

УДК 33

DOI: 10.34670/AR.2022.31.65.002

Использование данных для когортного анализа при прогнозировании макроэкономических показателей методами машинного обучения

Калашников Владимир Андреевич

Аспирант,
ассистент департамента анализа данных и машинного обучения,
младший научный сотрудник Инжинирингового центра,
Финансовый университет при Правительстве Российской Федерации,
125167, Российская Федерация, Москва, просп. Ленинградский, 49/2;
e-mail: vakalashnikov@fa.ru

Иванов Максим Валерьевич

Магистрант,
Национальный исследовательский университет «Высшая школа экономики»,
101000, Российская Федерация, Москва, ул. Мясницкая, 20;
e-mail: mvivanov_5@edu.hse.ru

Аннотация

Актуальность темы исследования обусловлена тем, что на протяжении десятилетий прогнозирование является важным исследовательским вопросом. Для обеспечения точности построенных прогнозов ведется постоянный поиск моделей, методов и переменных, оказывающих существенное влияние на полученные результаты. В статье исследуется проблема прогнозирования различных макроэкономических показателей с использованием данных для когортного анализа. Рассматриваются различия между методами прогнозирования на данных для когортного анализа и итоговых данных. Используются алгоритмы машинного обучения на наборах данных, содержащих финансовые показатели, для проверки эффективности прогнозирования макроэкономических показателей. На основании информации из базы данных Центрального банка Российской Федерации проводится самостоятельное исследование с применением методов машинного обучения. Посредством анализа метрик RMSE и MAE показывается, что составление прогнозов на данных для когортного анализа в некоторых случаях обладает преимуществами перед традиционно используемыми пересмотренными данными. Точность прогнозов повышается за счет использования методов машинного обучения. Указывается на то, что методы машинного обучения доказали свою эффективность для больших наборов данных для когортного анализа, которые слишком многочисленны для анализа человеком.

Для цитирования в научных исследованиях

Калашников В.А., Иванов М.В. Использование данных для когортного анализа при прогнозировании макроэкономических показателей методами машинного обучения // Экономика: вчера, сегодня, завтра. 2022. Том 12. № 8А. С. 7-15. DOI: 10.34670/AR.2022.31.65.002

Ключевые слова

Когортный анализ, прогнозирование, макроэкономические показатели, машинное обучение, индексы промышленного производства.

