

A Study on the Test and Visualization of Change in Structures Associated with the Occurrence of Non-Stationary of Long-Term Time Series Data Based on Unit Root Test

Jaeseong Yoo[†] · Jaegul Choo^{††}

ABSTRACT

Structural change of time series means that the distribution of observations is relatively stable in the period of constituting the entire time series data, but shows a sudden change of the distribution characteristic at a specific time point. Within a non-stationary long-term time series, it is important to determine in a timely manner whether the change in short-term trends is transient or structurally changed. This is because it is necessary to always detect the change of the time series trend and to take appropriate measures to cope with the change. In this paper, we propose a method for decision makers to easily grasp the structural changes of time series by visualizing the test results based on the unit root test. Particularly, it is possible to grasp the short-term structural changes even in the long-term time series through the method of dividing the time series and testing it.

Keywords : Time Series, Non-Stationary, Unit Root Test, Visualization

Unit Root Test를 기반으로 한 장기 시계열 데이터의 Non-Stationary 발생에 따른 구조 변화 검정 및 시각화 연구

유재성[†] · 주재걸^{††}

요 약

시계열의 구조 변화란, 전체 시계열 자료를 구성하는 기간에서 관측치들의 분포가 상대적으로 안정적이다가, 특정 시점에서 분포 특성의 급격한 변화를 보이는 것을 의미한다. 비정상(non-stationary) 장기 시계열 안에서도, 단기적인 추세의 변화가 일시적인 것인지, 아니면 구조적으로 변한 것인지를 적시에 판단하는 것은 중요하다. 이는 시계열 추세의 변화를 상시 감지하여, 변화에 맞는 적절한 대응을 할 필요가 있기 때문이다. 본 연구에서는 단위원 검정법을 기반으로 한 검정 결과를 시각화함으로써, 의사결정자가 시계열의 구조 변화를 손쉽게 파악할 수 있는 방안을 제시하였다. 특히 시계열을 분할한 후 검정하는 방법을 통해, 장기 시계열일 때에도 단기 구조 변화를 파악할 수 있도록 하였다.

키워드 : 시계열, 비정상성, 단위원 검정, 시각화

1. 서 론

시계열의 구조 변화(structural break)란, 전체 시계열 자료를 구성하는 기간에서 관측치들의 분포가 상대적으로 안정적

이다가, 특정 시점에서 분포 특성의 급격한 변화를 보이는 것을 의미한다[1]. KTX 정차 이후의 지방중소도시의 개별산업 구조 변화, 갑작스러운 폭염으로 인한 전력 사용량 변화, 통화정책 시그널링에 따른 주식 거래량 변화 등 다양한 도메인에서 시계열의 구조 변화는 흔하게 발생하고 있다. 이 구조 변화 시점을 추정하기 위해 다양한 방법을 활용할 수 있으나, 대체로 난이도가 높아 숙련된 분석가가 아니면 이들을 다루기 힘들며, 추정 방법, 기간 등의 파라미터 등을 적절하게 정하지 못함으로써 그릇된 검정 및 결론을 도출하는 오류가 발생할 수 있다. 특히 장기 시계열일수록 데이터가 전체적으로 안정적인 가능성이 작아져, 구조 변화를 검정하기가 더욱 어려워진다.

※ 본 연구는 한국전력공사의 2018년 착수 에너지 거점대학 클러스터 사업에 의해 지원되었음(과제번호:R18XA05).

※ 이 논문은 2018년도 한국정보처리학회 추계학술발표대회에서 'Unit Root Test를 기반으로 한 장기 시계열 데이터의 non-stationary 발생에 따른 추세 변화 검정 및 시각화 연구'의 제목으로 발표된 논문을 확장한 것이다.

