крок 3. Експеримент 6 – сценарне прогнозування та оцінювання області

застосовності моделей (підрозділ 4.3.4, табл. 4.19–4.21).

Сценарій задається не траєкторією цільового показника, а узгодженим набором

припущень щодо чинників. Базою для всіх трьох сценаріїв є узгоджений прогноз

контингенту за видами пенсій з кроку 2: мікрорівневі оцінки виступають обмеженнями для макрорівневої моделі (вертикальне узгодження).
Результат: out/scen.json
"""

```
import json
import warnings
import numpy as np
import pandas as pd
import statsmodels.api as sm
from scipy import stats
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from statsmodels.tsa.holtwinters import Holt
from config import OUT, DEM, CATS, YEAR_MACRO, HORIZON,
FORECAST_YEARS
warnings.filterwarnings('ignore')
```

2) Позначення сценаріїв

```
SCENARIO_NAMES = {'S1': 'інерційний',
                  'S2': 'помірного погіршення',
                  'S3': 'несприятливий'}
```

3) Основна процедура сценарних розрахунків

```
def main():
```

4) Завантаження результатів попередніх кроків

```
    main_json = json.load(open(OUT / 'main.json'))
    df = pd.read_csv(OUT / 'clean.csv')
    lo, hi = YEAR_MACRO
    d = df[(df.year >= lo) & (df.year <=
hi)].reset_index(drop=True)
    y = d.soc_gdp.values.astype(float)
    scaler = StandardScaler().fit(d[DEM])
    Z = scaler.transform(d[DEM])
    pca = PCA().fit(Z)
    K = pca.transform(Z)
    macro = sm.OLS(y, sm.add_constant(
        np.column_stack([K[:, :2], d.cpi_1l.values]))).fit()
```

5) Екстраполяція демографічного фону (дані про населення є і за 2022 р.)

```
    hist = df[(df.year >= 2015) & (df.year <= 2022)]
    pop_hist = hist.pop_tot.values.astype(float)
    pop_base = np.concatenate([[pop_hist[-1]],
                                Holt(pop_hist,
initialization_method='estimated')
                                .fit().forecast(HORIZON - 1)])
    urb_hist = hist.share_urb.values.astype(float)
    urb_path = np.concatenate([[urb_hist[-1]],
                                Holt(urb_hist,
initialization_method='estimated')
                                .fit().forecast(HORIZON - 1)])
    le21 = float(df.loc[df.year == 2021, 'le_all'].iloc[0])
    le19 = float(df.loc[df.year == 2019, 'le_all'].iloc[0])
    gap21 = float(df.loc[df.year == 2021, 'le_gap'].iloc[0])
    # узгоджений прогноз контингенту за категоріями з кроку 2
    base = {c: np.array(v, dtype=float) for c, v in
```

```
main_json['struct']['rec'].items() }
```

6) Складання одного сценарію у просторі вихідних показників

```
def build(dis_multiplier, pop_scale, le_path, gap_path,
          cpi_path):
    """Складання сценарію у просторі вихідних показників."""
    cat = {c: base[c].copy() for c in CATS}
    # додатковий приріст чисельності одержувачів пенсій по
    інвалідності
    cat['pens_dis'] = cat['pens_dis'] * np.linspace(1.0,
dis_multiplier, HORIZON)
    total = sum(cat.values())
    pop = pop_base * pop_scale
    X = pd.DataFrame({
        'pens_tot': total, 'pens_age': cat['pens_age'],
'pens_dis': cat['pens_dis'],
        'pens_surv': cat['pens_surv'], 'pens_serv':
cat['pens_serv'],
        'pens_soc': cat['pens_soc'], 'pop_tot': pop,
'share_urb': urb_path,
        'le_all': le_path, 'le_gap': gap_path,
'support_ratio': pop / total,
    }) [DEM]
    # проєкція сценарію на простір головних компонент
    Kf = pca.transform(scaler.transform(X))[:, :2]
    design = np.column_stack([np.ones(HORIZON), Kf[:, 0],
Kf[:, 1], cpi_path])
    pred = design @ macro.params
    var = np.einsum('ij,jk,ik->i', design, macro.cov_params(),
        design) + macro.mse_resid
    tc = stats.t.ppf(0.95, macro.df_resid) # 90-
відсотковий інтервал
    se = np.sqrt(var)
    return {'cat': cat, 'tot': total, 'pop': pop, 'burden':
total / pop * 1000,
        'K1': Kf[:, 0], 'K2': Kf[:, 1], 'cpi': cpi_path,
        'pred': pred, 'lo': pred - tc * se, 'hi': pred +
tc * se}
```

