УДК 004.415.3:681.6

**Козир С.В., аспірант**
**Науковий керівник: Молоканова В.М., д.т.н., професор кафедри системного аналізу та управління**
(*Національний технічний університет "Дніпровська політехніка", м. Дніпро, Україна*)

# АЛГОРИТМ ПОДІЛУ ПОЧАТКОВОЇ ВИБІРКИ ПРИ ПРОГНОСТИЧНОМУ МОДЕЛЮВАННІ

Прогнозування найчастіше використовується на етапах аналізу реалізації стратегій розвитку організації через проекти. При реалізації проектів прогнозування використовується для оцінки можливих результатів їх відхилень від цілей. При цьому важливо оцінити ефективність передбачення побудованої моделі і особливо при прогнозуванні тестових зразків, які не використовувались при навчанні моделі. Однією із метрик валідності моделі може бути $Q_2$. Ця метрика аналогічна коефіцієнту детермінації $R^2$, але з тією різницею, що $Q_2$ використовується до контрольної послідовності замість навчальної ($R^2$) [1].

$$Q_2 = 1 - \frac{\sum_{i=1}^{n_t}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n_t}(Y_i - \bar{Y})^2}$$

де $n_t$ – кількість точок у контрольній послідовності, $\widehat{Y}_i$ – розраховане значення шуканої залежності в $i$-й точці контрольної послідовності, $\bar{Y}$ – середнє значення шуканої величини. Коли $Q_2$ наближається до 1, то якість моделі підвищується.

Коли ретроспектива незначна, та й ще представлена нестаціонарними часовими рядами, то досягти високих значень $Q_2$ складно. У випадку малої кількості точок експерименту для побудови прогностичних моделей може бути застосований метод групового врахування аргументів (МГВА). Апробація прогностичної моделі собівартості товарної вугільної продукції на відомих подіях минулого (46 точок передісторії) відбувалась із залученням бібліотеки Python GmdhPy, яка реалізує ітераційний МГВА з опорними функціями. Отримані моделі також відомі як самоорганізуючі поліноміальні нейронні мережі глибокого навчання. Коефіцієнти детермінації $R^2$ моделей із 4-ма та 9-ма предикторами отримано значно менше 1 [2]. Тому пошук прогностичних моделей із надійними характеристиками є актуальною задачею.

Нехай початкові дані зосереджені в матриці $A = (X_1, X_2, \dots, X_n, Y)$, де $X_i, i = \overline{1,n}$, і $Y$ – вектори-стовпчики розмірністю $m$, $X_i$ – вхідні фактори, $Y$ – вихідна характеристика. Задача полягає в ідентифікації залежності $Y = F(X_1, X_2, \dots, X_n)$ поліномом Колмогорова-Габора : $Y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i,j=1}^{n} a_{ij} x_i x_j + \sum_{i,j,k=1}^{n} a_{ijk} x_i x_j x_k + \dots$

Реалізація МГВА, в більшості випадків, пов'язана з необхідністю поділу генеральної сукупності даних на три вибірки – навчальну, тестову та контрольну. Найбільш поширеним, але не єдиним, є підхід, при якому в навчальну послідовність вибирають точки експериментів з великим значенням дисперсії, а в тестову та контрольну – з меншими. Це пояснюється тим, що область навчання повинна бути якнайширшою, а контрольні точки, в більшості своїй, знаходитися всередині неї [3].

*Алгоритм поділу* є таким:

*Крок 1*. Визначити процентне співвідношення між кількістю елементів у навчальній, тестовій та контрольної послідовностях.

*Крок 2*. Для кожного стовпчика $X_i, i = \overline{1,n}$, розрахувати середнє значення його елементів

$$x_{i_{\text{сер}}} = \frac{1}{m}\sum_{j=1}^{m} x_{ji},$$

та отримати середнє значення безлічі образів $(x_{1_{\text{сер}}}, x_{2_{\text{сер}}}, \dots \ , \ x_{n_{\text{сер}}})$

*Крок 3*. Знайти вибіркові дисперсії для кожного рядка таблиці за формулою

$$D_j = \frac{1}{n-1}\sum_{j=1}^{m}(x_{ji} - x_{i_{\text{сер}}})^2 \ , \ j = \overline{1,m},$$

*Крок 4*. Для впорядкування таблиці переставити рядки так, щоб першим був рядок з найбільшим значенням дисперсії, а останній – з найменшим.

*Крок 5*. Відповідно до результату кроку 1 розділити дані в таблиці на навчальну, тестову та контрольну послідовності.

