УДК 004.89:519.22](043.3)

Oleksii Zarudnii, PhD student
ORCID ID: https://orcid.org/0009-0008-7462-3899 ***e-mail:*** oleksii.zarudnyi@gmail.com

Institute of Telecommunications and Global Information Space of the NASU, Kyiv, Ukraine

USAGE OF INTELLECTUAL ANALYSIS OF ECOLOGICAL-ECONOMIC DATA IN THE PENSION INSURANCE SYSTEM AND FOR FORECASTING EXPENDITURES ON SOCIAL PROTECTION AND SOCIAL SECURITY

Abstract. *The paper is devoted to an actual scientific and applied problem – the development of methodology for applying mathematical models and data mining methods for actuarial calculations in mandatory state pension insurance system. The paper describes methodology for modeling changes in the number of pension recipients taking into account the impact of environmental factors, in particular air pollution. The basis of the proposed method is a multi-model approach, characterized by combination of data mining and probabilistic models in the form of Bayesian network, which are appropriate in conditions of statistical, parametric and structural uncertainty.*

The proposed approach describes the change in number of pension recipients, in particular for disability and breadwinner loss, under influence of air pollution from organic and inorganic compounds. The scientific novelty of the paper is in the use of an ensemble of models including probabilistic and statistical models in the form of Bayesian network and regression models, in the system of actuarial calculations of mandatory state pension insurance.

The paper considers several scenarios for the impact of pollutants on the growth of number of pension recipients. The indicator of the share of expenditures on social protection and social security of the population in the gross national product was chosen as the target variable of the process under study. Mathematical models were found to be adequate to the modeling process, and the Bayesian network classification error is about 20%. The model structure is built in Genie 2.0 modeling system. The principal component analysis, is used to reduce the data dimension. The proposed methodology can also be applied to other tasks of forecasting social protection and social security expenditures.

Key words: *expenses for social protection and social security, air pollution, Bayesian network, principal component analysis, the system of actuarial calculations.*

О.Б. Зарудний

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ, Україна

ЗАСТОСУВАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ЕКОЛОГО-ЕКОНОМІЧНИХ ДАНИХ У СИСТЕМІ ПЕНСІЙНОГО СТРАХУВАННЯ ТА ДЛЯ ПРОГНОЗУВАННЯ ВИТРАТ НА СОЦІАЛЬНИЙ ЗАХИСТ І СОЦІАЛЬНЕ ЗАБЕЗПЕЧЕННЯ

Анотація. *Стаття присвячена актуальній науково-прикладній проблемі – розробці методики застосування математичних моделей та методів інтелектуального аналізу даних для актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування. В роботі описано методику моделювання зміни контингенту одержувачів пенсій за урахування впливу екологічних чинників, зокрема забрудненості повітря. Основу пропонованої методики становить багатомодельний підхід, особливістю якого є поєднання інтелектуального аналізу даних та ймовірнісних моделей у формі мереж Байєса, застосування яких є доцільним в умовах статистичної, параметричної та структурної невизначеності.*

Пропонований підхід описує зміну контингенту одержувачів пенсій, зокрема по інвалідності та по втраті годувальника, під впливом екологічних чинників – зміни забруднення повітря органічними та неорганічними сполуками. Науковою новизною роботи є застосування ансамблю моделей, до якого входять ймовірнісно-статистичні моделі у формі мереж Байєса та регресійні моделі, у системі актуарних розрахунків загальнообов'язкового державного пенсійного страхування.

В роботі розглянуто декілька сценаріїв впливу забруднюючих речовин на зростання контингенту одержувачів пенсій. В якості цільової змінної досліджуваного процесу обрано показник частки видатків на соціальний захист та соціальне забезпечення населення у валовому національному продукті. Математичні моделі виявилися адекватними обраному для моделювання процесу, помилка класифікації мереж Байєса становить близько 20%. Структуру моделі побудовано в системі моделювання Genie 2.0. Для зменшення розмірності даних застосовано метод головних компонент. Пропонована методика може бути застосована і для інших задач прогнозування витрат на соціальний захист та соціальне забезпечення.

Ключові слова: *витрати на соціальний захист та соціальне забезпечення, забруднення повітря, мережа Байєса, метод головних компонентів, система актуарних розрахунків.*

https://doi.org/10.32347/2411-4049.2024.3.161-176

Вступ

Питання використання математичних моделей, методів інтелектуального аналізу даних тощо у актуарних розрахунках та для прогнозування державних видатків на соціальний захист та соціальне забезпечення, незважаючи на значну кількість досліджень, представлених у роботах вітчизняних та закордонних фахівців [1–9], залишається не вирішеним. Перш за все, це пов'язано із значною кількістю факторів, зокрема демографічних чинників, поточною соціально-економічною ситуацією в країні, наслідками військового конфлікту, що впливають на обсяги видатків, крім того, необхідно враховувати й особливості чинного законодавства. Крім того, останнім часом в публікаціях

міжнародних організацій, роботах науковців, пов'язаних із соціальним забезпеченням та соціальним захистом населення, особлива увага приділяється оцінюванню економічного впливу смертності та захворюваності внаслідок впливу екологічних факторів та кліматичних змін [8, 9]. Як зазначається у підсумковому звіті «Вплив на здоров'я та соціальні витрати, пов'язані із забрудненням повітря у великих містах України» [9], виконаному в рамках Програми розвитку ООН в Україні, «у великих містах України, де якість повітря явно не оптимальна, покращення якості повітря до рівня, що вважається «хорошим», могло б запобігти третині випадків усіх інсультів і випадків хронічної обструктивної хвороби легень» [9]. За даними проєкту «Світовий індекс якості повітря» [10], у 2019 році 42 900 передчасних смертей і 953 500 років життя з поправкою на інвалідність (DALY) були пов'язані з впливом PM2,5, а за даними Глобального тягаря хвороб (GBD), це відповідає близько 10% всіх випадків захворюваності та смертності. Тому актуальною є задача урахування екологічної складової у актуарних розрахунках у системі загальнообов'язкового державного пенсійного страхування та для прогнозування витрат на соціальний захист та соціальне забезпечення. Що в свою чергу потребує розроблення відповідного аналітичного інструментарію.