7) Припущення сценаріїв (табл. 4.19)

```
S = {
    'S1': build(1.00, np.ones(HORIZON),
        np.linspace(le21, le19, HORIZON),
np.repeat(gap21, HORIZON),
        np.repeat(106.0, HORIZON)),
    'S2': build(1.18, np.linspace(1.0, 0.985, HORIZON),
        np.repeat(le21, HORIZON), np.linspace(gap21,
gap21 + 0.4, HORIZON),
        np.array([120, 115, 112, 110, 109, 108, 108,
107, 107.0])),
    'S3': build(1.45, np.linspace(1.0, 0.960, HORIZON),
        np.linspace(le21, le21 - 1.2, HORIZON),
np.linspace(gap21, gap21 + 1.0, HORIZON),
        np.array([135, 128, 122, 118, 115, 113, 112,
111, 110.0])),
```

```

    }
8) Область застосовності: відстань Махаланобіса (формула (4.4))
    mu = K[:, :2].mean(axis=0)
    cov_inv = np.linalg.inv(np.cov(K[:, :2].T))
    critical = float(np.sqrt(stats.chi2.ppf(0.95, 2)))
    lo1, hi1 = K[:, 0].min(), K[:, 0].max()
    lo2, hi2 = K[:, 1].min(), K[:, 1].max()
    print('Діапазон K1 на навчальній вибірці: [%.2f; %.2f], K2:
[%.2f; %.2f]'
        % (lo1, hi1, lo2, hi2))
    print('Критичне значення відстані Махаланобіса (рівень 0,95):
%.2f' % critical)
9) Формування підсумкової структури результатів
    out = {}
    for tag, s in S.items():
        dm = [float(np.sqrt((np.array([a, b]) - mu) @ cov_inv @
(np.array([a, b]) - mu)))
            for a, b in zip(s['K1'], s['K2'])]
        in_range = [bool(lo1 <= a <= hi1 and lo2 <= b <= hi2)
            for a, b in zip(s['K1'], s['K2'])]
        out[tag] = {
            'name': SCENARIO_NAMES[tag], 'year': FORECAST_YEARS,
            'tot': np.round(s['tot'], 1).tolist(),
            'pop': np.round(s['pop'], 1).tolist(),
            'burden': np.round(s['burden'], 1).tolist(),
            'dis': np.round(s['cat']['pens_dis'], 1).tolist(),
            'K1': np.round(s['K1'], 3).tolist(), 'K2':
np.round(s['K2'], 3).tolist(),
            'cpi': list(map(float, s['cpi'])),
            'pred': np.round(s['pred'], 3).tolist(),
            'lo': np.round(s['lo'], 3).tolist(), 'hi':
np.round(s['hi'], 3).tolist(),
            'hull_d': np.round(dm, 2).tolist(), 'hull_in':
in_range,
        }
        print('\n%s - %s' % (tag, SCENARIO_NAMES[tag]))
        print('    контингент, тис. осіб :', np.round(s['tot'], 0))
        print('    навантаження на 1000    :', np.round(s['burden'],
1))
        print('    частка видатків, %% ВВП:', np.round(s['pred'],
2))
        print('    відстань dM                :', np.round(dm, 2))
        fact_tot = float(d.pens_tot.values[-1])
        fact_burden = fact_tot / float(d.pop_tot.values[-1]) * 1000
        print('\nФакт 2021 р.: контингент %.0f тис. осіб, навантаження
%.1f, Y = %.1f %%'
            % (fact_tot, fact_burden, y[-1]))
        print('Розкид контингенту у 2030 р.: %.1f тис. осіб (%.2f %%)'
            % (out['S3']['tot'][-1] - out['S1']['tot'][-1],
                (out['S3']['tot'][-1] / out['S1']['tot'][-1] - 1) *
100))
        json.dump(out, open(OUT / 'scen.json', 'w'),
ensure_ascii=False, indent=1)