Введение

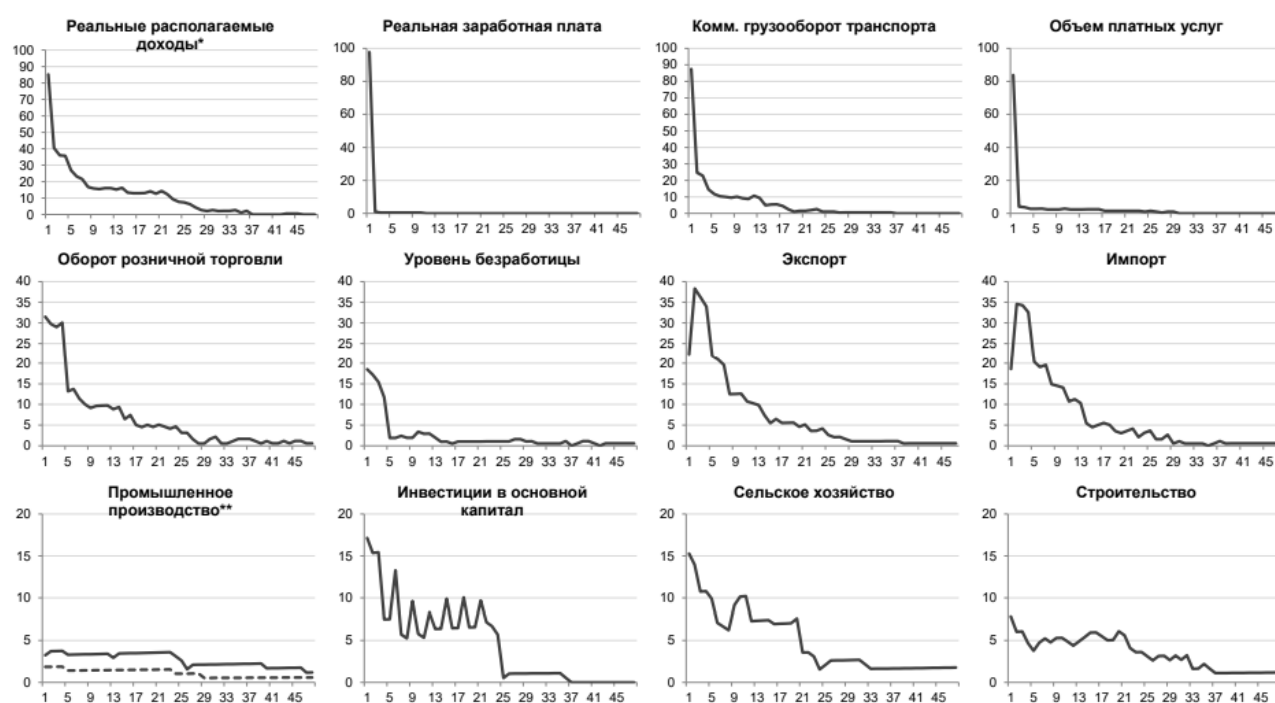
Методика выбора данных для прогнозирования различных макроэкономических показателей может оказывать существенное влияние на точность составленного прогноза. В настоящем исследовании авторы поднимают вопрос об использовании методов сбора информации, основанных на применении винтажей, т. е. данных, собранных на некоторый момент времени в прошлом. Построение моделей прогнозирования сталкивается с двумя основными проблемами: 1) выбором показателей, включаемых в анализ (обычно используются экспертные знания, но не существует методики, которая бы однозначно обосновывала первоначальный выбор показателей для анализа); 2) несбалансированным набором исторических данных, используемых для построения моделей. С быстрым развитием информационных технологий в 1990-х гг. начались интенсивные исследования по использованию методов искусственного интеллекта для прогнозирования. Классические эконометрические модели стали заменяться искусственными нейронными сетями и деревьями классификации. Инструменты машинного обучения лучше справляются с быстрорастущими наборами необработанных данных, т. е. содержащими ошибочные или неполные данные.

Для построения классических прогнозов в «реальном времени» чаще всего применяют предикторы, которые не соответствуют реальным показателям, используя неполную информацию о текущих экономических условиях. Многие ключевые статистические данные публикуются с задержками и часто пересматриваются. В результате модель показывает менее точные результаты, приближенные к фактическим показателям. Данная проблема может быть измерена количественно с помощью сравнения метрик качества прогноза. В зависимости от частоты пересмотров и горизонта прогнозирования влияние данных факторов снижает эффективность прогноза на 7-16% [Mamedli, Shibitov, www]. Большая часть краткосрочных экономических показателей (КЭП) могут пересматриваться в течение первого месяца с даты публикации, а некоторые – в течение двух лет после нее (рис. 1).

При этом различия исходных и пересмотренных значений достигают 3% в течение 48 месяцев, что может привести к серьезным отклонениям итоговых расчетов. Чаще всего пересматривают ВВП – ключевой макроэкономический показатель [Астафьева, Турунцева, 2021].

Можно сделать вывод о том, что в псевдореальном времени прогнозы могут быть нереалистичными, внушая при этом ложную уверенность. Впервые решение этой проблемы было предложено Ф. Дентоном и Дж. Куйпером в 1965 г. Исследователи в процессе анализа канадской экономики сделали вывод о том, что построение прогнозов на основе последней доступной информации приводит к серьезным отклонениям результатов прогноза [Denton,

Kuiper, 1965]. Затем У. Тривеллато и Э. Ретторе, исследуя макроэкономические показатели Италии, доказали наличие серьезных последствий из-за ошибок в модели одновременных уравнений. Они пришли к выводу о том, что использование данных для когортного анализа зачастую приводит к получению более точных прогнозов в сравнении с данными, прошедшими корректировку [Trivellato, Rettore, 1986]. Та же закономерность отмечается и более современными исследователями [Bec, Kanda, 2020; Croushore, Stark, 2003; Patterson, Heravi, 1991; Richardson, van Florenstein Mulder, Vehbi, 2021].



* Реальные располагаемые доходы рассчитаны по старой методологии.

** Пунктирная линия – при переходе на новые методологию и базы расчета пересмотры не учитывались.

Источники: Росстат, расчеты ДИП.

