

Piotr Wójcik

Katedra Data Science
Wydział Nauk Ekonomicznych
Uniwersytet Warszawski
ORCID: 0000-0003-1853-8784

Bartłomiej Wieczorek

Data Science Lab WNE UW
Wydział Nauk Ekonomicznych
Uniwersytet Warszawski
ORCID: 0009-0004-8975-8658

Analiza czynników realnej konwergencji z wykorzystaniem wyjaśnialnej sztucznej inteligencji

Abstrakt

Analiza czynników konwergencji od wielu lat pozostaje ważnym obszarem badań makroekonomicznych i ma istotne znaczenie dla polityki gospodarczej, w szczególności polityki spójności Unii Europejskiej. W artykule zaproponowano nowatorskie podejście do identyfikacji czynników konwergencji. Po pierwsze, artykuł podejmuje próbę identyfikacji potencjalnych nieliniowości za pomocą metod uczenia maszynowego. Po drugie, wykorzystuje nowoczesne metody, które pozwalają na wybór zmiennych ważnych dla konwergencji i identyfikację kształtu zależności niezależnie od przyjętej specyfikacji modelu. Oprócz metody LASSO, stosowanej wcześniej w badaniach konwergencji, wykorzystana została regresja wektorów nośnych, las losowy oraz algorytmy wzmacniające. Dodatkowo narzędzia wyjaśnialnej sztucznej inteligencji (XAI) pozwalają analizować znaczenie czynników konwergencji w sposób niezależny od typu modelu oraz interpretować to, czego modele nauczyły się z danych. Wyniki wskazują, że obserwowana we wcześniejszych badaniach konwergencja warunkowa mogła wynikać z niewłaściwej specyfikacji modeli liniowych. Stwierdzono, że w poprawnym podejściu nieliniowym początkowy PKB nie jest silnie powiązany ze wzrostem, potwierdzono natomiast występowanie konwergencji klubów.

Słowa kluczowe: konwergencja międzykrajowa, konwergencja warunkowa, czynniki, uczenie maszynowe, nieliniowość, XAI

The analysis of real convergence factors using explainable artificial intelligence

Abstract

The analysis of convergence factors has long remained an important area of macroeconomic research and has significant implications for economic policy, particularly European Union cohesion policy. This article proposes a novel approach to identifying convergence factors. First, it attempts to identify potential nonlinearities using machine learning methods. Second, it employs modern techniques that allow for the selection of variables relevant to convergence and the identification of the shape of relationships regardless of model specification. In addition to the LASSO method, previously applied in convergence studies, the analysis employs support vector regression, random forest, and boosted algorithms. Furthermore, explainable artificial intelligence (XAI) tools make it possible to assess the importance of convergence factors in a model-agnostic manner and to interpret what the models have learned from the data. The results suggest that the conditional convergence identified in previous studies may have resulted from the misspecification of linear models. It is found that, in a proper nonlinear framework, initial GDP is not strongly related to growth, while the existence of club convergence is confirmed.

Keywords: cross-country convergence, conditional convergence, factors, machine learning, nonlinearity, XAI

1. Wprowadzenie i przegląd literatury

Identyfikacja ważnych czynników konwergencji dochodów lub wzrostu gospodarczego między krajami jest tematem badań empirycznych od dziesięcioleci. Jedną z najczęściej stosowanych metod w badaniach empirycznych jest analiza warunkowej konwergencji typu *beta* (Barro i Sala-i-Martin, 2004). Polega ona na analizie związku między średniorocznym tempem wzrostu (jako zmienną objaśnianą) a dochodem początkowym, zazwyczaj z uwzględnieniem dodatkowych zmiennych objaśniających, interpretowanych jako czynniki konwergencji. Wybór tych czynników ma kluczowy wpływ na wnioskowanie o wystąpieniu konwergencji (Durlauf, 2009). Wnioski dotyczące znaczenia poszczególnych czynników warunkujących występowanie konwergencji często różnią się w zależności od stosowanych podejść empirycznych.

Levine i Renelt (1992) zastosowali wariant analizy wiarygodności zwany Analizą Ekstremalnych Granic (ang. Extreme Bound Analysis, EBA) do analizy konwergencji między krajami (Leamer, 1985). Ich podejście zostało później zakwestionowane przez Sala-i-Martina (1997), który wykazał, że EBA ma zbyt dużą moc, aby jakkolwiek zmienna mogła spełnić założenia występowania zależności. Zaproponował on odejście od „ekstremalnego testu” i raczej przypisanie pewnego poziomu ufności dla istotności każdego z czynników konwergencji (Sala-i-Martin, 1997), obliczanego za pomocą statystyk ważonych na podstawie wielu (milionów) regresji wzrostu w różnych konfiguracjach zmiennych objaśniających (uśrednianie bayesowskie). W przeciwieństwie do tego podejścia Hendry i Krolzig (2004) zaproponowali oszacowanie tylko jednej regresji, a następnie zastosowanie procedury „od ogółu do szczegółu”. Podejście to nie jest jednak możliwe do zastosowania, gdy liczba potencjalnych czynników konwergencji przekracza liczbę obserwacji. Sala-i-Martin et al. (2004) oraz Lay i Steel (2007) zaproponowali uśrednianie modeli bayesowskich z różnymi wartościami *a priori*, aby uwzględnić niepewność wyboru właściwych zmiennych. Większość wspomnianych powyżej podejść została zakwestionowana przez Ciccone i Jarocińskiego (2010), zarówno na gruncie teoretycznym, jak i empirycznym. Stwierdzili oni, że tylko poprawna identyfikacja czynników konwergencji pozwoli na sformułowanie wiarygodnych rekomendacji dla polityki gospodarczej. W kontekście narzędzi uczenia maszynowego niektórzy badacze (np. Varian, 2014) zastosowali względnie prostą metodę LASSO (Tibshirani, 1996) do regresji typu Barro. Jest to rozwinięcie klasycznego modelu regresji liniowej, umożliwiające automatyczną eliminację zmiennych niemających znaczącego wpływu na modelowane zjawisko. Pozwala ona również na oszacowanie modelu, nawet jeśli liczba zmiennych objaśniających znacząco przekracza liczbę obserwacji. Hofmarcher et al. (2015) wykazali, że metoda LASSO pozwala wyeliminować problem współliniowości zmiennych objaśniających. Co więcej, wyniki predykcji tempa wzrostu na nowych danych, nieużywanych do estymacji modeli, były lepsze niż w przypadku technik bayesowskiego uśredniania modelu.