```

```

    print('\nЗбережено:', OUT / 'scen.json')
if __name__ == '__main__':
    main()

```

В.6. Крок 4. Побудова рисунків

Вхідні дані: out/*.csv, out/*.json. Результат: fig/*.png.

Лістинг В.6 – 04_figures.py

1) Заголовок модуля, опис і підключення бібліотек

```
# -*- coding: utf-8 -*-
```

```
"""
```

Крок 4. Побудова рисунків 4.1–4.14 розділу.

Шрифт Liberation Serif метрично сумісний із Times New Roman і містить

кирилицю. Якщо його немає в системі, замініть значення

FONT_FAMILY.

Результат: fig/fig_4_1.png ... fig/fig_4_14.png

```
"""
```

```

import json
import warnings
import matplotlib
import numpy as np
import pandas as pd
matplotlib.use('Agg')
import matplotlib.pyplot as plt          # noqa: E402
from matplotlib import rcParams          # noqa: E402
from matplotlib.ticker import MaxNLocator # noqa: E402
from config import OUT, FIG, DEM         # noqa: E402
warnings.filterwarnings('ignore')
FONT_FAMILY = 'Liberation Serif'
rcParams['font.family'] = FONT_FAMILY
rcParams['font.size'] = 11
rcParams['axes.grid'] = True
rcParams['grid.alpha'] = 0.3
rcParams['axes.axisbelow'] = True
# палітра: темно-синій, блакитний, теракотовий, графітовий,
# зелений, вохра, фіолетовий
C = ['#3d5a80', '#98c1d9', '#ee6c4d', '#293241', '#81b29a',
     '#f2cc8f', '#6d597a']
CAT_KEYS = ['pens_age', 'pens_dis', 'pens_surv', 'pens_serv',
            'pens_soc']
CAT_TEXT = ['за віком', 'по інвалідності', 'у зв'язку з
            втратою\пгодувальника',
            'за вислугу років', 'соціальні']
XNAMES = ['X%d' % i for i in range(1, 12)]

```

2) Службова функція збереження рисунка

```

def save(fig, name):
    fig.tight_layout()
    fig.savefig(FIG / (name + '.png'), dpi=200,
bbox_inches='tight', facecolor='white')
    plt.close(fig)
    print(name)
def main():
    df = pd.read_csv(OUT / 'clean.csv')
    M = json.load(open(OUT / 'main.json'))