† 준 회 원 : 고려대학교 컴퓨터학과 박사수료

†† 비 회 원 : 고려대학교 컴퓨터학과 조교수

Manuscript Received : December 17, 2018

First Revision : January 28, 2019

Second Revision : March 8, 2019

Accepted : April 10, 2019

* Corresponding Author : Jaegul Choo(jchoo@korea.ac.kr)

본 연구에서는 구조 변화가 있을 시 의사결정자가 이를 어렵지 않게 파악, 시의적절한 대응을 할 수 있도록 돕는 시각화 방법을 제시하였다. 특히 이 개발 방법은 시계열을 분할한 후 검정하는 방법을 통해, 장기 시계열일 때에도 단기 구조 변화를 파악할 수 있도록 하였다.

2. 관련 연구

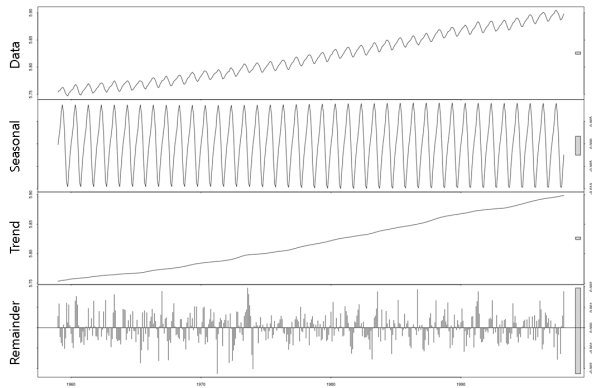


Fig. 1. Seasonal-trend Variation Decomposition using Loess by Cleveland et al.

전통적으로 시계열의 변동 요인을 분할(decomposition)한 다음 이를 선 그래프로 시각화한 것을 관찰하는 방법이 오랫동안 사용됐다. 1884년에 처음으로 추세와 계절변동 제거가 시도된 이래로 Hooker[2], Spencer[3], Anderson과 Nochmals[4], Macauley[5], Copeland[6]에 의해 완성해 나갔으며, Fig. 1과 같이 Loess를 사용한 계절-추세변동 분해를 시도한 Cleveland et al.[7]의 방법은 국가 통계를 다루는 기관에서 널리 사용되었다. 만일 데이터의 성분이 정확하게 분해된다면, 원래 정상 시계열이 갖는 계절 요소와 추세 요소가 분리된 remainder 데이터는 일반적인 상황이라면 단순한 오차의 분산이기 때문에 정규분포 특성을 가지게 된다. 따라서 이 remainder 데이터의 평균, 표준편차 대비 크게 벗어나는 한순간의 데이터를 이상치로 탐지할 수 있고, 만일 한순간이 아닌 어느 정도 지속하고 있다면 구조 변화로 볼 수 있다. 물론 분해된 추세(trend) 요소를 직접 관찰함으로써 비정상성을 판단할 수도 있다.

한편 Henry et al.[8]은 M-정상 과정(M-stationary process)을 기반으로 순환변동(cyclical variation)이 시간에 따라 선형적으로 변하는 비정상 시계열(non-stationary)의 분해를 시도하기도 하였다.

또 다른 전통적인 방법으로는 Chow와 Lin[9], Denton[10]의 참조 시계열(related time series)의 자기 상관 함수(auto-correlation function)를 시각화한 방법이 있다. 자기 상관은 k 개 시차 안에 있는 계열 값 사이의 상관관계로, 현재 및 지난 계열 값 사이의 연관성 측도이다. 자기 상관 함수로 계산된 k 개 시차 안에서의 자기 상관 값과 시차 상관 그림을 함께 그



Fig. 2. Three Main Patterns of the Auto-correlation (A) The Case of Indicates that There is an Autoregressive Term in the Data, (B) The Case of Indicates that There is a Higher Order Autoregressive Term in the Data, (C) The Case of Indicates that There is a Moving Average Term in the Data

린다. 이론적으로는 정상 시계열이면 확률값들의 계열의 모든 자기 상관계수는 0이 되어야 하지만, 현실의 유한표본 하에서는 이러한 경우가 거의 나타나지 않으므로 Fig. 2와 같은 시차 상관 그림을 통해 패턴을 구체적으로 살펴보게 되는데, 주기(cycle)를 자기 상관으로 정한 경우 시계열 자체의 비정상성(non-stationary)을 엿볼 수 있다. 그러나 확정적 신호(deterministic signal)나 랜덤 과정(random process)에 따라 자기 상관 함수가 다르게 정의될 수 있다는 어려움이 있으며, 무엇보다도 사용자가 자기 상관 함수를 이해하지 않으면 시각화 결과도 이해하기 쉽지 않다는 문제가 있다.