Для розв'язання даної задачі запропоновано методику використання інтелектуального аналізу даних та математичного моделювання (ансамблів моделей, зокрема таких, що поєднують ймовірнісні моделі у формі мереж Байєса та регресійні моделі) для розширення переліку вхідних показників для проведення актуарних розрахунків за рахунок включення у модель еколого-економічних чинників – показників забруднення повітря.

Отже, тема роботи є актуальною, має практичну значущість та наукову новизну.

Постановка задачі

Метою дослідження є удосконалення існуючих методик прогнозування витрат на соціальний захист та соціальне забезпечення.

Задачі дослідження:

- виконати огляд сучасних методів моделювання та прогнозування витрат на соціальний захист та соціальне забезпечення в умовах невизначеності, спричиненої впливом різних груп чинників, в тому числі екологічних;
- проаналізувати дані, що використовуються для аналізу та прогнозування актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування;
- розробити методику побудови комп'ютерних моделей для опису та прогнозування чинників, що впливають на зростання видатків на соціальний захист та соціальне забезпечення в умовах невизначеності, спричиненої погіршенням екологічної ситуації у країні, зокрема зростанням забрудненості повітря органічними та неорганічними сполуками;
- удосконалити методику моделювання кількості одержувачів пенсій, зокрема пенсій по інвалідності в умовах зростання забрудненості повітря органічними та неорганічними сполуками;
- побудувати ймовірнісно-статистичні моделі у формі мережі Байєса, які описують вплив екологічних чинників на динаміку контингенту одержувачів пенсій;
- виконати апробацію розробленої методики на статистичних даних;

- проаналізувати якість прогнозів, отриманих за побудованими моделями, зробити висновки щодо можливості їх використання у роботі установ та організацій системи соціального захисту та соціального забезпечення;
- запропонувати шляхи покращення та практичного застосування вказаної методики.

Аналіз останніх досліджень і публікацій

Питання розроблення сучасного аналітичного інструментарію, призначеного до використання при проведенні актуарних розрахунків у системі пенсійного забезпечення та прогнозуванні витрат на соціальний захист і соціальне забезпечення, є актуальним для багатьох країн, про що свідчить значна кількість публікацій як фахівців-актуаріїв, так і науковців [1–9, 11–14]. Багато методик передбачають використання регресійних, економічних та макроекономічних моделей тощо, які призначені до використання у країнах із сталим економічним розвитком, усталеною системою пенсійного забезпечення та соціального захисту населення. Адаптувати вказані методики до використання в Україні в повній мірі практично не можливо, адже країна охоплена війною, що значно та швидко збільшує витрати на соціальний захист та соціальне забезпечення, тривають реформи системи пенсійного забезпечення та соціального захисту, країна є географічно великою, її регіони мають різний рівень соціально-економічного розвитку, а деякі з них знаходяться у зоні активних бойових дій. Іншим, не менш проблемним, є питання збору статистичних даних, достовірності отриманої інформації, необхідної для побудови складних моделей, використовуваних у актуарних розрахунках у системі державного пенсійного забезпечення. Тому потрібні нові, дещо спрощені підходи до виконання актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування, прогнозування витрат на соціальний захист та соціальне забезпечення. Зокрема, у дослідженні «Глобальний тягар хвороб» [15] запропонована статистична модель, у якій враховано, що внаслідок погіршення стану здоров'я і ймовірного настання інвалідності людина втрачає можливість брати активну участь у соціально-економічних процесах і робити свій внесок, що впливає як на збільшення соціальних витрат, так і на втрати людського капіталу, а також і на економічне зростання країни. Дане дослідження вирізняється тим, що в ньому використані регресійні моделі, які відображають вплив 87 факторів ризику, і найбільш вагомими з них виявились показники забруднення повітря. З огляду на поточну екологічну ситуацію в Україні та зростання рівня забрудненості повітря внаслідок бойових дій, урахування еколого-економічних чинників у задачі прогнозування витрат на пенсії, соціальний захист та соціальне забезпечення є обґрунтованим та доцільним.

Теоретичні основи дослідження

Методика проведення актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування, затверджена постановою Кабінету Міністрів України від 16 грудня 2004 року №1677 [16], яка врегульовує, в тому числі, й питання побудови коротко-, середньо- та довгострокового прогнозів фінансового стану пенсійної системи. Також, в ній визначається, що у системі проведення актуарних розрахунків використовуються методики, які «базуються на поєднанні методів теорії імовірності, математичного аналізу,

математичної статистики, теорії стохастичного аналізу, теорії складних відсотків, диференційних рівнянь та оптимізації». В сучасних умовах, що характеризуються наявністю невизначеностей різного типу, динамічним перебігом подій, необхідністю роботи з короткими вибірками різнорідних показників, важливо забезпечити системність використання різних методів аналізу, обробки даних, урахування можливих сценаріїв розвитку ситуації, вирішувати задачі прогнозування та підтримки прийняття рішень [17].

Одним із шляхів вирішення даної проблеми є впровадження багатомодельного підходу, основаного на поєднанні інтелектуального аналізу даних, використанні регресійних моделей, ймовірнісно-статистичних методів, сценарного аналізу, методів експертних оцінок тощо [17, 18].

Регресійні моделі широко застосовують для аналізу та прогнозування процесів різної природи [17, 19]. Методики їх використання представлені у багатьох роботах як закордонних, так і вітчизняних фахівців [19–24]. Основною перевагою даного підходу є простота, що й зумовило його використання для виявлення зв'язків в даних, тестування гіпотез, аналізу та прогнозування часових рядів тощо. У регресійному аналізі вирізняють аналітичний підхід та прогнозування. За прогнозного підходу знаходять найбільш значимі параметри моделей, а за аналітичного – досліджують силу та напрям зв'язків між регресорами та відгуком.