Рисунок 1 - Частота пересмотров показателей, % [Горностаев и др., www]

Благодаря достижениям в области вычислительной мощности методы машинного обучения недавно были предложены в качестве альтернативы моделям регрессии временных рядов, обычно используемым центральными банками для прогнозирования ключевых макроэкономических переменных. Модели машинного обучения особенно подходят для обработки больших наборов данных, когда количество потенциальных регрессоров больше, чем количество доступных наблюдений. Таким образом, исследуемый вопрос является актуальным.

Описание данных

Основой для настоящего исследования является база данных Центрального банка России, из которой были получены ключевые макроэкономические параметры с 1993 г., при этом из-за неполноты данных используемых показателей за точку отчета был принят 2007 г. Эта база данных содержит значения ежемесячного сборника «Краткосрочные экономические показатели Российской Федерации», в котором публикуется информация о социально-

экономической ситуации в стране, включая ценовую, производственную, социальную и финансовую статистику.

Сохранение актуальных показателей обеспечено за счет создания матрицы, состоящей из данных за все предыдущие периоды в формате «дата публикации – показатели». Модель исследования основана на месячных данных, поэтому были построены соответствующие когорты (винтажи) (рис. 2).

DAT	2006-04	2006-05	2006-06	2006-07	2006-08	2006-09	2006-10	2006-11	2006-12	2007-01	2007-02	2007-03	2007-04
01.04.2006	3,87531	3,903709	3,947098	3,945551	3,959716	3,967595	3,973121	4,01637	3,980417	3,976655	3,97662	4,031548	4,075837
01.05.2006		3,948714	4,002265	3,974922	3,988667	3,998744	4,014533	4,060248	4,020921	4,031634	4,025936	4,048653	4,074679
01.06.2006			4,280833	4,224306	4,236826	4,249073	4,27887	4,329182	4,285262	4,238144	4,292476	4,289512	4,378965
01.07.2006				4,024998	4,031612	4,023614	4,091294	4,159887	4,069841	4,094641	4,1096	4,131548	4,052756
01.08.2006					4,058645	4,101002	4,125779	4,229508	4,136085	4,214368	4,162778	4,160818	4,232463
01.09.2006						4,272325	4,29902	4,362002	4,342026	4,329902	4,355743	4,354062	4,321415
01.10.2006							4,230452	4,300282	4,326972	4,416035	4,316447	4,319372	4,424176
01.11.2006								4,602945	4,606954	4,695889	4,553234	4,530247	4,51325
01.12.2006									4,333661	4,474216	4,251931	4,26525	4,360662
01.01.2007										4,791384	5,908245	5,689122	5,607198
01.02.2007											5,875732	5,454432	5,386748
01.03.2007												5,407855	5,339501
01.04.2007													5,376093

Рисунок 2 - Примеры данных

Следует отметить «треугольную» структуру используемых данных, что связано с увеличением на каждый период по горизонтали последующего винтажа. Главное отличие винтажей от обычных датасетов состоит в отражении при когортном анализе динамики пересмотров для уточнения показателей, поэтому бывают случаи, когда за одни и те же даты значения показателей могут быть различны. На рис. 3 представлены показатели, которые использовались для построения винтажных и классических прогнозов, а также их свойства.

Название показателя	Переменная	Динамика	Вид показателя
Процентная ставка межбанковского рынка	IR	Абсолютный	Классический
Уровень безработицы	UNEP	Абсолютный	Винтаж
Индекс цен на нефть марки Brent	BRENT	Абсолютный	Классический
Ввод в действие новых жилых домов	DWELL	Абсолютный	Винтаж
Индекс РТС	RTS	Абсолютный	Классический
Валютный курс	USD	Абсолютный	Классический
Индекс промышленного производства	IPP	Процентный	Винтаж
Индекс потребительских цен	CPI	Процентный	Винтаж
Индекс реальных зарплат	RWAGE	Процентный	Винтаж
Оборот розничной торговли	RETAIL	Процентный	Винтаж

Рисунок 3 - Перечень исследуемых показателей

Описание эксперимента

Исследование состояло из этапов, представленных на рис. 4.