```

```

S = json.load(open(OUT / 'scen.json'))
cv = pd.read_csv(OUT / 'e1_cv.csv')
fc = pd.read_csv(OUT / 'e1_forecast.csv')
3) 4.1 динаміка населення і контингенту
fig, ax = plt.subplots(figsize=(7.2, 3.6))
ax.plot(df.year, df.pop_tot / 1000, color=C[0], lw=2,
label='Населення, млн осіб')
ax.set_ylabel('Населення, млн осіб', color=C[0])
ax.tick_params(axis='y', labelcolor=C[0])
ax2 = ax.twinx()
ax2.grid(False)
ax2.plot(df.year, df.pens_tot / 1000, color=C[2], lw=2,
label='Пенсіонери, млн осіб')
ax2.set_ylabel('Пенсіонери, млн осіб', color=C[2])
ax2.tick_params(axis='y', labelcolor=C[2])
ax.axvspan(2013.5, 2015.5, color='0.75', alpha=.35)
ax.annotate('структурний\nпрозрів 2014', xy=(2017.5, 47.5),
fontsize=9,
ha='center', color='0.25')
ax.set_xlabel('Рік')
ax.set_title('Чисельність населення та контингенту одержувачів
пенсій, '
'1990-2022 pp.', fontsize=11)
h1, l1 = ax.get_legend_handles_labels()
h2, l2 = ax2.get_legend_handles_labels()
ax.legend(h1 + h2, l1 + l2, loc='lower left', fontsize=9,
framealpha=.9)
save(fig, 'fig_4_1')
4) 4.2 структура контингенту
dd = df[(df.year >= 1995) & (df.year <=
2021)].dropna(subset=CAT_KEYS)
fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.4))
axs[0].stackplot(dd.year, [dd[c] / 1000 for c in CAT_KEYS],
labels=CAT_TEXT, colors=C[:5], alpha=.9)
axs[0].set_ylabel('млн осіб')
axs[0].set_xlabel('Рік')
axs[0].set_title('Контингент за видами пенсій', fontsize=10)
axs[0].legend(fontsize=7.5, loc='lower left', ncol=2)
share = dd[CAT_KEYS].div(dd[CAT_KEYS].sum(axis=1), axis=0) *
100
for i, c in enumerate(CAT_KEYS):
axs[1].plot(dd.year, share[c], color=C[i], lw=1.8,
label=CAT_TEXT[i].replace('\n', ' '))
axs[1].set_ylabel('% контингенту')
axs[1].set_xlabel('Рік')
axs[1].set_yscale('log')
axs[1].set_title('Структура контингенту (логарифмічна шкала)',
fontsize=10)
axs[1].legend(fontsize=7, loc='lower left', framealpha=.85)
save(fig, 'fig_4_2')
5) 4.3 видатки та ІСЦ
fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.2))
axs[0].plot(df.year, df.soc_gdp, color=C[0], lw=2, marker='o',

```

```

ms=3)
    axs[0].set_ylabel('% до ВВП')
    axs[0].set_xlabel('Рік')
    axs[0].set_title('Державні видатки на соціальний захист',
fontsize=10)
    axs[1].semilogy(df.year, df.cpi, color=C[2], lw=2, marker='s',
ms=3)
    axs[1].axhline(100, color='0.4', ls='--', lw=1)
    axs[1].set_ylabel('ІЦЦ, % до попереднього року')
    axs[1].set_xlabel('Рік')
    axs[1].set_title('Індекс споживчих цін (логарифмічна шкала)',
fontsize=10)
    save(fig, 'fig_4_3')
6) 4.4 кореляційна матриця
    cols = DEM + ['cpi']
    labels = XNAMES + ['X12']
    corr = df[(df.year >= 2002) & (df.year <=
2021)][cols].corr().values
    fig, ax = plt.subplots(figsize=(5.6, 4.8))
    im = ax.imshow(corr, cmap='RdBu_r', vmin=-1, vmax=1)
    ax.set_xticks(range(len(labels)))
    ax.set_xticklabels(labels, fontsize=8)
    ax.set_yticks(range(len(labels)))
    ax.set_yticklabels(labels, fontsize=8)
    for i in range(len(labels)):
        for j in range(len(labels)):
            ax.text(j, i, '%.2f' % corr[i, j], ha='center',
va='center', fontsize=5.6,
                    color='white' if abs(corr[i, j]) > 0.6 else
'black')
    ax.grid(False)
    fig.colorbar(im, ax=ax, shrink=.85)
    ax.set_title('Кореляційна матриця факторних ознак, 2002-2021
pp.', fontsize=10)
    save(fig, 'fig_4_4')
7) 4.5 діаграма «кам'янистого осипу»
    ev = np.array(M['ev'])
    x = np.arange(1, len(ev) + 1)
    fig, ax = plt.subplots(figsize=(6.4, 3.2))
    ax.bar(x, ev * 100, color=C[1], edgecolor=C[0], label='частка
поясненої дисперсії')
    ax.plot(x, np.cumsum(ev) * 100, color=C[2], marker='o', ms=4,
lw=2,
            label='накопичена частка')
    ax.axhline(80, color='0.35', ls='--', lw=1.2)
    ax.text(len(ev) - 3, 81.5, 'критерій 80 %', fontsize=8,
color='0.3')
    ax.set_xlabel('Номер головної компоненти')
    ax.set_ylabel('%')
    ax.set_xticks(x)
    ax.legend(fontsize=8, loc='center right')
    ax.set_title('Діаграма «кам'янистого осипу» головних
компонент', fontsize=10)