Регресійний аналіз часто застосовують разом із кореляційним аналізом, інколи навіть застосовують термін «кореляційно-регресійний аналіз», оскільки ці методи доповнюють один одного. Однак, якщо за допомогою кореляційного аналізу досліджується напрямок та щільність зв'язку між незалежними змінними, то в регресійному аналізі досліджується форма залежності між незалежними змінними.

У спеціалізованій літературі регресію визначають як залежність математичного очікування (середнього значення) [19, 23] незалежної змінної від однієї або декількох інших змінних.

Нехай, у точках x_n незалежної змінної *x* отримані виміри Y_n [22], і потрібно знайти залежність середнього значення величини $\overline{Y}$ від величини *x*, тобто, маємо

$$\overline{Y}(x) = f(x|a),$$

де *a* – вектор невідомих параметрів a_i [22].

Функція $f(x|a)$ – це функція регресії. Припускають, що $f(x|a)$ є лінійною функцією параметрів *a*, має вигляд (1) [32]:

$$f(x|a) = \sum_{i=1}^{I} a_i \varphi_i(x), \qquad (1)$$

де $f_i(x)$ – задані функції.

Тоді, матрицю $A_{ni} = f_i(x_n)$ називають регресійною матрицею [22].

Для визначення параметрів a_i зазвичай використовують метод найменших квадратів [22].

При пошуку функції регресії у вигляді (1) необхідно визначити кількість членів *I* у сумі (1).

Якщо *I* – мале, не можна досягти прийнятного опису $\overline{Y}(x)$, якщо велике – є ризик виникнення великих статистичних помилок функції регресії [22].

Парні лінійні регресійні моделі описують лінійну залежність між двома змінними [22]. У загальному випадку парна вибіркова регресійна модель [19] має вигляд (2):

$$Z = a_0 + a_1 X + E, \tag{2}$$

де Z – вектор значень залежної змінної, $Z = [z_1, z_2, \ldots, z_n]$; X – вектор значень незалежної змінної $X = [x_1, x_2, \ldots, x_n]$; a_0, a_1 – невідомі параметри регресійної моделі; E – вектор значень випадкових величин $E = [e_1, e_2, \ldots, e_n]$, його наявність зумовлена випадковими збуреннями, недосконалою структурою моделі, похибками вимірів та неточністю обчислень оцінок параметрів моделі.

Для дослідження зв'язку між $k+1$ змінними (k регресорів плюс один відгук) використовують $k-$мірну поверхню для прогнозування, тобто, у загальному випадку в модель регресії можна включати і більш складні зв'язки – квадратичні, кубічні, поліноміальні, перетин змінних [19].

Рівняння регресії із нелінійним зв’язком відгуку та регресорів має вигляд (3):

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_1^2 + \beta_3 \cdot X_2 + \beta_4 \cdot X_2^2 + \varepsilon\ . \tag{3}$$

Модель множинної лінійної регресії, на відміну від моделі лінійної регресії, дозволяє досліджувати зв'язки з декількома регресорами, що, звичайно, ускладнює вибір кращої моделі та подальшу її інтерпретацію, але створює умови для виявлення складніших залежностей.

Як зазначають [17, 25], при побудові моделей еколого-соціально-економічних процесів кращі результати можна отримати, використовуючи ансамблі моделей (при багатомодельному підході). Тобто в умовах, які характерні для більшості соціально-еколого-економічних систем, коли задачу складно формалізувати, а явище чи процес описує значна кількість чинників, серед яких важко визначити найбільш значимі та встановити причинно-наслідкові зв’язки між вхідними даними та вихідними показниками, а прогнози їх розвитку є підґрунтям для прийняття важливих управлінських рішень. Кращі результати, як зазначають [17], отримані за використання ансамблів моделей, до складу яких входять регресійні моделі в поєднанні із мережами Байєса, деревами рішень, когнітивними моделями.

Використання мереж Байєса, як зазначають фахівці [26–32], актуальне, якщо необхідно побудувати адекватну модель, в умовах, коли відсутня вичерпна інформація про досліджуване явище чи об’єкт, силу та характер впливу на них внутрішніх та зовнішніх чинників, їх динаміку, взаємодію з іншими процесами та системами.

Мережа Байєса являє собою пару [32] $<G, B>$, де G – це направлений ациклічний граф, а B – це множина параметрів, які визначають мережу, або множина таблиць умовних ймовірностей вершин: $P = P\left(X^{(i)} \middle| pa\left(X^{(i)}\right)\right), i = 1...N$ для кожного можливого значення $x^{(i)} \in X^{(i)}$ та $pa(X^{(i)}) \in Pa(X^{(i)})$, де $Pa(X^{(i)})$ – множина батьків змінної $X^{(i)} \in G$.

Кожна змінна $X^{(i)} \in G$ є вершиною мережі Байєса. Повна спільна ймовірність БМ обчислюється за формулою (4) [32]:

$$P(X^{(1)},...,X^{(N)}) = \prod_{i=1}^{N} P(X^{(i)} \left| Pa(X^{(i)}) \right.). \quad (4)$$

З математичної точки зору, мережа Байєса – це модель подання існуючих і не існуючих імовірнісних залежностей. При цьому зв'язок $A \to B$ є причинним, коли подія A – причина виникнення B, тобто коли існує механізм, відповідно до якого значення, набуте A, впливає на значення, набуте B.

Правило Байєса [32] має вигляд (5):

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}. \quad (5)$$

Дана залежність описує причинно-наслідкові зв’язки між спостереженнями та гіпотезами. Ймовірності $p(H)$ та $p(E \mid H)$ є апріорними, тобто задаються до початку спостережень, а ймовірність $P(H \mid E)$ є апостеріорною.

Переваги байєсівського методу полягають в тому, що апріорні ймовірності можна уточнювати (оновлювати), відповідно до характеру перебігу досліджуваного процесу, що дає можливість уточнювати ймовірності подій за надходження додаткової інформації.

