A Bagging Approach with a Scaled Logit Transformation for Improving Predictive Performance in Non-stationary Time Series Analysis

Young Eun Jeon[a] • Yongku Kim[b] • Jung-In Seo[a]

[a] Department of Data Science, Andong National University, Andong, Korea

[b] Department of Statistics, Kyungpook National University, Daegu, Korea

1. Abstract

Backgrounds

- Bagging is well known for significantly improving predictive performance in time series analysis.
- However, **traditional block bootstrapping techniques, such as a moving block bootstrap (MBB), struggle with non-stationary time series due to requiring a stationary assumption.**

Purpose

- This study aims to improve the predictive performance of non-stationary time series.
- To achieve this goal, we propose **a bagging approach with a bootstrapping algorithm featuring a scaled logit transformation.**

Results and Conclusions

- The applicability of our approach is verified through the analysis of three types of non-stationary time series datasets with different frequencies.
- **Our approach has superior predictive performance in terms of uncertainty, compared to a bagging approach with Box-Cox transformation.**

2. Methodology

Schematic Diagram

- A overall prediction framework for our approach is illustrated as a schematic diagram in Fig1.

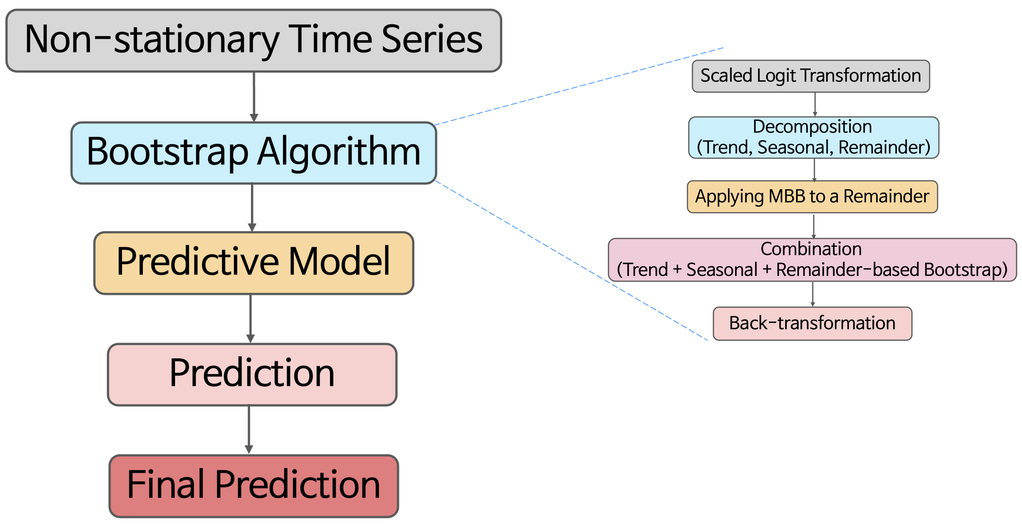


Fig 1. Prediction framework

- The main steps in our prediction framework are the applications of a scaled logit transformation and MBB, and the two techniques are described in the following section.

Scaled Logit Transformation

- Let $T_1, \dots, T_n$ be a time series dataset with non-stationarity such as trend or seasonality.
- Given that all data points lie within the interval (a, b), a scaled logit transformation is defined as follows:

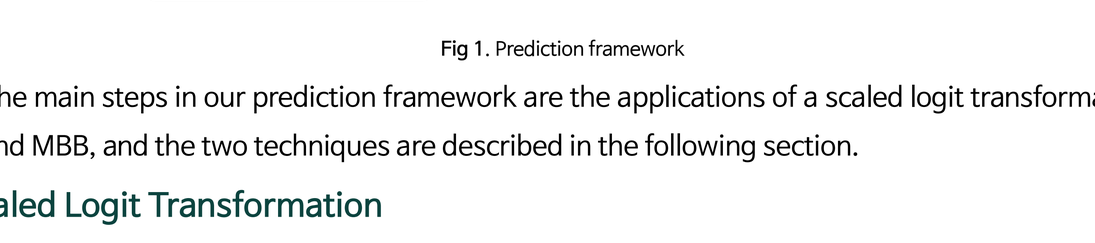
$$T_t^* = \log\left(\frac{T_t - a}{b - T_t}\right), \quad t = 1, \dots, n, \qquad (1)$$

where a and b are constants that need to be determined from the characteristics of the dataset.