```

```

    save(fig, 'fig_4_5')
8) 4.6 навантаження на компоненти
    load = pd.DataFrame(M['load'])
    idx = np.arange(len(load))
    w = .38
    fig, ax = plt.subplots(figsize=(6.8, 3.4))
    ax.barh(idx - w / 2, load['K1'], w, color=C[0],
            label='K1 (%.2f %%)' % (ev[0] * 100))
    ax.barh(idx + w / 2, load['K2'], w, color=C[2],
            label='K2 (%.2f %%)' % (ev[1] * 100))
    ax.set_yticks(idx)
    ax.set_yticklabels(XNAMES, fontsize=9)
    ax.invert_yaxis()
    ax.axvline(0, color='0.3', lw=1)
    ax.set_xlabel('Навантаження')
    ax.legend(fontsize=8)
    ax.set_title('Навантаження факторних ознак на головні
компоненти', fontsize=10)
    save(fig, 'fig_4_6')
9) 4.7 фактичні/модельні значення та залишки
    mf = M['macro_fit']
    fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.2))
    axs[0].plot(mf['year'], mf['actual'], color=C[3], lw=2,
marker='o', ms=4,
                label='фактичні')
    axs[0].plot(mf['year'], mf['fit'], color=C[2], lw=2, ls='--',
marker='s', ms=4,
                label='модельні')
    axs[0].set_xlabel('Рік')
    axs[0].set_ylabel('% до ВВП')
    axs[0].legend(fontsize=9)
    axs[0].set_title('Зіставлення фактичних і модельних значень',
fontsize=10)
    axs[1].bar(mf['year'], mf['resid'],
                color=[C[0] if r >= 0 else C[2] for r in
mf['resid']])
    axs[1].axhline(0, color='0.3', lw=1)
    sd = np.std(mf['resid'], ddof=4) # 4 оцінені
параметри моделі
    axs[1].axhline(2 * sd, color='0.5', ls=':')
    axs[1].axhline(-2 * sd, color='0.5', ls=':')
    axs[1].set_xlabel('Рік')
    axs[1].set_ylabel('В. п.')
    axs[1].set_title('Залишки моделі ( $\pm 2\sigma$ )', fontsize=10)
    save(fig, 'fig_4_7')
10) 4.8 важливість ознак
    rf, gb, pm = M['ml_imp']['rf'], M['ml_imp']['gb'],
M['ml_perm']
    keys = list(rf)
    names = XNAMES + ['X12(-1)']
    x = np.arange(len(keys))
    w = .27
    fig, ax = plt.subplots(figsize=(7.2, 3.2))

```

```

    ax.bar(x - w, [rf[k] for k in keys], w, color=C[0],
label='випадковий ліс (MDI)')
    ax.bar(x, [gb[k] for k in keys], w, color=C[2],
label='градієнтний бустинг (MDI)')
    ax.bar(x + w, [pm[k] for k in keys], w, color=C[4],
label='перестановочна важливість')
    ax.set_xticks(x)
    ax.set_xticklabels(names, fontsize=8.5)
    ax.set_ylabel('Відносна важливість')
    ax.legend(fontsize=8)
    ax.set_title('Оцінки важливості факторних ознак', fontsize=10)
    save(fig, 'fig_4_8')