Головне припущення теорії побудови мереж Байєса полягає в тому, що події є вичерпними ($\cup_{i=1}^{n} H_i = \Omega$) і не перетинаються (6). При виконанні цих умов ймовірність події E можна обчислити за допомогою умовних ймовірностей.

$$p(E) = \sum_{i=1}^{n} p(E \cap H_i) = \sum_{i=1}^{n} p(E \mid H_i) \cdot p(H_i). \quad (6)$$

Підставивши даний вираз в формулу (4), отримаємо формулу, яка є основою для побудови мереж Байєса (7) [32]:

$$p(H_k \mid E) = \frac{p(E \mid H_k) \cdot p(H_k)}{\sum_{i=1}^{n} p(E \mid H_i) \cdot p(H_i)}. \quad (7)$$

Побудова мережі Байєса охоплює комплекс задач, зокрема, пошук оптимальної структури мережі [29] – направленого ациклічного графа, що найбільш адекватно відповідає навчальним даним або досліджуваному процесу, обчисленням значень таблиць умовних ймовірностей для кожної вершини графа.

Інформацію про топологію графа (структуру) і параметри кожної вершини. можна отримати з навчальних даних. Найбільш складною задачею є отримання топології мережі. Особливо складно отримати правильну топологію у випадку, коли деякі вершини приховані або дані є некоректними чи неповними.

Для виконання точних розрахунків, пов'язаних з вибором моделі, необхідно обчислити $P(D) = \sum_G P(D \mid G)$, що є задачею експоненційної складності.

Для вирішення цієї проблеми застосовують байєсовий інформаційний критерій (8) [32]:

$$\log(P(G \mid D)) \approx \log(P(D \mid G, \hat{\theta}_G)) - \frac{\log(N)}{2} \cdot \dim(G), \qquad (8)$$

де N – число моделей, $\dim(G)$ – розмір моделі (кількість вільних параметрів), $\hat{\theta}_G$ – максимально правдоподібна оцінка параметрів, $-\frac{\log(N)}{2}\dim(G)$ – штрафуюча компонента за надмірну складність моделі.

Оскільки формула Байєса, фактично, є моделлю раціонального вибору в умовах неточної або неповної інформації, що особливо актуально для даного дослідження, що й зумовило вибір мереж Байєса до використання у складі ансамблю моделей.

Отже, обрані до використання у даному дослідженні методи та моделі є популярними методами інтелектуального аналізу даних і можуть бути ефективно використані для дослідження факторів, що впливають на обсяги видатків на соціальний захист та соціальне забезпечення, виявлення зв'язків в даних, в тому числі, й прихованих, та дають змогу передбачити позитивний (негативний) вплив окремих чинників тощо. Все це дозволить покращити якість управлінських рішень, оптимальніше використовувати наявні фінансові ресурси.

Методика дослідження

Дослідження виконано на матеріалах Державної служби статистики України [33] та порталу «Відкриті дані» [34]. Всього розглянуто 50 показників, що стосуються пенсійного забезпечення, соціально-економічного розвитку та екології. У дослідженні використано математичні моделі та методи інтелектуального аналізу для підготовки даних, виявлення закономірностей, відбору найбільш значимих показників та для прогнозного моделювання.

Основними етапами методики є:

1. Збір та попередня обробка статистичних показників, відбір найбільш значимих для дослідження.
2. Типологічне групування вхідних даних.
3. Зменшення простору вхідних змінних (метод головних компонент) для кожної групи даних.
4. Відбір регресорів та побудова регресійних моделей.
5. Оцінювання якості прогнозів, побудованих за різними моделями, вибір кращої моделі.
6. Побудова мережі Байєса для оцінювання різних сценаріїв впливу екологічних чинників на зміну витрат на пенсійне забезпечення.
7. Аналіз побудованих сценаріїв.
8. Підготовка варіантів управлінських рішень.

Результати дослідження

Вхідними даними є державні видатки на соціальний захист та соціальне забезпечення, макроекономічні показники, такі як валовий внутрішній продукт, індекс споживчих цін, демографічні показники та екологічні – стан забруднення повітря різними органічними та неорганічними сполуками в розрахунку на одну особу населення. Перелік показників представлений у табл. 1.

Таблиця 1. Перелік змінних, використаних у роботі

Позначення	Показник
Target	Державні видатки на соціальний захист та соціальне забезпечення, % до ВВП
x01	Всього, населення тис. чол.
x02	Середня очікувана тривалість життя при народженні (обидві статі), роки
x03	Індекс споживчих цін, %
x04	Валовий внутрішній продукт, тис. грн
x05	Чисельність пенсіонерів (на початок року), тис. чол.
x06	Чисельність одержувачів пенсій за віком, тис. чол.
x07	Чисельність одержувачів пенсій по інвалідності
x08	Чисельність одержувачів пенсій по втраті годувальника
x09	Чисельність одержувачів пенсій за вислугу років, тис. чол.
x10	Чисельність одержувачів соціальних пенсій, тис. чол.
x11	Чисельність отримувачів довічного утримання суддів, чол.
x12	Викиди речовин у вигляді суспендованих твердих частинок на одну особу, кг
x13	Викиди речовин у вигляді суспендованих твердих частинок більше 2,5 мкм та менше 10 мкм на одну особу, кг
x14	Викиди речовин у вигляді суспендованих твердих частинок 2,5 мкм та менше на одну особу, кг
x15	Викиди діоксиду азоту на одну особу, кг
x16	Викиди аміаку на одну особу, кг
x17	Викиди діоксиду сірки на одну особу, кг
x18	Викиди оксиду вуглецю на одну особу, кг
x19	Викиди неметанових летких органічних сполук на одну особу, кг

Фактори, представлені в табл. 1, було розподілено на три групи, що відповідають напрямкам – пенсійне забезпечення, макроекономіка, екологія. Для групи змінних, що описують чисельність одержувачів пенсій (змінні x_5–x_{11}) методом головних компонент [35], зменшується простір вхідних змінних. Визначено, що два головних компоненти описують 80% інформації про пенсійне забезпечення.

Для групи екологічних чинників (змінні x_{12}-x_{19}), використовуючи метод головних компонент, визначено дві змінні, вони описують 80% інформації про викиди шкідливих речовин та забруднення повітря.