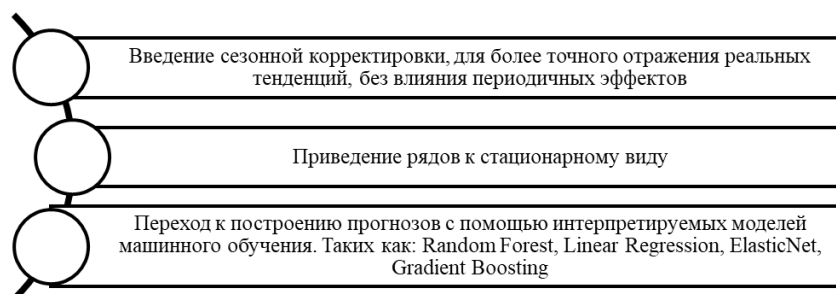


Рисунок 4 - Этапы исследования

Поскольку количество наблюдений было небольшим, данные были разделены в соотношении 2 к 1 на тестовую и тренировочную выборки. Сравнение прогнозов проводилось на основе классических метрик задач регрессии MAE и RMSE.

В табл. 1 и 2 представлены результаты исследования в метрике MAE.

Таблица 1 - Результаты исследования MAE на тренировочной выборке¹

Lags	Metrika	Linear Regression	Random Forest	ElasticNet	Gradient Boosting
0	Vintage_MAE	0,02631	0,01160	0,02816	0,00001
	Classic_MAE	0,02525	0,01275	0,02697	0,00045
1	Vintage_MAE	0,02487	0,01403	0,02816	0,00001
	Classic_MAE	0,02370	0,01204	0,02627	0,00045
2	Vintage_MAE	0,02333	0,01360	0,02784	0,00001
	Classic_MAE	0,02262	0,01250	0,02593	0,00045
3	Vintage_MAE	0,02270	0,01278	0,02749	0,00001
	Classic_MAE	0,02200	0,01324	0,02585	0,00045
4	Vintage_MAE	0,02240	0,01218	0,02750	0,00001
	Classic_MAE	0,02070	0,01188	0,02541	0,00045

Таблица 2 - Результаты исследования MAE на тестовой выборке²

Lags	Metrika	Linear Regression	Random Forest	ElasticNet	Gradient Boosting
0	Vintage_MAE	0,04388	0,09899	0,10221	0,09582
	Classic_MAE	0,03212	0,09928	0,04589	0,08970
1	Vintage_MAE	0,02536	0,09087	0,10223	0,09535
	Classic_MAE	0,01930	0,08600	0,02755	0,08649
2	Vintage_MAE	0,02168	0,09087	0,02779	0,09714
	Classic_MAE	0,02864	0,08967	0,01991	0,05339
3	Vintage_MAE	0,03378	0,09946	0,01466	0,09631
	Classic_MAE	0,03067	0,08103	0,01933	0,09939
4	Vintage_MAE	0,03411	0,09520	0,02259	0,09507
	Classic_MAE	0,03341	0,09187	0,02447	0,07540

¹ Построено по данным: [Базы данных, www].

² Построено по данным: [Там же].

Сравнение полученных на тестовой и тренировочной выборках значений показало слишком большую разницу для Gradient Boosting и Random Forest, что позволяет говорить о переобучении даже с учетом использования механизмов автоподбора параметров. Кроме того, модели регрессии характеризовались более низкими значениями, нежели сложные алгоритмы независимо от лага. При этом лучшие результаты получены на втором и третьем лагах, а после четвертого лага показатели начали ухудшаться.

В табл. 3 и 4 представлены результаты исследования в метрике RMSE на тестовой и тренировочной выборках.

Таблица 3 - Результаты исследования RMSE на тестовой выборке³

Lags	Metrika	Linear Regression	Random Forest	ElasticNet	Gradient Boosting
0	Vintage_RMSE	0,03081	0,01511	0,03275	0,00003
	Classic_RMSE	0,03053	0,01626	0,03143	0,00056
1	Vintage_RMSE	0,02975	0,01780	0,03275	0,00003
	Classic_RMSE	0,02909	0,01594	0,03086	0,00056
2	Vintage_RMSE	0,02861	0,01727	0,03252	0,00003
	Classic_RMSE	0,02765	0,01653	0,03058	0,00056
3	Vintage_RMSE	0,02790	0,01616	0,03223	0,00003
	Classic_RMSE	0,02675	0,01735	0,03028	0,00056
4	Vintage_RMSE	0,02738	0,01470	0,03223	0,00003
	Classic_RMSE	0,02526	0,01543	0,02968	0,00056