```

11) 4.9 ретропрогнози макромоделей

```

P, yrs, act = M['macro_cv_pred'], M['macro_cv_years'],
M['macro_cv_actual']
fig, ax = plt.subplots(figsize=(6.8, 3.3))
ax.plot(yrs, act, color=C[3], lw=2.4, marker='o', ms=5,
label='факт')
for tag, lab, col in [('pca2cpi', 'ГК+ІСЦ(-1)', C[2]),
                      ('rf', 'випадковий ліс', C[4]),
                      ('gb', 'градієнтний бустинг', C[5]),
                      ('naive', 'наївний', C[1])]:
    ax.plot(yrs, P[tag], lw=1.6, ls='--', marker='.',
color=col, label=lab)
    ax.set_xlabel('Рік')
    ax.set_ylabel('% до ВВП')
    ax.legend(fontsize=8, ncol=2)
    ax.set_title('Ретропрогнози на ковзному вікні (h = 1)',
fontsize=10)
    save(fig, 'fig_4_9')

```

12) 4.10 точність моделей контингенту

```

fig, ax = plt.subplots(figsize=(6.8, 3.2))
x = np.arange(len(cv))
w = .26
for i, col in enumerate(['MAPE h=1*', 'MAPE h=2*', 'MAPE
h=3*']):
    ax.bar(x + (i - 1) * w, cv[col], w, color=C[i],
label='h = %d' % (i + 1))
    ax.set_xticks(x)
    ax.set_xticklabels(cv['Модель'], fontsize=8, rotation=15)
    ax.set_ylabel('MAPE, %')
    ax.legend(fontsize=8, title='горизонт', title_fontsize=8)
    ax.set_title('Точність моделей контингенту, початки прогнозу
2015-2020 pp.',
fontsize=10)
    save(fig, 'fig_4_10')

```

13) 4.11 прогноз контингенту з інтервалами

```

hist = df[(df.year >= 1995) & (df.year <= 2021)]
fig, ax = plt.subplots(figsize=(7.2, 3.4))
ax.plot(hist.year, hist.pens_tot / 1000, color=C[3], lw=2,
label='фактичні дані')
ax.plot(fc.year, fc.fc / 1000, color=C[2], lw=2, ls='--',
label='ARIMAX(1,1,0) з утрочанням')

```

```

    ax.fill_between(fc.year, fc.lo95 / 1000, fc.hi95 / 1000,
color=C[2], alpha=.14,
                    label='95 % інтервал')
    ax.fill_between(fc.year, fc.lo80 / 1000, fc.hi80 / 1000,
color=C[2], alpha=.26,
                    label='80 % інтервал')
    ax.plot(fc.year, fc.holt / 1000, color=C[0], lw=1.6, ls=':',
label='Holt')
    ax.axvline(2021.5, color='0.5', lw=1)
    ax.set_xlabel('Рік')
    ax.set_ylabel('млн осіб')
    ax.legend(fontsize=8)
    ax.set_title('Прогноз чисельності контингенту одержувачів
пенсій до 2030 р.',
                fontsize=10)
    save(fig, 'fig_4_11')

```

14) 4.12 узгодження прогнозів

```

st = M['struct']
yy = st['years']
fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.2))
axs[0].plot(yy, np.array(st['bu']) / 1000, color=C[0], lw=2,
marker='o', ms=4,
            label='«знизу вгору» (сума категорій)')
axs[0].plot(yy, np.array(st['td']) / 1000, color=C[2], lw=2,
ls='--', marker='s',
            ms=4, label='«згори вниз» (прямий прогноз)')
axs[0].set_xlabel('Рік')
axs[0].set_ylabel('млн осіб')
axs[0].legend(fontsize=8)
axs[0].set_title('Два способи прогнозування контингенту',
fontsize=10)
axs[1].bar(yy, st['disc_pct'], color=C[1], edgecolor=C[0])
axs[1].axhline(0, color='0.3', lw=1)
axs[1].set_xlabel('Рік')
axs[1].set_ylabel('%')
axs[1].set_title('Неузгодженість прогнозів, % до прямого
прогнозу', fontsize=10)
for a in axs:
    a.xaxis.set_major_locator(MaxNLocator(integer=True))
save(fig, 'fig_4_12')