Для визначення характеру залежностей у даних, що використовуються у дослідженні, була побудована серія регресійних моделей. Змінні, які мають бути використані для побудови фінальної моделі, визначено на основі значення *p*-значення статистики *xi*-квадрат для оцінок коефіцієнтів моделі у вигляді лінійної регресії. В якості критичного значення міри рівня довіри використано 10%.

Кращою серед побудованих моделей виявилась модель, регресорами якої є середня очікувана тривалість життя при народженні (обидві статі) (x_2), індекс споживчих цін (x_3), валовий внутрішній продукт (x_4), головна компонента, що описує кількість одержувачів пенсій (Pension), головна компонента, яка описує макроекономічні показники (Econom), головна компонента (Emissions) описує екологічні чинники. Статистичні характеристики побудованої моделі: коефіцієнт детермінації R^2 = 0,81, скорегований коефіцієнт детермінації R^2 = 0.71, середньоквадратична похибка RMSE = 0,52, середня абсолютна процентна похибка MAPE = 4,92%. Тобто, відхилення реального значення від прогнозованого становить, в середньому, 4,92% (рис. 1).

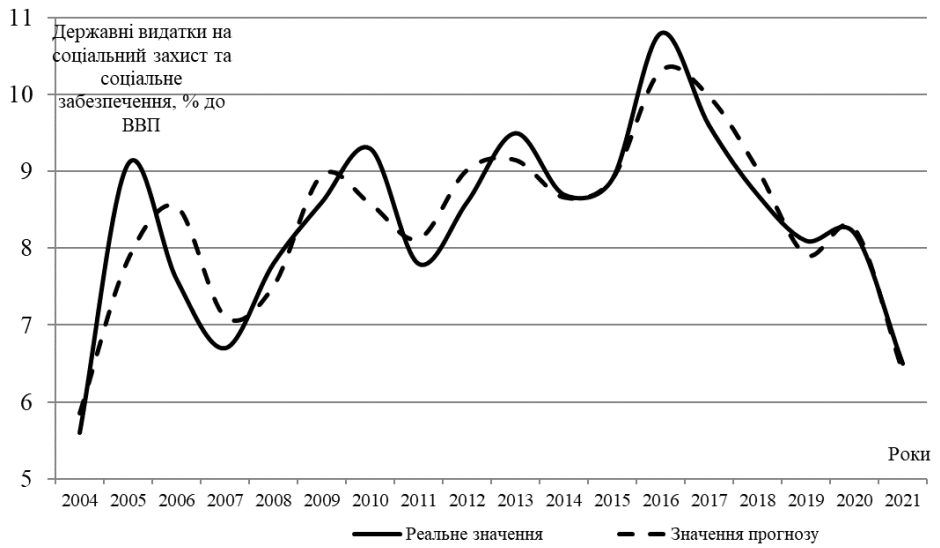

Рис. 1. Прогнозовані та реальні значення показника частки державних видатків на соціальний захист та соціальне забезпечення у ВВП, %

Отже, як видно з табл. 2, усі значення Pr > |t| менші 0,1, тобто при значенні міри рівня довіри 10% модель валідна.

Для побудови мережі Байєса (рис. 2) було використано програмне забезпечення Genie 2 [36]. Побудова топології мережі здійснювалась в автоматичному режимі із використанням алгоритму Greedy Thick Thinning (модифікація алгоритму K2).

Були накладені обмеження: кожна вершина-нащадок може мати не більше восьми батьківських вершин. Статистичні характеристики побудованої моделі – похибка помилкової класифікації (MISC) дорівнює 20%.

Таблиця 2. Значення оцінок коефіцієнтів моделі

Оцінка параметра моделі					
Змінна	Кількість ступенів свободи	Оцінка параметра моделі	Стандартна похибка	Значення *t*-статистики Стьюдента	*p*-значення статистики *xi*-квадрат
Вільний член	1	71.28906	26.36255	2.70	0.0205
x_2	1	-0.73005	0.34617	-2.11	0.0587
x_3	1	-0.07780	0.02516	-3.09	0.0103
x_4	1	-0.00000147	4.53494E-7	-3.24	0.0079
Pension	1	-1.46774	0.70558	-2.08	0.0617
Econom	1	1.36263	0.40293	3.38	0.0061
Emissions	1	-1.92650	0.61879	-3.11	0.0099

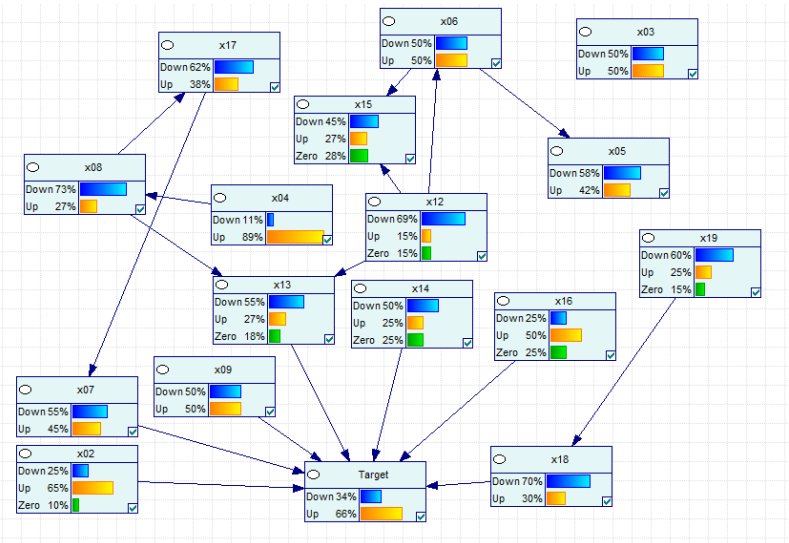

Рис. 2. Топологія мережі Байєса для моделювання сценаріїв впливу забруднення повітря на зростання обсягів видатків на пенсійну складову витрат на соціальний захист та соціальне забезпечення

Як показало дослідження, виконане із використанням мережі Байєса, представленої на рис. 2, можна розглянути декілька сценаріїв. Один із варіантів – збільшені кількості одержувачів пенсій по інвалідності за збільшення рівня забруднення повітря речовинами у вигляді суспендованих твердих частинок різного розміру. В цьому випадку, ймовірність збільшення частки державних видатків на соціальний захист та соціальне забезпечення у ВВП збільшується до 75%. За іншого, менш ймовірного сценарію, при зменшенні рівня забруднення повітря речовинами у вигляді суспендованих твердих частинок різного розміру, кількості пенсій по інвалідності, по втраті годувальника ймовірність зменшення частки державних видатків на соціальний захист та соціальне забезпечення у ВВП зменшується до 49%.