비정형 데이터 분석이 활성화된 이래로 통계분석 방법 외에도 군집 분석이나 규칙 기반 분석을 활용하여 구조 변화를 탐지하는 연구가 활발하게 진행되었다. Wei et al.[11]은 이메일 데이터로부터 토픽 모델링을 수행한 후, 토픽들을 주제별로 군집 분석을 한 후, 다음 토픽 기반으로 이메일 내용을 시각적으로 분석하는 연구를 수행하였다. Glance와 Ruge[12]은 특정 상품에 관한 온라인 리뷰를 분석하여 토픽을 추출, 토픽 맵을 구축한 다음, 사용자가 검색을 통해 상품에 대한 리뷰를 시각적으로 분석할 수 있는 연구를 진행하였다. 기계학습을 활용한 방법론의 개선을 위한 연구도 활발하였는데, Sawon P. et al.[13]은 서포트 벡터 머신, Cheung et al.[14]은 비음수 행렬 분해, Godfrey와 Gashler[15]은 신경망 모델을 이용한 시계열의 분해, 구조 변화를 살펴보는 방법을 각각 제안하였다.

3. 구조 변화 및 추세 검정

3.1 장기 시계열 안의 단기 구조 변화

시계열은 시간에 대한 난수의 순서로 정의할 수 있다. 구체적으로, 다음과 같은 확률과정으로 표현할 수 있다.

$$\{y(s, t), s \in \mathbb{N}, t \in \mathbb{T}\}$$

여기서 $t \in \mathbb{T}$, $y(\cdot, t)$ 는 확률공간 \mathcal{B} 상의 확률변수이고, 확률과정의 실현은 시간 $t \in \mathbb{T}$ 에 관한 각 $s \in \mathbb{N}$ 에 대해 $y(s, \cdot)$ 로 주어진다. 따라서, 우리가 실제로 관찰하는 데이터는, 알려지지 않은 확률과정의 실현, 즉 데이터 생성 과정이라고 할 수 있다.

$$\{y\}_{t=1}^T = \{y_1, y_2, \dots, y_t, \dots, y_{T-1}, y_T\},$$

$$(t = 1, \dots, T \text{INT})$$

시계열 분석의 한 가지 목적은, 이 데이터 생성 과정의 탐지와 관련이 있다. 이 과정은 이미 실현된 데이터로부터 기본 구조를 추론함으로써 진행된다. 추론된 구조가 정상 프로세스(stationary process)라면 다음과 같은 정의를 할 수 있다.

$$E[y_t] = \mu < \infty, \forall t \text{INT}$$

$$E[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j, \forall t, j \text{INT}$$

시계열에 대한 정상성을 더욱 엄격히 정의한다면, 다음과 같이 정의할 수도 있다.

$$F\{y_1, y_2, \dots, y_t, \dots, y_T\} = F\{y_{1+j}, y_{2+j}, \dots, y_{t+j}, \dots, y_{T+j}\}$$

여기서 $F\{\cdot\}$ 는 결합분포함수이다. 따라서, 프로세스가 유한한 모멘트로 엄격하게 고정되어 있다면, 공분산 또한 고정되어 있어야 한다.

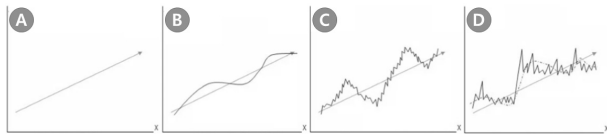


Fig. 3. Changes due to Variation Elements of Time Series Data, (A) Long-term trend, (B) (A) with Cyclical Variations, (C) (B) and Seasonal Variations, (D) (C) with Random Variations

그러나 대부분의 장기 시계열 데이터는 이러한 안정적인 생성 과정을 따르지 않는다.