Wadą wszystkich wymienionych metod (również LASSO) jest to, że zakładają one liniowość zależności między tempem wzrostu a początkowym dochodem i poszczególnymi czynnikami konwergencji. Levine i Renelt (1992) przewidywali, że rzeczywista zależność między tempem wzrostu a początkowym dochodem może być nieliniowa. Istnienia nieliniowych zależności nie powinno się ignorować. He i Xu (2019) pokazują, że zidentyfikowanie zmiennej jako statystycznie istotnego czynnika wzrostu w specyfikacji liniowej może być wynikiem niewłaściwej specyfikacji modelu. Gdyby poprawna postać modelu była nieliniowa, zmienna ta mogłaby nie być istotnie powiązana ze wzrostem.

Niniejszy artykuł przyjmuje nowatorskie podejście. Jego głównym celem jest identyfikacja ważnych czynników konwergencji dochodowej i rozszerzenie istniejących badań empirycznych na dwa sposoby. Po pierwsze, artykuł podejmuje próbę identyfikacji potencjalnych nieliniowości za pomocą metod uczenia maszynowego. Po drugie, wykorzystuje nowoczesne metody, które pozwalają na identyfikację ważnych zmiennych objaśniających i identyfikację kształtu relacji ze zmienną objaśnianą bez żadnych założeń *a priori*. Oprócz wspomnianej metody LASSO wykorzystaliśmy regresję wektorów nośnych (SVR, Vapnik, 1995) oraz dwa podejścia zespołowe bazujące na drzewach regresyjnych – las losowy (Breiman, 2001) oraz jeden z wariantów algorytmu wzmacnianego (XGBoost, Chen, 2016). Do strojenia hiperparametrów modeli wykorzystana została walidacja krzyżowa Leave-One-Out, oparta na minimalizacji metryki średniego błędu bezwzględnego (MAE). Ponieważ oprócz skutecznej predykcji równie ważna jest zdolność do interpretacji tego, czego model nauczył się z danych, zastosowaliśmy także wybrane narzędzia wyjaśnialnej sztucznej inteligencji (XAI).

W artykule poddajemy weryfikacji dwie hipotezy badawcze. Po pierwsze, zakładamy, że między tempem wzrostu a początkowym dochodem zachodzi relacja nieliniowa. Po drugie, w związku z tym, narzędzia uczenia maszynowego, które pozwalają odkryć potencjalną nieliniowość relacji, wyjaśniać zróżnicowanie tempa wzrostu między krajami z większą dokładnością niż modele liniowe.

Narzędzia uczenia maszynowego nadal umożliwiają interpretację modelu i ocenę ważności cech. Dzięki temu pozwolą nam na lepszą identyfikację czynników konwergencji, a tym samym będą pomocne w formułowaniu rekomendacji dla polityki gospodarczej.

W niniejszym artykule analizowaliśmy dwa zbiory danych, często wykorzystywane we wcześniejszych badaniach, aby umożliwić porównanie wyników ze stosowanymi wcześniej metodami. Pierwszy zbiór danych został po raz pierwszy wykorzystany w pracy Fernandez et al. (2001). Drugi zbiór danych został przedstawiony w pracy Sala-i-Martin et al. (2004).

Artykuł ma następującą strukturę: w drugiej części krótko opisano metody modelowania zastosowane w kolejnych częściach; w trzeciej części przedstawiono wyniki analizy czynników konwergencji na podstawie dwóch wyżej wymienionych zbiorów danych i porównano je z wynikami innych badaczy; ostatnia sekcja zawiera podsumowanie i wnioski.

2. Metody

W tej sekcji przedstawiono pokrótce w sposób nietechniczny i intuicyjny algorytmy uczenia maszynowego, które zostaną wykorzystane w części empirycznej. Przed zastosowaniem algorytmów wszystkie zmienne zostały wystandaryzowane, zgodnie z sugestiami literatury.

2.1. LASSO (Least Absolute Shrinkage and Selector Operator)

LASSO (Tibshirani, 1996) to jedna z kilku metod regularyzacji. W wersji podstawowej można ją postrzegać jako rozszerzenie Metody Najmniejszych Kwadratów (MKN, ang. *ordinary least squares*, OLS). Różni się od MKN funkcją kosztu – nie tylko minimalizuje sumę kwadratów reszt, ale także uwzględnia sumę wartości bezwzględnych parametrów modelu liniowego jako dodatkowe ograniczenie. Dodanie takiej kary w procesie optymalizacji powoduje poszukiwanie parametrów, które dobrze pasują do danych, ale dodatkowo są jak najmniejsze (najbliższe zera). Wartości parametrów przy mniej ważnych zmiennych będą się kurczyć w kierunku zera, niektóre z nich będą nawet w wynikowym modelu równe zeru. Kosztem