UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneva Switzerland
04 April 2023

Time Series Outlier Detection using Metadata and Data Machine Learning in Statistical Production

Olivier Sirello, Bilyana Bogdanova, Bianca Ligani, Alexis Maurin, Ismail Mustafi (Bank for International Settlements (BIS))
olivier.sirello@bis.org

Abstract

The detection of anomalies in macroeconomic and financial time series is often regarded as a challenging but crucial task to produce high-quality official statistics. Well-designed and performant outlier checks are a key component of statistical production pipelines and they play a crucial role to improve the process efficiency.

Several methods exist for detecting outliers, ranging from traditional techniques to new machine learning algorithms (ML). The latter are particularly useful to evaluate the presence of contextual outliers taking stock of the cluster dynamics. However, many ML algorithms face several challenges when it comes to the analysis of macro and financial time series in a statistical production environment, for instance due to the existence of variable-length time series which may negatively affect clustering outcomes.

In this paper, we propose a novel approach to identify time series outliers in a statistical production environment. Our suggested solution is innovative in two respects. On the one hand, we exploit time series metadata, on top of data, to identify the clusters needed to perform contextual outlier detection. In doing so, we rely on a mix of unsupervised ML algorithms (e.g., Affinity Propagation, DBSCAN, Isolation Forest) which present the advantage of not requiring labelled data sets. To the best of our knowledge, we are the first to combine both data and metadata for anomaly detection checks for macroeconomic and financial time series in official statistics. On the other hand, we also ensure the suitability of our solution for production purposes, which entails, for example, the use of specific ML techniques to take into account variable-length time series in the clustering process (e.g., Dynamic Time Warping).