Таблица 4 - Результаты исследования RMSE на тестовой выборке⁴

Lags	Metrika	Linear Regression	Random Forest	ElasticNet	Gradient Boosting
0	Vintage_RMSE	0,0467	0,1006	0,1035	0,0975
	Classic_RMSE	0,0358	0,1008	0,0483	0,0913
1	Vintage_RMSE	0,0303	0,0924	0,1036	0,0975
	Classic_RMSE	0,0304	0,0876	0,0307	0,0881
2	Vintage_RMSE	0,0328	0,0924	0,0310	0,1017
	Classic_RMSE	0,0438	0,0912	0,0230	0,0562
3	Vintage_RMSE	0,0475	0,1009	0,0187	0,0985
	Classic_RMSE	0,0471	0,0827	0,0227	0,1009
4	Vintage_RMSE	0,0483	0,0967	0,0290	0,0972
	Classic_RMSE	0,0497	0,0934	0,0276	0,0772

Результаты, представленные в таблицах, показывают одинаковую тенденцию, в соответствии с которой лучшие значения наблюдаются на втором и третьем лагах.

Заключение

В процессе проведения исследования использовались различные совокупности лагов и предикторов, при этом лучшие результаты достигнуты при лагах, равных двум, и на наборе, который включает CPI, IR, USD, UNEMP, RWAGE. Лучшими признаны алгоритмы, которые основаны на регрессии, в первую очередь модель ElasticNet, для которой значения метрики

³ Построено по данным: [Там же].

⁴ Построено по данным: [Там же].

были в два раза ниже наивного метода. Несмотря на то, что более сложные алгоритмы демонстрировали худшие результаты, это можно объяснить возможным неверным выбором лагов и предикторов для них.

Полученные результаты не позволяют утверждать о том, что данные для когортного анализа уступают традиционным. Так, второй лаг ElasticNet продемонстрировал значение RMSE 0,03097 на данных для когортного анализа и 0,22967 на классических данных, при этом для Linear Regression данные для когортного анализа более точны – 0,03281 в сравнении с 0,04384 для классического набора. Для дальнейших исследований может быть перспективным исследование на основе рассмотрения дополнительных объясняющих переменных и более сложных методов анализа данных.

Библиография

1. Астафьева Е.В., Турунцева М.Ю. Пересмотры ВВП: данные и оценка // Экономический журнал ВШЭ. 2021. Т. 25. № 1. С. 65-101.
2. Базы данных. URL: https://cbr.ru/hd_base/
3. Горностаев Д. и др. База данных пересмотров макроэкономических показателей в России. URL: https://cbr.ru/ec_research/ser/wp76/
4. Bec F., Kanda P. Is inflation driven by survey-based, VAR-based or myopic expectations? An empirical assessment from US real-time data // The North American journal of economics and finance. 2020. Vol. 51. DOI: 10.1016/j.najef.2019.101050
5. Croushore D., Stark T. A real-time data set for macroeconomists: does the data vintage matter? // Review of economics and statistics. 2003. Vol. 85. No. 3. P. 605-617.
6. Denton F.T., Kuiper J. The effect of measurement errors on parameter estimates and forecasts: a case study based on the Canadian preliminary national accounts // Review of economics and statistics. 1965. Vol. 47. No. 2. P. 198-206.
7. Mamedli M., Shibitov D. Forecasting Russian CPI with data vintages and machine learning techniques. URL: http://www.cbr.ru/eng/ec_research/forecasting-russian-cpi-with-data-vintages-and-machine-learning-techniques/
8. Patterson K.D., Heravi S.M. Are different vintages of data on the components of GDP co-integrated? Some evidence for the United Kingdom // Economics letters. 1991. Vol. 35. No. 4. P. 409-413. DOI: 10.1016/0165-1765(91)90011-9
9. Richardson A., van Florenstein Mulder T., Vehbi T. Nowcasting GDP using machine-learning algorithms: a real-time assessment // International journal of forecasting. 2021. Vol. 37. No. 2. P. 941-948. DOI: 10.1016/j.ijforecast.2020.10.005
10. Trivellato U., Rettore E. Preliminary data errors and their impact on the forecast error of simultaneous-equations models // Journal of business and economic statistics. 1986. Vol. 4. No. 4. P. 445-453.