В цьому дослідженні МГВА із застосуванням алгоритму поділу реалізовано в MATLAB (рис. 1).

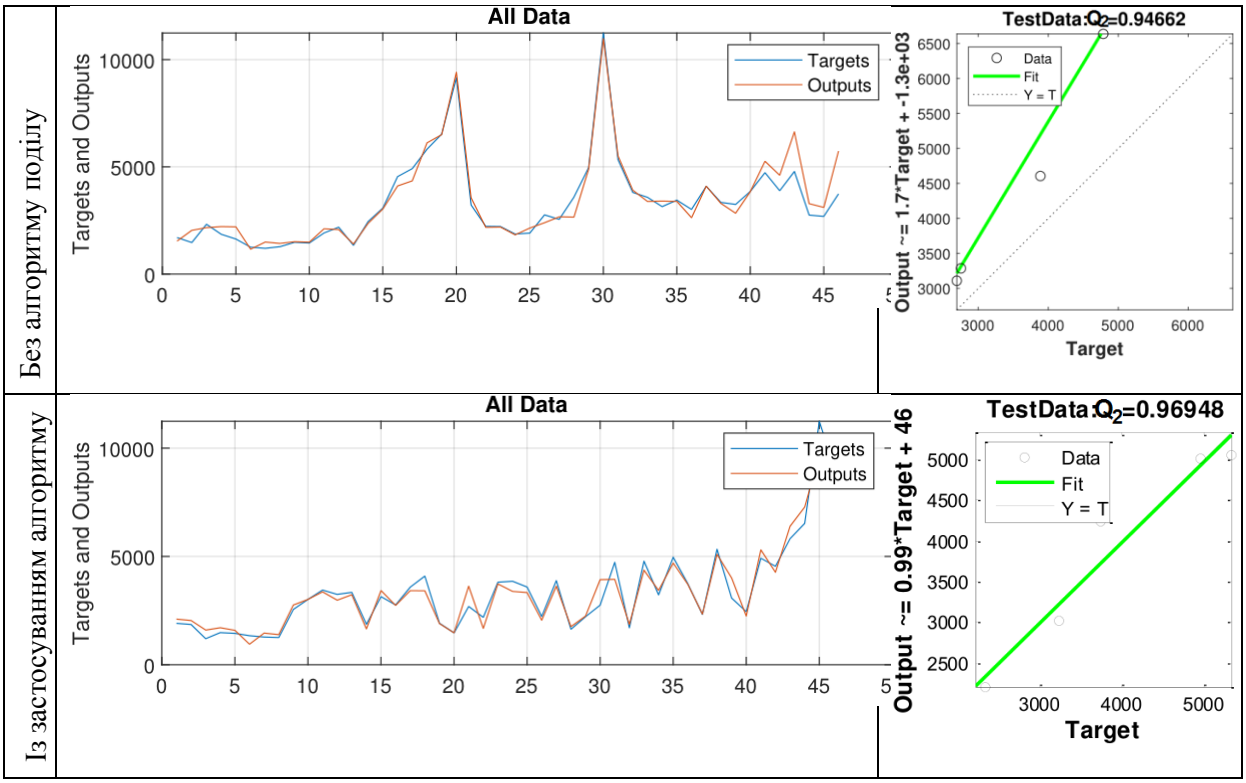


Рисунок 1 – Візуалізація результатів обчислень для всієї вибірки та графіки регресії для тестової вибірки.

**Висновки**. Задля покращення ефективності передбачення при побудові прогностичних моделей реалізації проектів краще застосовувати алгоритм поділу вибірки за дисперсіями. Після чого дані часових рядів упорядкувати в природний часовий порядок і реалізувати прогнозування.

## Перелік посилань

1. Majdi I. Radaideh, Tomasz Kozlowski, ”Analyzing nuclear reactor simulation data and uncertainty with the group method of data handling,” Nuclear Engineering and Technology, Vol. 52, Issue 2, 2020, pp. 287-295, ISSN 1738-5733.

2. Algorithm for Selecting and Comparing of Situations Features of Intelligent Decision-Making Support System/ S. Kozyr ,V. Sliesariev // 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), 2021, pp. 657-661, DOI: 10.1109/ACIT52158.2021.9548528

3. Снитюк В. Є. Прогнозування. Моделі, методи, алгоритми. [Текст]: Навч. посібник. /В.Є. Снитюк. – К.: «Маклаут», 2008. – 364 с.