УДК 519.688

К.т.н., доцент Люшенко Л.А., студент Васильковський К.В.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

МЕТОД ПРОГРАМНОГО ПРОГНОЗУВАННЯ МАКРОЕКОНОМІЧНИХ ПОКАЗНИКІВ НА ОСНОВІ КОРЕЛЯЦІЙНОГО АНАЛІЗУ

Abstract

Lesya A. Liushenko, assoc. prof., PhD; Konstantin Vasilkovskiy, student
Method of program forecasting macroeconomic indicators based on correlation analysis

This article presents a method of program forecasting of macroeconomic indicators using expert judgments. Static data are consistent with the overall time scale. Stationarity and seasonality are evaluated, and automatic expert selection of a forecasting method based on ARIMA, SARIMA, or ensemble trees is performed. Forecasting is a two-step process: first, the base variables are predicted, and then, based on them, the derived macro indicators are predicted using correlation models. The accuracy of the forecast is estimated by comparing it with the macroindicators actually obtained, and then the forecasting models are refined.

Вступ

Проблемою існуючих методів прогнозування економічних макропоказників є відсутність універсального підходу. Це пов’язано з динамічним розвитком економік різних країн, впливом кризових економічних, політичних та технологічних чинників. Тому для підвищення точності і швидкодії прогнозування необхідно впроваджувати експертний підхід для вибору методу прогнозування економічних макропоказників.

Методи прогнозування такі, як ARIMA, SARIMA, Random Forest і Gradient Boosting мають суттєві недоліки при прогнозуванні на різних наборах часових рядів. Методи ARIMA, SARIMA невідповідно працюють на рядах з значною нелінійністю та високою волатильністю. . В свою черго Random Forest і Gradient Boosting мають значне зниження швидкодії на лінійних, сталих наборах даних.

У роботі подано метод і програмне забезпечення на основі статистичного аналізування: приведення рядів до спільної періодичності та обробка даних з різними термінами збирання [3]. Оцінюється стаціонарність та сезонність, а також виконується автоматичний експертний вибір методу

прогнозування на основі ARIMA, SARIMA або Random Forest і Gradient Boosting [1, 2, 5].

Постановка задачі

Метою роботи є розроблення методу для програмного прогнозування макроекономічних показників на основі статистичного та кореляційного аналізування. Цей метод передбачає уніфікацію періодичності та приведення до спільної часової шкали статистичних даних, побудову кореляційних моделей, автоматичний вибір методів прогнозування ARIMA, SARIMA, SARIMAX, Random Forest та Gradient Boosting [1, 5].

Прогнозування відбувається з наступних етапів: виявляються кореляційні залежності та базові змінні; вибирається метод прогнозування та прогнозуються базові змінні, а потім на їх основі прогнозуються похідні макропоказники за допомогою кореляційних моделей. Точність прогнозу оцінюється шляхом порівняння його з фактично отриманими макропоказниками, а потім уточнюються моделі прогнозування.

Термінологія

Кореляційний відбір змінних — процедура вибору предикторів на основі статистично значущих зв’язків із цільовим рядом. Використовуються коефіцієнти Пірсона та Спірмана для оцінки лінійної та рангової залежностей з урахуванням лагових зсувів.

Ансамблеві моделі — методи машинного навчання, що поєднують велику кількість базових моделей, найчастіше дерев рішень. До таких моделей належать Random Forest і Gradient Boosting, які дозволяють відтворювати складні нелінійні взаємозв’язки між змінними.

Бутстреп-агрегація (bagging) — метод ансамблювання, що полягає у побудові кількох моделей на різних випадкових підвибірках з поверненням (бутстреп-зразках) із навчальної вибірки. Отримані передбачення об'єднуються шляхом «голосування», або усереднення.

Статистичне аналізування — сукупність методів дослідження часового ряду, що включає виявлення трендів, сезонності, лагових взаємозв’язків, стаціонарності та основних статистичних властивостей даних. У контексті статті використовується для підготовки рядів до прогнозування та обґрунтованого вибору моделі.