시계열 자료의 변수는 장기간에 걸쳐 추세변동(trend variation), 순환변동(cyclical variation), 계절변동(seasonal variation), 그리고 불규칙변동(irregular variation)이 동시에 일어나기 때문에, 시계열이 장기적일수록 구조 추론이 점점 어려워지는 현상이 발생한다. 단기 시계열의 경우 안정적인 시계열로 간주하고 분석해도 큰 무리가 없는 반면, 장기 시계열의 경우 이러한 변동 요인들로 인해 Fig. 3과 같이 비정상(non-stationary)의 특징을 지니게 된다[16].

이러한 비정상 장기 시계열 안에서, 단기적으로 추세의 변화가 일시적인 것인지, 아니면 구조적으로 변한 것인지를 적시에 판단하는 것은 중요하다. 이는 시계열 구조의 변화를 상시 감지하여, 변화에 맞는 적절한 대응을 할 필요가 있기 때문이다.

본 연구에서는 장기 시계열이 주어진 상황에서, 단위근 검정법을 기반으로 Fig. 4와 같은 단기적 구조 변화를 감지하여, 이러한 변화가 얼마나 지속할 것인지를 시각적으로 판단할 수 있도록 하였다.

여기서 추세의 변화(trend change)란, 우선 시계열이 다음과 같다고 하자.

$$y_t = \beta_0 + \beta_1|\theta - t|_- + \beta_2|\theta - t|_+ + \epsilon_t$$

여기서,

$$|\theta - t|_- = \begin{cases} \theta - t & \text{if } t < \theta \\ 0 & \text{else} \end{cases}, |\theta - t|_+ = \begin{cases} \theta - t & \text{if } t > \theta \\ 0 & \text{else} \end{cases}$$

이때, $|\theta - t|_-$ 와 $|\theta - t|_+$ 가 같지 않다면, 이를 추세가 변화되었다고 하며, θ 가 변화 시점(change point)이 된다.

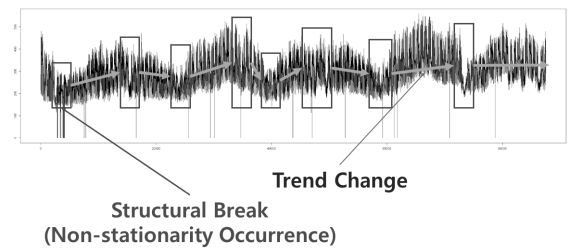


Fig. 4. Short-term Structural Changes in Long-term Time Series. Short-term Trends before and after Structural Changes Tend to Change

3.2 단위근 프로세스

대표적인 비정상 시계열은 랜덤워크(random walk) 과정으로부터 발생한다. 상수항 또는 추세선이 없는 랜덤워크를 따르는 시계열은 다음과 같이 표현할 수 있다.

$$y_t = y_{t-1} + \epsilon_t, \epsilon_t \sim N(0, 1)$$

$$E(y_t) = E[\epsilon_t + \epsilon_{t-1} + \dots] = 0$$

$$Var(y_t) = Var[\epsilon_t + \epsilon_{t-1} + \dots] = \sum_{t=1}^{INF} \sigma^2 = \infty$$

즉, 위 시계열은 분산이 무한히 커지면서 영구적 기억(infinite memory)을 갖는 특징이 있다. 위 시계열의 1차 차분값은 백색 잡음(white noise)을 가진다.

$$\Delta y_t = y_t - y_{t-1} = \epsilon_t$$

또한, 상수항을 갖는(추세선을 갖는) 랜덤워크 과정은 다음과 같이 표현할 수 있다.

$$y_t = a + y_{t-1} + \epsilon_t, \epsilon_t \sim N(0, 1)$$

$$E(y_t) = E[a + \epsilon_t + a + \epsilon_{t-1} + \dots] = \sum_{t=1}^{INF} a = \infty$$

$$Var(y_t) = Var[y_0 + a + \epsilon_1 + \dots + a + \epsilon_t] = \sum_{t=1}^{INF} \sigma^2 = \infty$$

즉, 위 시계열은 평균과 분산이 무한히 커지면서, 만일 상수항 a 가 0보다 크면 상방으로, 0보다 작으면 하방으로 흘러가게 되는 특징이 있다. 위 시계열은 확률적 추세를 갖는 시계열이라고 할 수 있다.