- This study proposes the use of $b = T_{max} + 3\sigma$ based on the normality assumption of time series.
  - T_{max} : Maximum value of $T_1, \dots, T_n$
  - σ : Standard deviation

MBB

1) Divide n data points into $n - l + 1$ overlapping blocks with $l(< n)$ data points
   - l : Block size
2) Draw randomly blocks



Predictive Models

Autoregressive Integrated Moving Average (ARIMA) Model

$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - B)^d T_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)\varepsilon_t, \quad \varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$$

- ✓ p : The order of the autoregression term
- ✓ d : The number of differences required to make a stationary time series
- ✓ q : The order of the moving average term
- ✓ c : Constant
- ✓ WN: White noise

SARIMA Model

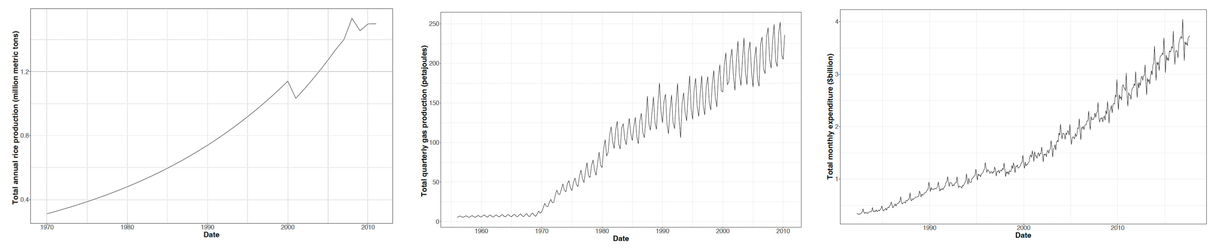
$$(1 - \phi_1 B - \cdots - \phi_p B^p)(1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps})(1 - B^s)^D (1 - B)^d T_t = c + (1 + \theta_1 B + \cdots + \theta_q B^q)(1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs})\varepsilon_t$$

- ✓ P : The order of the seasonal autoregression term
- ✓ D : The number of seasonal differences required to make a stationary time series
- ✓ Q : The order of the seasonal moving average term
- ✓ s : Seasonal frequency

3. Application

Real Datasets

- Three non-stationary time series datasets with different frequencies from R package **fpp2** are used.
  a) *guinearice* : Total annual rice production from 1970 to 2011 in Guinea
  b) *qgas* : Total quarterly gas production between the first quarter of 1956 and the second quarter of 2010 in Australia
  c) *auscafe* : Total monthly expenditure on eating out, including cafes, restaurants, and takeaway food services, between April 1982 and September 2017 in Australia



(a) *guinearice* (b) *qgas* (c) *auscafe*

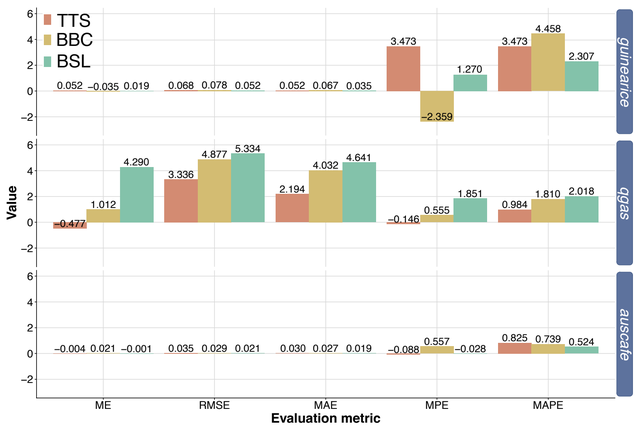
Fig 2. Line plot for three non-stationary time series datasets

Setting

- Data Partition
  - Test dataset : The last five observations
  - Training dataset : The remaining observations
- Number of Bootstrap Datasets : 1,000