Висновки та перспективи подальших досліджень

Проведення актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування врегульоване чинним законодавством та базується на поєднанні методів теорії імовірності, математичного аналізу, математичної статистики, теорії стохастичного аналізу, теорії складних відсотків, диференційних рівнянь та оптимізації. В даному дослідженні запропоновано нову методику прогнозного моделювання, яка передбачає використання ансамблів моделей, може бути застосована у існуючій системі аналізу та прогнозування витрат на соціальний захист та соціальне забезпечення, в тому числі й для проведення актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування. Як свідчать результати виконаного дослідження, прийнятних результатів прогнозування державних витрат на соціальний захист та соціальне забезпечення можна досягти завдяки застосуванню запропонованої методики, в основу якої покладено ансамбль з регресійних моделей та байєсівських мереж.

Серед регресійних моделей кращою виявилась комбінована модель із включенням регресорів за методом backward selection [5], коли з повної моделі вилучаються регресори, для яких *p*-значення статистики *xi*-квадрат більше за 0,1. Причому для пониження розмірності груп вхідних регресорів – кількість одержувачів пенсій, макроекономічні показники, екологічні чинники, було використано метод головних компонентів, за результатами роботи якого в модель було включено отримані головні компоненти, які враховують варіабельність даних більшу за 80%. Для побудованої фінальної моделі у вигляді регресії на головних компонентах отримано статистичні характеристики – коефіцієнт детермінації 0,81, середньоквадратична похибка 0,52, середня абсолютна процентна похибка 4,92%. Тобто в середньому на кожному кроці при прогнозуванні за отриманою моделлю реальне значення відрізняється від прогнозу на 4,92%.

У подальших дослідженнях для прогнозного моделювання доцільно застосовувати ансамблі різнотипних моделей, зокрема регресійні моделі та мережі Байєса у поєднанні, наприклад, з методом групового врахування аргументів, штучним інтелектом та іншими методами і моделями інтелектуального аналізу даних.

СПИСОК ЛІТЕРАТУРИ

1. Дубініна С.В., Бідюк П.І. (2017). Застосування методів інтелектуального аналізу даних до розв'язання задач актуарного моделювання та оцінювання фінансових ризиків. *Системні дослідження та інформаційні технології.* 1. 49–64 doi: 10.20535/SRIT.2308-8893.2017.1.04

2. 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

3. Larochelle J.-P., Carlson P., Cote V. C., Lu Y., Shapiro N., Tam A., Thusu V., Zhang A. (2023). *Predictive Analytics and Machine Learning. Practical Applications for Actuarial Modeling (Nested Stochastic)*. Schaumburg: Society of Actuaries. Research Institute. Retrieved from https://www.soa.org/49ae74/globalassets/assets/files/resources/research-report/2023/predictive-analytics-and-machine-learning.pdf

4. Iyer S. Stochastic Actuarial Modelling of a Defined-Benefit Social Security Pension Scheme: An Analytical Approach. (2008) *Annals of Actuarial Science.* 3(1-2), 127-185. https://doi.org/10.1017/S174849950000049X
5. McCrea R., King R., Graham L., Börger L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution.*Vol.14, Issue 1, 4-11. doi: 10.1111/2041-210X.14050
6. Frees E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. New York: Cambridge University Press.
7. Gupta R. Y., Mudigonda S. S., Baruah P. K., Kandala P. K. (2020). 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).* Vol. 9, Issue 3. 699-706. doi: 10.35940/ijrte.C4686.099320
8. Караєва Н. В. (2018). Методологічні аспекти та програмні засоби оцінки ризику здоров'ю населення при несприятливому впливі факторів навколишнього середовища. *Системи управління, навігації та зв'язку*. Вип. 1(47), 164-169. doi: 10.26906/SUNZ.2018.1.164
9. Естілл Я. Вплив на здоров'я та соціальні витрати, пов'язані із забрудненням повітря у великих містах України. United Nations Development Programme. Звіт. Retrieved from: https://www.undp.org/sites/g/files/zskgke326/files/2023-03/Health%20impacts%20and%20social%20costs%20associated%20with%20air%20pollution%20in%20larger%20urban%20areas%20of%20Ukraine%20%28UA%29.pdf
10. Проєкт «Всесвітній індекс якості повітря». Retrieved from: https://aqicn.org та https://waqi.info.
11. González Parra, G., Arenas, A. J. (2014). A mathematical model for social security systems with dynamical systems. *Ingeniería Y Ciencia,* 10(19), 33–53. doi: 10.17230/ingciencia.10.19.2
12. Iyer S. (1999). *Actuarial mathematics of social security pensions.* International Labour Organization. Retrieved from: https://www.issa.int/sites/default/files/documents/publications/2Actuarial_mathematics_of_ss_pensions_en-29172.pdf
13. Lähderanta T., Salonen J., Möttönen J., Sillanpää M.J. (2022). Modelling old-age retirement: An adaptive multi-outcome LAD-lasso regression approach. *International Social Security Review*: Vol. 75, Issue 1. 3-29. doi: 10.1111/issr.12287
14. Black E., Lattyak C. G., Chairperson V., Stone L. K. (2023). *Senior Pension Fellow Modeling – for Pension Actuaries.* Washington: The American Academy of Actuaries. Retrieved from: https://www.actuary.org/sites/default/files/2023-01/Modeling_Practice_Note.pdf
15. Global Burden of Disease. Retrieved from: https://www.healthdata.org/research-analysis/gbd
16. Постанова Кабінету Міністрів України від 16 грудня 2004 р. № 1677 «Методика проведення актуарних розрахунків у системі загальнообов'язкового державного пенсійного страхування». Retrieved from: https://www.kmu.gov.ua/npas/10301286
17. Trofymchuk, O., Bidiuk, P., Terentiev, O., Prosyankina-Zharova, T. (2019). *Decision Support Systems for Modelling, Forecasting and Risk Estimation*. Riga: LAP Lambert Academic Publishing.
18. Shapovalenko, N. (2021). A Suite of Models for CPI Forecasting. *Visnyk of the National Bank of Ukraine,* 252, 4-36. doi: 10.26531/vnbu2021.252.01
19. Бідюк П. І., Романенко В. Д., Тимощук О. Л. (2013). *Аналіз часових рядів.* Київ: НТУУ КПІ.
20. Seber G. A. F., Wild C. J. (1989). *Nonlinear Regression.* New York: John Wiley & Sons, Inc.
21. Hurvich C. M., Simonoff J S., Tsai C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society.* Vol. 60, Issue 2, 271–293.
22. Fahrmeir L., Kneib T., Lang S., Marx B. D. (2021). *Regression Models, Methods and Applications.* Berlin: Springer-Verlag. doi: 10.1007/978-3-662-63882-8