```

15) 4.13 сценарії контингенту й навантаження

```

cols = {'S1': C[4], 'S2': C[5], 'S3': C[2]}
h2 = df[(df.year >= 2015) & (df.year <= 2021)]
fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.3))
axs[0].plot(h2.year, h2.pens_tot / 1000, color=C[3], lw=2,
label='факт')
axs[1].plot(h2.year, h2.pens_per1000_calc, color=C[3], lw=2,
label='факт')
for k in ['S1', 'S2', 'S3']:
    lab = '%s - %s' % (k, S[k]['name'])
    axs[0].plot(S[k]['year'], np.array(S[k]['tot']) / 1000,
color=cols[k],
                lw=1.9, ls='--', label=lab)

```

```

        axs[1].plot(S[k]['year'], S[k]['burden'], color=cols[k],
lw=1.9, ls='--',
                    label=lab)
    axs[0].set_ylabel('млн осіб')
    axs[0].set_title('Контингент одержувачів пенсій', fontsize=10)
    axs[1].set_ylabel('осіб на 1000 населення')
    axs[1].set_title('Пенсійне навантаження', fontsize=10)
    for a in axs:
        a.set_xlabel('Рік')
        a.axvline(2021.5, color='0.6', lw=1)
        a.legend(fontsize=7.5)
        a.xaxis.set_major_locator(MaxNLocator(integer=True))
    save(fig, 'fig_4_13')
16) 4.14 сценарії видатків і область застосовності
    h3 = df[(df.year >= 2012) & (df.year <= 2021)]
    fig, axs = plt.subplots(1, 2, figsize=(9.6, 3.3))
    axs[0].plot(h3.year, h3.soc_gdp, color=C[3], lw=2, marker='o',
ms=4, label='факт')
    for k in ['S1', 'S2', 'S3']:
        axs[0].plot(S[k]['year'][:5], S[k]['pred'][:5],
color=cols[k], lw=1.9,
                    ls='--', marker='.', label='%s - %s' % (k,
S[k]['name']))
        axs[0].fill_between(S['S1']['year'][:5], S['S1']['lo'][:5],
S['S3']['hi'][:5],
                        color='0.75', alpha=.25)
        axs[0].axvline(2021.5, color='0.6', lw=1)
        axs[0].set_xlabel('Рік')
        axs[0].set_ylabel('% до ВВП')
        axs[0].legend(fontsize=7.5)
        axs[0].set_title('Сценарні прогнози частки видатків, 2022-2026
pp.', fontsize=10)
    for k in ['S1', 'S2', 'S3']:
        axs[1].plot(S[k]['year'], S[k]['hull_d'], color=cols[k],
lw=1.9, marker='.',
                    label=k)
        axs[1].axhline(2.45, color=C[2], ls='--', lw=1.4)
        axs[1].text(2022.2, 2.6, 'межа області навчання (95 %)',
fontsize=8, color=C[2])
        axs[1].set_xlabel('Рік')
        axs[1].set_ylabel('відстань Махаланобіса у просторі (K1, K2)')
        axs[1].legend(fontsize=8)
        axs[1].set_title('Показник виходу за область навчання моделі',
fontsize=10)
    for a in axs:
        a.xaxis.set_major_locator(MaxNLocator(integer=True))
    save(fig, 'fig_4_14')
    print('\nРисунки збережено до', FIG)
if __name__ == '__main__':
    main()

```