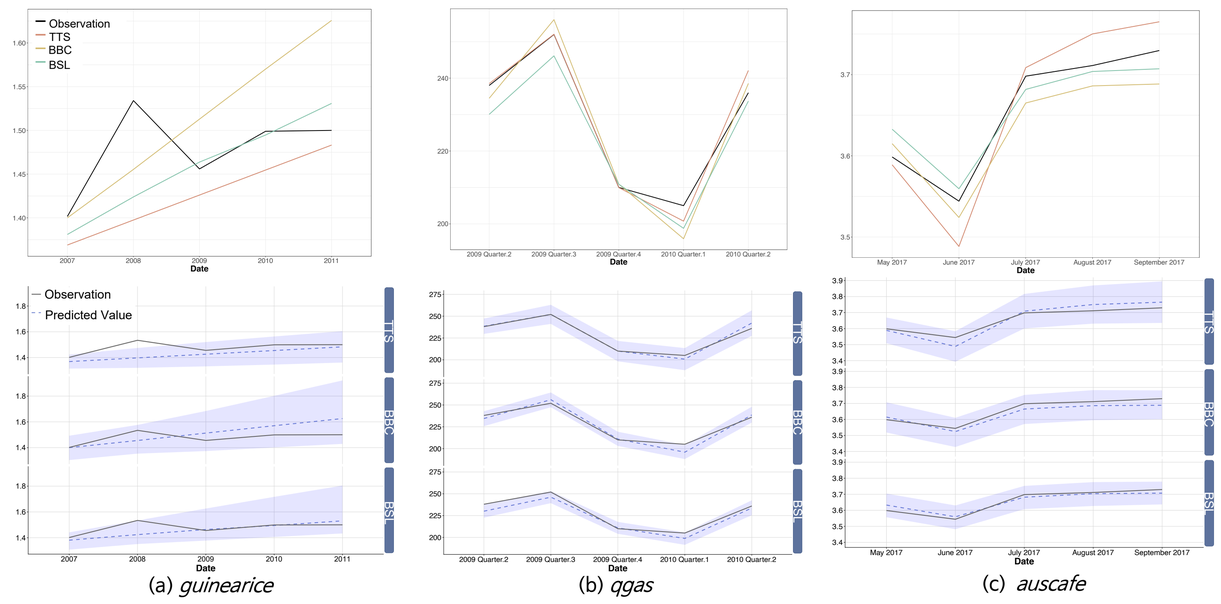
Results

- Our approach : BSL
- Traditional approach
  - TTS : Traditional time series approach without bootstrapping
  - BBC : Bagging approach with the Box-Cox transformation



- Let z_t and $\hat{z}_t$ are observed and predicted values, respectively, and $e_t = z_t - \hat{z}_t$ is the error.
  - Mean Error (ME) : $\frac{1}{m}\sum_{t=1}^{m} e_t$
  - Root Mean Square Error (RMSE) : $\sqrt{\frac{1}{m}\sum_{t=1}^{m} e_t^2}$
  - Mean Absolute Error (MAE) : $\frac{1}{m}\sum_{t=1}^{m} |e_t|$,
  - Mean Percentage Error (MPE) : $\frac{100}{m}\sum_{t=1}^{m} \frac{e_t}{z_t}$
  - Man Absolute Percentage Error (MAPE) : $\frac{100}{m}\sum_{t=1}^{m} \left|\frac{e_t}{z_t}\right|$

Fig 3. Bar plots for five evaluation metrics results



(a) *guinearice* (b) *qgas* (c) *auscafe*

Fig 4. Plots for prediction results

4. Discussions and Conclusions

- For *guinearice* and *auscafe* datasets, the values of the BSL are closer to zero than those of the other approaches in all evaluation metrics.
  ⇒ The proposed approach is useful in improving accuracy for annual and monthly time series.
- For *qgas* and *auscafe* datasets, the predictive intervals of the BSL have smaller length than other approaches.
  ⇒ The proposed approach improves predictive performance by reducing uncertainty for quarterly and monthly time series.
- **Based on these results, the proposed approach is anticipated to yield more accurate and reliable predictions in time series analysis, particularly for real-world datasets exhibiting non-stationary and complex dependency structures.**