23. Abubakari A. G. (2022). Actuarial Measures, Regression, and Applications of Exponentiated Fr´echet Loss Distribution. *International Journal of Mathematics and Mathematical Sciences*. Vol. 2022 1-17. https://doi.org/10.1155/2022/3155188
24. Thrane C. (2020). *Applied Regression Analysis Doing, Interpreting and Reporting*. New York: Taylor & Francis Group. https://doi.org/10.4324/9780429443756
25. Григорків М.В. (2020). *Динамічні моделі еколого-економічних систем в умовах соціально-економічної кластеризації: монографія.* Тернопіль: «Економічна думка ТНЕУ».
26. Pearl, J. (2000). *Causality: models, reasoning, and inference.* Cambridge University Press.
27. Jensen, F.V. (2001). *Bayesian networks and decision graphs.* New York: Springer. doi:10.1007/978-1-4757-3502-4
28. Spiegelhalter, D., Dawid, P., Lauritzen, S., Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8 (3), 219–247.
29. Lauritzen, S. L., Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal Royal Statistics Society, series B (Methodology).* 50 (2), 157-194.
30. Spirtes, P., Glymour, C., Scheines, R. (1993). *Causation, Prediction and Search. Part of the book series: Lecture Notes in Statistics* (LNS, vol. 81). Berlin: Springer Verlag. doi:10.1007/978-1-4612-2748-9
31. Spirtes, P., Glymour C., Scheines, R. (1991). From probability to causality. *Philosophical Studies*, 64, 1–36. doi:10.1007/BF00356088
32. Згуровський, М.З., Бідюк, П.І., Терентьєв, О.М., Просянкіна-Жарова Т.І. (2015). *Байєсівські мережі у системах підтримки прийняття рішень.* Київ: ТОВ «Видавниче Підприємство «Едельвейс».
33. Державна служба статистики України. Навколишнє природне середовище. Retrieved from: https://www.ukrstat.gov.ua
34. Щоденні та щомісячні спостереження за забрудненням атмосферного повітря. Retrieved from: https://diia.data.gov.ua/
35. Чугаєвська С.В., Ковтун Н.В. (2022). *Основи статистичного моделювання: навч. посібник.* Житомир: Видавництво ПП "Рута".
36. Genie 2.0. Retrieved from https://www.bayesfusion.com/genie/

Стаття надійшла до редакції 17.05.2024 і прийнята до друку після рецензування 23.08.2024

REFERENCES

1. Dubinina, S.V., & Bidiuk, P.I. (2017). Zastosuvannya metodiv intelectualnogo analizu danyh intelectualnogo analizu do rozv'azannya zadach aktuarnogo modeluvannya ta otzinuvannya finansovyh ryzykiv. *Systemny doslidgenya ta infornatsiyni tehnologii*, 1, 49–64. https://doi.org/10.20535/SRIT.2308-8893.2017.1.04 [in Ukrainian].
2. 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
3. Larochelle, J.-P., Carlson, P., Cote, V. C., Lu, Y., Shapiro, N., Tam, A., Thusu, V., & Zhang, A. (2023). *Predictive Analytics and Machine Learning. Practical Applications for Actuarial Modeling (Nested Stochastic).* Schaumburg: Society of Actuaries. Research Institute. Retrieved from https://www.soa.org/49ae74/globalassets/assets/files/resources/research-report/2023/predictive-analytics-and-machine-learning.pdf
4. Iyer, S. (2008). Stochastic Actuarial Modelling of a Defined-Benefit Social Security Pension Scheme: An Analytical Approach. *Annals of Actuarial Science*, 3(1-2), 127-185. https://doi.org/10.1017/S174849950000049X