이 두 랜덤워크 과정은, 단위근(unit root)을 갖는 시계열의 예라고 할 수 있다.

$$y_t = \beta y_{t-1} + \epsilon_t, \quad -1 < \beta \leq 1$$

만일 $\beta=1$ 이라면, 위 랜덤워크 과정은 단위근을 가진다.

$$(1 - \beta L)y_t = a + \epsilon_t$$

여기서 L 은 lag operator ($L^n y_t = y_{t-n}$)이다.

단위근이라는 용어는 lag operator의 다항식의 근을 의미하는 것이다($1 - \beta L = 0$). 시계열에 따라서는 1개 이상의 단위근을 갖는 경우도 존재한다.

$-1 < \beta < 1$, $\beta \neq 1$ 인 경우는 정상(stationary) 시계열이다.

3.3 단위근 검정

시계열의 정상성(stationarity) 여부를 단위근 존재 여부를 이용하여 검정하는 방법을 단위근 검정 방법이라고 하며, autoregressive(AR)(1) 하에서의 방법은 다음과 같다.

$$y_t = \rho y_{t-1} + \nu_t, \quad \nu_t \sim N(0, \sigma_\nu^2)$$

AR(1) 모형은 랜덤워크와 달리 ρ 가 y_{t-1} 사이에 간혀있다. 여기서 $\rho = 1$ 이면 단위근을 가진다. 이때 y_t 는 상수항이 없는 랜덤워크 $y_t = y_{t-1} + \nu_t$ 이며, 비정상(non-stationary)이라고 할 수 있다. 반면에 $|\rho| < 1$ 이면, 정상(stationary) 시계열이다.

시계열 y_t 가 비정상인가 여부를 다음과 같은 가설을 통해 판단할 수 있다.

$$H_0 : \rho = 1, \quad H_1 : |\rho| < 1$$

이러한 검정을 위한 검정 통계량은 다음과 같은 다소 변형된 식으로부터 얻어진다.

$$y_t - y_{t-1} = \rho y_{t-1} - y_{t-1} + \nu_t$$

$$\Delta y_t = (\rho - 1)y_{t-1} + \nu_t = \gamma y_{t-1} + \nu_t$$

여기서 $\gamma \equiv \rho - 1$

$$\Rightarrow \begin{cases} H_0 : \rho = 1 \\ H_1 : \rho < 1 \end{cases} \text{ or } \begin{cases} H_0 : \gamma = 0 \\ H_1 : \gamma < 1 \end{cases}$$

Fig. 5는 단위근 검정 과정을 순서도로 도식화한 것이다[17].

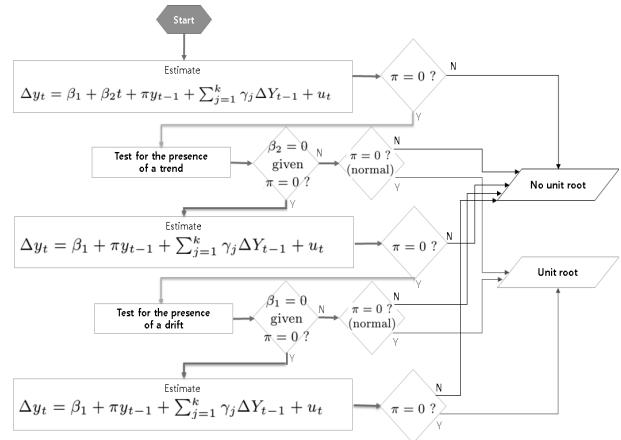


Fig 5. Procedure of Unit Root Test

1) Adjusted Dickey-Fuller 검정(ADF 검정)

