BIS

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

Bilyana Bogdanova, Bianca Ligani, Alexis Maurin, Ismail Mustafi, **Olivier Sirello** – Bank for International Settlements

UNECE Machine Learning for Official Statistics

Geneva, 5 June 2023

At a glance

- Outlier checks are typically performed relying on traditional statistical methods, looking at each series individually
- However, statistical agencies often pool data and metadata on related indicators from multiple sources or countries
- Knowing the *context* of the series (eg related series sharing the same shape) may help to improve outlier detection, especially in case of shocks propagating across indicators and countries

➢ How machine learning can help to improve outlier checks by performing *contextual* anomaly detection?

Outline

- Introduction
- Background and motivation
- Clustering metadata and data
- Outlier detection
- Preliminary results

Introduction

- BIS Data Bank is a data warehouse hosting more than sixty thousand macroeconomic and financial time series (mostly submitted by central banks)
- New ML-based method to run contextual outlier detection harvesting large availability of cross-country data and metadata
- Multiple trade-offs: data quality, process efficiency, manual intervention and domain-specific knowledge
- Preliminary results show that the new method outperforms traditional threshold-based rules and highlight its potential to increase data quality and process efficiency

Background and motivation

- BIS Data Bank identifies outliers when the absolute difference between the observation i and the moving average MM over the rolling window of length k is greater or equal to its standard deviation σ over the window of same length k multiplied by a factor

$$\left|i - MM\left(i\right)_k\right| \geq \sigma\left(i\right)_k \times b$$

- Thresholds are often not appropriate for time series with linear breaks

- Contextual outlier checks are better suited for outlier detection on related cross-country data

  Contextual anomalies are those "points which can be normal in a certain context, while detected as anomaly in another context" (Braei and Wagner 2020)

- ML clustering required to define the *context* against which performing checks

Clustering metadata and data

- Clustering time series leveraging both metadata and data

- Metadata

  - Cluster time series on topic attributes and time series dimensions (eg topic name, title and unit code). For example, we expect series belonging to topic "interest rate, policy rate" to be close to "interest rate, official discount rate"

  - Two stages
    1. Jaro-Winkler distance matrix
       - Common characters, transpositions and agreeing initial characters between strings (Winkler 1999)
    2. Affinity propagation
       - Data points coalesce around exemplars until convergence (Frey and Dueck 2007)

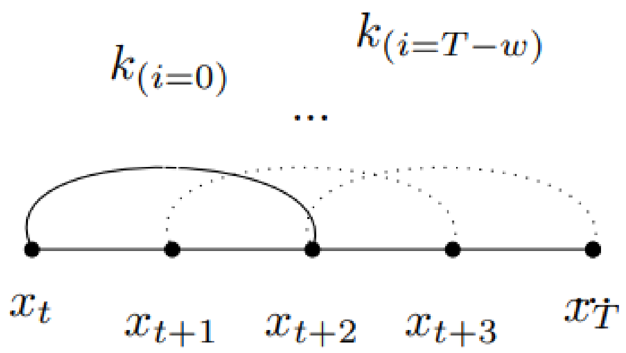
Clustering metadata and data (cont'd)

- Data
  - Cluster variable time series based on their shape (eg to account for variable-length, time shifts or data gaps)
  - Two stages
    1. Dynamic Time Warping
       - Similarity based on shape and temporal dynamics of time series (Bagnall et al. 2016)
       - Minimization of cumulative distance of each mapped pairs of points to derive the warping paths
       - To prevent pathological mappings, we apply a cap on the maximal shift using Sakoe-Chiba band (Dau et al. 2018)
    2. Density-based spatial clustering of applications with noise (DBSCAN)
       - Unsupervised cluster discovery as high vs low density areas
       - Epsilon-neighbourhood containing a minimum number of points (Ester et al. 1998)

Contextual outlier detection

- For each cluster, data point $x_{t,S1}$ is flagged as outlier if it deviates significantly from its neighbouring points, $x_{t-1,S1}$ … $x_{t-n,S1}$ as well as from the data points belonging to related series $x_{t-1,S2}$ … $x_{t-n,SZ}$

- We rely on DBSCAN to capture outliers, ie data points that lie in low-density areas. We also apply a sliding window of size *w* up to when the number of iterations *i* matches the difference between the total number of periods and the window size

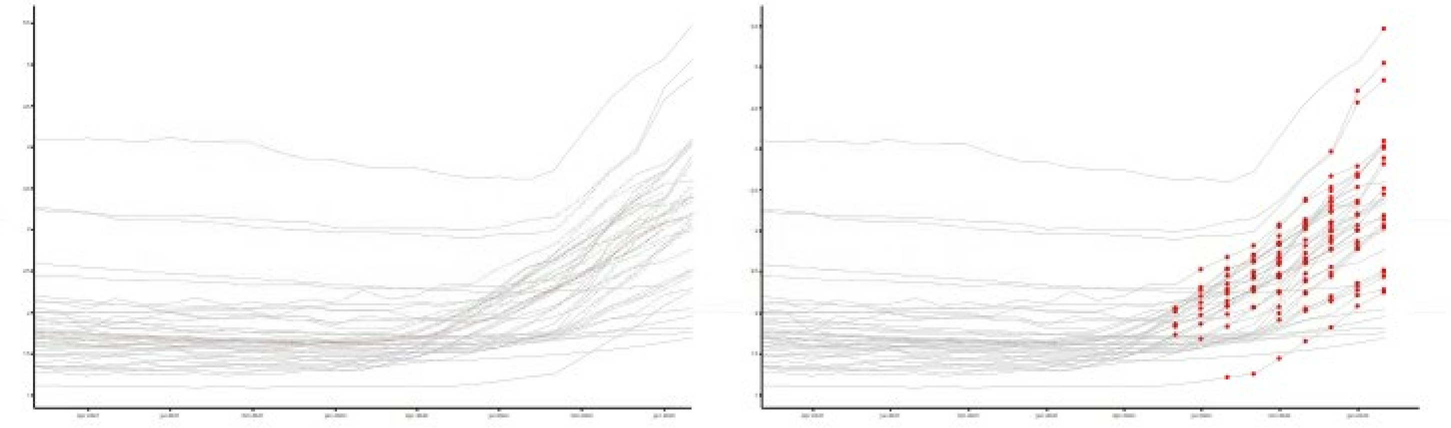


- Because differences in levels may exist, we differentiate the time series (min-max scaler)

Application and preliminary results

- We test this method on a subset of the BIS DataBank (668 monthly financial time series, spanning from January 1919 and May 2023 and 1.99% missing points on average)

- With low tuning effort, we manage to cluster around two-thirds of the sample and remove false outliers flagged by the current check (ie no outliers expected)



(a) Contenxtual (b) Current

Figure 1: Comparison between contextual and current outlier checks during the sharp increase in interest rates since mid-2022

Conclusion and way forward

- ML methods for contextual outlier detection are a powerful tool to enhance traditional outlier detection checks
- Better accuracy generally entails less manual intervention to get rid of false positives
- More effort needed in three directions:
  - Better assessment of the mix of individual and contextual outlier checks
  - Further testing and robustness checks
  - ML pipeline in the context of statistical production