Cohort data in the forecasting of macroeconomic indicators by using machine learning methods

Vladimir A. Kalashnikov

Postgraduate,
Assistant at the Department of data analysis and machine learning,
Junior Researcher at the Engineering Center,
Financial University under the Government of the Russian Federation,
125167, 49/2, Leningradsky ave., Moscow, Russian Federation;
e-mail: vakalashnikov@fa.ru

Maksim V. Ivanov

Master's Degree Student,
Higher School of Economics – National Research University,
101000, 20, Myasnitskaya str., Moscow, Russian Federation;
e-mail: mvivanov_5@edu.hse.ru

Abstract

The research is relevant due to the fact that forecasting has been an important research issue for decades. There is a constant search for models, methods and variables that have a significant impact on the results obtained with a view to ensuring the accuracy of forecasts. The article deals with the problem of forecasting various macroeconomic indicators by using cohort data. It aims to reveal differences between cohort-based forecasting methods and final data. The authors of the article carry out an analysis of studies substantiating the relevance of using cohort data in forecasting. The article uses machine learning algorithms on datasets containing financial indicators to test the effectiveness of forecasting macroeconomic indicators. Taking into account information from the database of the Central Bank of the Russian Federation, the authors conduct independent research using machine learning methods. The results of the analysis of RMSE and MAE metrics show that cohort-based forecasting in some cases has advantages over traditionally used revised data. The article demonstrates that the accuracy of forecasts is improved by using machine learning methods. Having considered cohort data in the forecasting of macroeconomic indicators by using machine learning methods, the authors conclude that machine learning methods are effective for big cohort data.

For citation

Kalashnikov V.A., Ivanov M.V. (2022) Ispol'zovanie dannykh dlya kogortnogo analiza pri prognozirovanii makroekonomicheskikh pokazatelei metodami mashinnogo obucheniya [Cohort data in the forecasting of macroeconomic indicators by using machine learning methods]. *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 12 (8A), pp. 7-15. DOI: 10.34670/AR.2022.31.65.002

Keywords

Cohort analysis, forecasting, macroeconomic indicators, machine learning, industrial production indices.

References

1. Astaf'eva E.V., Turuntseva M.Yu. (2021) Peresmotry VVP: dannye i otsenka [GDP revisions: data and assessment]. *Ekonomicheskii zhurnal VShE* [The HSE economic journal], 25 (1), pp. 65-101.
2. *Bazy dannykh* [Databases]. Available at: https://cbr.ru/hd_base/ [Accessed 08/08/22].
3. Bec F., Kanda P. (2020) Is inflation driven by survey-based, VAR-based or myopic expectations? An empirical assessment from US real-time data. *The North American journal of economics and finance*, 51. DOI: 10.1016/j.najef.2019.101050
4. Croushore D., Stark T. (2003) A real-time data set for macroeconomists: does the data vintage matter? *Review of economics and statistics*, 85 (3), pp. 605-617.
5. Denton F.T., Kuiper J. (1965) The effect of measurement errors on parameter estimates and forecasts: a case study based on the Canadian preliminary national accounts. *Review of economics and statistics*, 47 (2), pp. 198-206.
6. Gornostaev D. et al. *Baza dannykh peresmotrov makroekonomicheskikh pokazatelei v Rossii* [The database of the revisions of macroeconomic indicators in Russia]. Available at: https://cbr.ru/ec_research/ser/wp76/ [Accessed 08/08/22].

-
7. Mamedli M., Shibitov D. *Forecasting Russian CPI with data vintages and machine learning techniques*. Available at: http://www.cbr.ru/eng/ec_research/forecasting-russian-cpi-with-data-vintages-and-machine-learning-techniques/ [Accessed 08/08/22].
 8. Patterson K.D., Heravi S.M. (1991) Are different vintages of data on the components of GDP co-integrated? Some evidence for the United Kingdom. *Economics letters*, 35 (4), pp. 409-413. DOI: 10.1016/0165-1765(91)90011-9
 9. Richardson A., van Florenstein Mulder T., Vehbi T. (2021) Nowcasting GDP using machine-learning algorithms: a real-time assessment. *International journal of forecasting*, 37 (2), pp. 941-948. DOI: 10.1016/j.ijforecast.2020.10.005
 10. Trivellato U., Rettore E. (1986) Preliminary data errors and their impact on the forecast error of simultaneous-equations models. *Journal of business and economic statistics*, 4 (4), pp. 445-453.