주어진 시계열 y_t 의 비정상성 여부를, 이 y_t 를 차분하여 이를 다시 y_{t-1} 에 회귀하여 얻는 계수의 추정치가 0인지 여부를 검정하는 문제로 귀결된다. 그러나 통상적인 t 검정 통계량은 귀무가설($H_0 : \gamma = 0$) 하에서 t 분포를 따르지 않으며, 근사적으로도 정규분포를 따르지 않는다. γ 에 대한 t 값을 Dickey-Fuller (DF) 검정[18] 통계량이라고 하며, 이 통계량에 대한 임계값(critical value)을 Dickey and Fuller가 표로 제시하였다. 만일 귀무가설이 기각될 경우, 정상 시계열이라고 볼 수 있으며, 통상적인 t 검정을 적용할 수 있게 된다.

DF 검정은 랜덤워크 과정이 상수항을 가지는 경우, 비확률 추세를 포함하고 있는 경우 등을 고려하여 다음의 세 가지 경우에 대해 각각의 귀무가설을 검정할 수 있다.

$$\Delta y_t = \gamma y_{t-1} + \nu_t$$

$$\Delta y_t = \alpha_0 + \gamma y_{t-1} + \nu_t$$

$$\Delta y_t = \alpha_0 + \alpha_1 t + \gamma y_{t-1} + \nu_t$$

한편, DF 검정을 위한 위 세 가지 모형 설정 모두, 오차항에 자기 상관성이 있지 않다는 가정이 전제된다. 오차항에 자기 상관성이 있을 경우를 고려하기 위해 다음과 같은 모형을 설정하여 이루어지는 단위 검정을 adjusted DF 검정이라고 한다.

$$\Delta y_t = \pi y_{t-1} + \sum_{j=1}^k \gamma_j \Delta y_{t-j} + u_{1t}$$

$$\Delta y_t = \alpha_0 + \pi y_{t-1} + \sum_{j=1}^k \gamma_j \Delta y_{t-j} + u_{2t}$$

$$\Delta y_t = \alpha_0 + \alpha_1 t + \pi y_{t-1} + \sum_{j=1}^k \gamma_j \Delta y_{t-j} + u_{3t}$$

이때 귀무가설은 모두 $\gamma = 0$, 즉 해당 시계열이 비정상(non-stationary)이라는 것이고, 대립가설은 $\gamma < 0$, 즉 해당 시계열이 정상 시계열이라는 의미가 된다.

2) Phillips-Perron 검정(PP 검정)

DF 검정의 가정은, 오차항이 독립적이며 동일한 분포를 따른다는 것이다. 또한, ADF 검정은 설명변수에 시차를 갖는 분 값을 포함함으로써 자기 상관의 문제를 고려하고 있다.

PP 검정[19]은, 시차를 갖는 차분 값의 포함 없이 자기 상관을 고려하는 방법을 제시하였다. PP 검정은 다음 두 가지 회귀분식을 고려한다.

- i) $y_t = \mu + \alpha y_{t-1} + \epsilon_t$
 ii) $y_t = \mu + \beta[t - (1/2)T] + \alpha y_{t-1} + \epsilon_t$

위 Equation (i)에 대해서 검정 통계량은 다음과 같다.

$$\begin{aligned} Z(\hat{\alpha}) &= T(\hat{\alpha} - 1) - \hat{\lambda}'\bar{m}_{yy} \\ Z(\tau_{\hat{\alpha}}) &= (\hat{s}/\hat{\sigma}_{T\ell})t_{\hat{\alpha}} - \hat{\lambda}'_{T\ell}/\bar{m}_{yy}^{1/2} \\ Z(\tau_{\hat{\mu}}) &= (\hat{s}/\hat{\sigma}_{T\ell})t_{\hat{\mu}} - (\hat{\lambda}'_{T\ell}m_y)/(\bar{m}_{yy}^{1/2}m_{yy}^{1/2}) \end{aligned}$$

여기서 $m_y = T^{-3/2}\sum y_t$, $\bar{m}_{yy} = T^{-2}\sum (y_t - \bar{y})^2$, $m_{yy} = T^{-2}\sum y_t^2$, 그리고 $\hat{\lambda} = 0.5(\hat{\sigma}_{T\ell}^2 - \hat{s}^2)$ 이다.