Існуючі методи

Моделі ARIMA, SARIMA та SARIMAX моделюють залежність поточного значення часового ряду від його лагів і похибок прогнозу. SARIMA враховує регулярну сезонність, а SARIMAX — зовнішні регресори [1]. Побудова передбачає приведення ряду до стаціонарності, ідентифікацію

порядків та оцінювання параметрів методом максимальної правдоподібності. Якість моделі перевіряють аналізом залишків, зокрема тестом Льюнга–Бокса [4].

Методи машинного навчання, такі як Random Forest і Gradient Boosting, формують ансамблі моделей на основі дерев рішень, здатні моделювати складні нелінійні залежності між змінними. Random Forest базується на бутстреп-агрегації та випадковому відборі ознак, тоді як Gradient Boosting будує дерева послідовно, мінімізуючи похибку попередніх моделей [5]. У задачах прогнозування використовують лагові та сезонні ознаки. Надмірне пристосування до навчальних даних контролюють перевіркою на послідовних часових відрізках без використання майбутніх значень.

Опис запропонованого методу

Запропоновано комбінований метод прогнозування макроекономічних показників, що реалізується з восьми етапів (рис. 1).

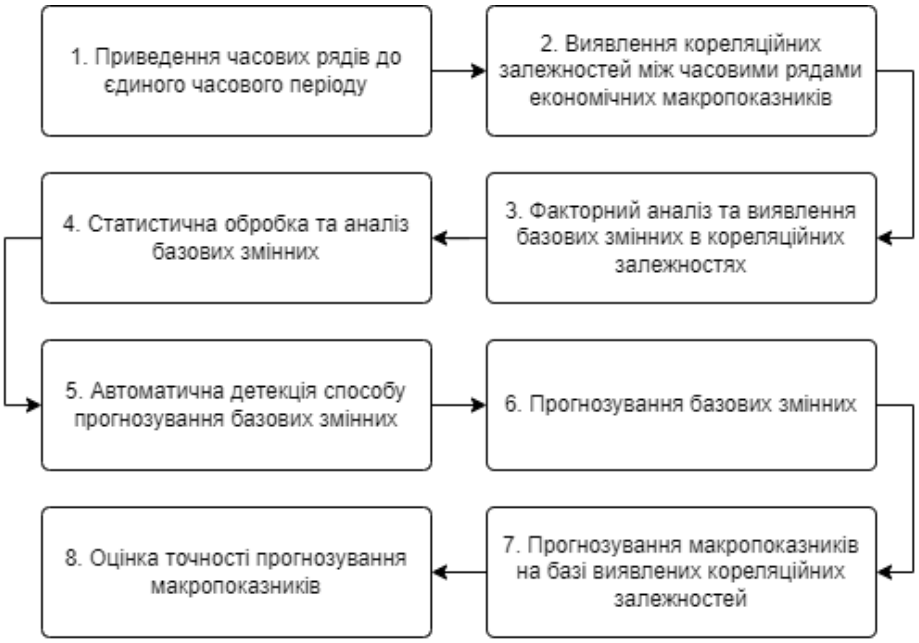

Рис. 1. Комбінований метод прогнозування макроекономічних показників

На першому етапі часові ряди приводяться до однакової періодичності. Квартальні й річні ряди трансформують у місячні зі збереженням підсумків.

Другим етапом визначаються кореляційні залежності між часовими рядами. На рядах приведених до стаціонарності обчислюються коефіцієнти кореляції Пірсона та Спірмана з урахуванням лагових зсувів. Коефіцієнт Пірсона використовується для оцінювання лінійного зв’язку між змінними, тоді як коефіцієнт Спірмана застосовують у разі порушення нормальності

розподілу або за наявності монотонної нелінійної залежності. Перехресний аналіз дозволяє уточнити часові зсуви.

На третьому етапі виявляються базові змінні при кореляційному аналізі. Проводиться факторний аналіз та оцінки, статистичної значущості вибіркових коефіцієнтів кореляції, лінійність зав'язків, побудова довірчих інтервалів.

Четвертим етапом виконують статистичний аналіз базових змінних. оцінюють варіацію, перевіряють стаціонарність і регулярну сезонність та, за потреби. Перевіряють мультиколінеарність і уточнюють лагові зсуви. Отримують набір показників для вибору методу прогнозування.