5. McCrea, R., King, R., Graham, L., & Börger, L. (2023). Realising the promise of large data and complex models. *Methods in Ecology and Evolution*, 14(1), 4-11. https://doi.org/10.1111/2041-210X.14050
6. Frees, E.W. (2010). *Regression Modeling with Actuarial and Financial Applications.* New York: Cambridge University Press.
7. Gupta, R.Y., Mudigonda, S.S., Baruah, P.K., & Kandala, P.K. (2020). 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)*, 9(3), 699-706. https://doi.org/10.35940/ijrte.C4686.099320
8. Karaeva, N.V. (2018). Metodologichny aspekty ta programni zasoby orzinky riziku zdorov'u naselennya pry nespriyatlyvomu vplivy daktoriv navkolyshnyogo seredovysa. *Systemy upravlinnya, navigatzii ta zv'yazku*, 1(47), 164-169. https://doi.org/10.26906/SUNZ.2018.1.164 [in Ukrainian].
9. Estill, Ya. Vplyv na zdorov'ya ta sotzialny vytraty, pov'yazany iz zabrudnennyam povitrya u velykyh mistah Ukrainy. United Nations Development Programme. Report. Retrieved from: https://www.undp.org/sites/g/files/zskgke326/files/2023-03/Health%20impacts%20and%20social%20costs%20associated%20with%20air%20pollution%20in%20larger%20urban%20areas%20of%20Ukraine%20%28UA%29.pdf [in Ukrainian].
10. Proekt «Vsesvitniy index yakosti povitrya». Retrieved from: https://aqicn.org and https://waqi.info [in Ukrainian].
11. González Parra, G., & Arenas, A.J. (2014). A mathematical model for social security systems with dynamical systems. *Ingeniería Y Ciencia,* 10(19), 33–53. https://doi.org/10.17230/ingciencia.10.19.2
12. Iyer, S. (1999). *Actuarial mathematics of social security pensions.* International Labour Organization. Retrieved from: https://www.issa.int/sites/default/files/documents/publications/2Actuarial_mathematics_of_ss_pensions_en-29172.pdf
13. Lähderanta, T., Salonen, J., Möttönen, J., & Sillanpää, M.J. (2022). Modelling old-age retirement: An adaptive multi-outcome LAD-lasso regression approach. *International Social Security Review*, 75(1), 3-29. https://doi.org/10.1111/issr.12287
14. Black, E., Lattyak, C.G., Chairperson, V., & Stone, L.K. (2023). *Senior Pension Fellow Modeling – for Pension Actuaries.* Washington: The American Academy of Actuaries. Retrieved from: https://www.actuary.org/sites/default/files/2023-01/Modeling_Practice_Note.pdf
15. Global Burden of Disease. Retrieved from: https://www.healthdata.org/research-analysis/gbd
16. Postanova Kabinetu Ministriv Ukrainy vid 16 grudnya 2004 r. № 1677 «Metodyka provedennya aktuarnyh rozrahunkiv u systemi zagaknoobov'yazkovogo derzavnogo pebsiynogo strahuvanya». Retrieved from: https://www.kmu.gov.ua/npas/10301286
17. Trofymchuk, O., Bidiuk, P., Terentiev, O., & Prosyankina-Zharova, T. (2019). *Decision Support Systems for Modelling, Forecasting and Risk Estimation.* Riga: LAP Lambert Academic Publishing.
18. Shapovalenko, N. (2021). A Suite of Models for CPI Forecasting. *Visnyk of the National Bank of Ukraine*, 252, 4-36. https://doi.org/10.26531/vnbu2021.252.01
19. Bidiuk, P.I., Romanenko, V.D., & Tymoszhuk, O.L. (2013). *Analyz chasovyh ryadiv.* Kyiv: NTUU KPI [in Ukrainian].
20. Seber, G.A.F., & Wild, C.J. (1989). *Nonlinear Regression.* New York: John Wiley & Sons, Inc.
21. Hurvich, C.M., Simonoff, J.S., & Tsai, C.L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society,* 60(2), 271–293.
22. Fahrmeir, L., Kneib, T., Lang, S., & Marx, B.D. (2021). *Regression Models, Methods and Applications.* Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-662-63882-8

23. Abubakari, A.G. (2022). Actuarial Measures, Regression, and Applications of Exponentiated Fr´echet Loss Distribution. *International Journal of Mathematics and Mathematical Sciences*, 2022, 1-17. https://doi.org/10.1155/2022/3155188
24. Thrane, C. (2020). *Applied Regression Analysis Doing, Interpreting and Reporting*. New York: Taylor & Francis Group. https://doi.org/10.4324/9780429443756
25. Grygorkiv, M.V. (2020). *Dynamichny modely ekologo-ekonomichnyh system v umovah sotzialno-ekonomichnoi klasteryzatzii: monographia*. Ternopil: «Ekonomichna dumka. TNEU» [in Ukrainian].
26. Pearl, J. (2000). *Causality: models, reasoning, and inference.* Cambridge University Press.
27. Jensen, F.V. (2001). *Bayesian networks and decision graphs.* New York: Springer. https://doi.org/10.1007/978-1-4757-3502-4
28. Spiegelhalter, D., Dawid, P., Lauritzen, S., & Cowell, R. (1993). Bayesian analysis in expert systems. *Statistical Science*, 8 (3), 219–247.
29. Lauritzen, S. L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal Royal Statistics Society, series B (Methodology)*, 50 (2), 157-194.
30. Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction and Search. Part of the book series: Lecture Notes in Statistics* (LNS, vol. 81). Berlin: Springer Verlag. https://doi.org/10.1007/978-1-4612-2748-9
31. Spirtes, P., Glymour, C., & Scheines, R. (1991). From probability to causality. *Philosophical Studies*, 64, 1–36. https://doi.org/10.1007/BF00356088
32. Zgurovs'kyj, M.Z., Bidjuk, P.I., Terent'jev, O.M., & Prosjankina-Zharova, T.I. (2015). *Bajjesivs'ki merezhi u systemah pidtrymky pryjnjattja rishen'.* Kyi'v: TOV «Vydavnyche Pidpryjemstvo «Edel'vejs» [in Ukrainian].
33. Derzavna sluzba statystyky Ukrainy. Navkolyshnye pryrodne seredovyshe. Retrieved from: https://www.ukrstat.gov.ua [in Ukrainian].
34. Shodenny ta shomisyachni sposteregenya za zabrudnennyam atmosfernogo povitrya. Retrieved from: https://diia.data.gov.ua/ [in Ukrainian].
35. Chugaevska, S.V., & Kovtun, N.V. (2022). *Osnovy statystychnogo modeluvannya: navch. posibnyk.* Gytomyr: Vydavnytstvo PP "Ruta" [in Ukrainian].
36. Genie 2.0. Retrieved from https://www.bayesfusion.com/genie/

The article was received 17.05.2024 and was accepted after revision 23.08.2024

Зарудний Олексій Борисович
здобувач, Інститут телекомунікацій і глобального інформаційного простору НАН України
Адреса робоча: 03186 Україна, м. Київ, Чоколівський бульвар, 13
ORCID ID: https://orcid.org/0009-0008-7462-3899 ***e-mail:*** oleksii.zarudnyi@gmail.com