П'ятим етапом автоматично обирають метод прогнозування для базових змінних. ARIMA застосовують при лінійній динаміці. SARIMA обирають за наявності регулярної сезонності. SARIMAX використовують, коли обґрунтовано вплив зовнішніх пояснювальних змінних. За виражених нелінійностей застосовують Random Forest або Gradient Boosting. Перевірку виконують за часом без використання майбутніх значень.

Шостим етапом будують прогнози базових змінних на заданий горизонт, обраним способом на попередньому етапі.

Сьомий етап передбачає побудову прогнозу цільових показників на основі раніше зпрогнозованих базових змінних, що використовуються як зовнішні регресори. Вибір зовнішніх регресорів обґрунтовано результатами попереднього кореляційного аналізу. Використовуючи визначені базові змінні, їхні лаги, наявність тренду та сезонності, виконують побудову прогнозу на заданому горизонті.

На заключному етапі оцінюють точність прогнозу. Основним критерієм виступає MASE, додатково обчислюють sMAPE і RMSE на тестовій вибірці [2]. Для моделей типу ARIMA аналізують залишки, зокрема за допомогою тесту Льюнга–Бокса, та за потреби переглядають склад змінних та лагові структури[4].

При тестуванні порівнювався розроблений комбінований метод прогнозування з класичними методами ARIMA та SARIMA, а також Gradient Boosting і Random Forest. Для кількох щомісячних часових рядів із різною динамікою макропоказників було побудовано прогнози з горизонтом до 3 місяців. Отримані результати показують, що на заданому горизонті комбінований метод забезпечує найменше значення MASEM— 0,31, тоді як для Random Forest отримано — 0,36, для Gradient Boosting — 0,39, для SARIMA — 0,47, для ARIMA — 0,97. Це свідчить про те, що розроблений метод є більш точним для прогнозування економічних макропоказників.

За симетричною відсотковою похибкою sMAPE усереднене відхилення прогнозів від фактичних значень для комбінованого способу становило

приблизно 1,2%, тоді як для Random Forest і Gradient Boosting — близько 1,8–1,9%, для SARIMA — близько 1,4%, а для ARIMA — понад 5%.

Комбінований метод є більш витратним: середній час виконання прогнозу становить близько 1,9 с, тоді як для Random Forest — 1,0 с, для Gradient Boosting та SARIMA — близько 0,45 с, для ARIMA — 0,08 с.

Висновки

У дослідженні запропоновано комбінований підхід до прогнозування макроекономічних показників, що поєднує кореляційне аналізування, аналіз статистичних властивостей рядів та автоматизований вибір методів прогнозування (ARIMA, SARIMA, SARIMAX, Random Forest, Gradient Boosting). Підхід забезпечує врахування лагових залежностей, регулярної сезонності й мультиколінеарності. Точність прогнозу підвищується завдяки окремому моделюванню базових змінних і використанню їхніх прогнозів як зовнішніх регресорів у моделі цільового показника, попередньому кореляційному визначенню інформативних лагів, перевірці та усуненню мультиколінеарності.

Окрім того, запропонований комбінований метод програмного прогнозування економічних макропоказників за рахунок автоматичного експертного підбору методу прогнозування забезпечує підвищення точності порівняно з окремими моделями, в той же час потребує дещо більших обчислювальних витрат.

У подальших дослідженнях передбачається застосування підходу до ширшого кола макроекономічних показників і вивчення стійкості прогнозу до зміни складу регресорів.

Література

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., Ljung, G. M. Time Series Analysis: Forecasting and Control. 5th ed. Wiley, 2015.
2. Hyndman, R. J., Koehler, A. B. Another Look at Measures of Forecast Accuracy. International Journal of Forecasting, 2006, 22(4), 679–688.
3. McKinney, W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. 2nd ed., O'Reilly Media, 2018.
4. Ljung, G. M., Box, G. E. P. On a Measure of Lack of Fit in Time Series Models. Biometrika, 1978, 65(2), 297–303.
5. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009. — 745 p.