또한, Equation (ii)에 대해서 검정 통계량은 다음과 같다.

$$\begin{aligned} Z(\tilde{\alpha}) &= T(\tilde{\alpha} - 1) - \tilde{\lambda}'/M \\ Z(t_{\tilde{\alpha}}) &= (\tilde{s}/\tilde{\sigma}_{T\ell})t_{\tilde{\alpha}} - \tilde{\lambda}'_{T\ell}/M^{1/2} \\ Z(t_{\tilde{\mu}}) &= (\tilde{s}/\tilde{\sigma}_{T\ell})t_{\tilde{\mu}} - (\tilde{\lambda}'_{T\ell}m_y)/[M^{1/2}(M + m_y^2)]^{1/2} \\ Z(t_{\tilde{\beta}}) &= (\tilde{s}/\tilde{\sigma}_{T\ell})t_{\tilde{\beta}} - \left[\tilde{\lambda}'_{T\ell} \left(\frac{1}{2}m_y - m_{ty} \right) \right] / [(M/12)^{1/2}\bar{m}_{yy}^{1/2}] \\ M &= (1 - T^{-2})m_{yy} - 12m_{ty}^2 + 12(1 + T^{-1})m_{ty}m_y \\ &\quad - (4 + 6T^{-1} + 2T^{-2})m_y^2 \end{aligned}$$

여기서 $m_{ty} = T^{-5/2}\sum ty_t$ 이다.

3) Elliott-Rothenberg-Stock 검정(ERS 검정)

앞선 두 가지 단위근 검정의 단점은 실제 데이터 생성 과정이 계수가 1에 가까운 AR(1) 과정이면 검정력이 낮아진다는 것이다. ERS 검정[20]은 Dickey-Fuller 단위근 검정을 변형시켜 검정력을 향상하려는 방법이다. 이 방법의 기각역은 Elliott, Rothenberg and Stock이 표로 제시하였다.

4) Schmidt-Phillips 검정(SP 검정)

DF 검정의 또 다른 단점은 불필요한 매개 변수(즉, 결정론적 회귀 계수)가 확정되지 않는다는 것 또는 확정되었다고 해도 대립가설하에서 다른 해석을 하게 된다는 것이다. Schmidt and Phillips[21]는 귀무가설과 대립가설하에서 동일한 일련의 불필요한 매개 변수를 정의하는 Lagrange multiplier (LM) 검정 방법을 제안했다. 또한, 이 검정 방법에서는 선형 추

세보다 높은 다항식을 고려한다.

$$\begin{aligned} y_t &= \alpha + Z_t\delta + x_t, \quad Z_t = (t, t^2, \dots, t^p) \\ x_t &= \pi x_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2) \end{aligned}$$

검정 통계량은 위 회귀 식을 실행함으로써 구성된다.

$$\Delta y_t = \Delta Z_t\delta + u_t$$

우선 $\tilde{\psi}_x = y_t - Z_t\bar{\delta}$ (여기서 $\bar{\delta}$ 는 δ 의 추정량)를 계산한다. 다음으로 $\tilde{S}_t = y_t - \tilde{\psi}_x - Z_t\bar{\delta}$ 를 정의한다. 마지막으로 검정을 위한 회귀 식을 다음과 같이 고려한다.

$$\Delta y_t = \Delta Z_t\gamma + \phi\tilde{S}_{t-1} + \nu_t, \quad \nu_t \text{는 오차항}$$

SP 검정의 검정 통계량은 $Z(\rho) = \tilde{\rho}/\hat{\omega}^2 = (T\tilde{\phi})/\hat{\omega}^2$ 이며, $\hat{\omega}^2$ 는 다음과 같이 계산한다.

$$\hat{\omega}^2 = \left[T^{-1}\sum_{i=1}^T \hat{\epsilon}_i^2 \right] / \left[T^{-1}\sum_{i=1}^T \hat{\epsilon}_i^2 + 2T^{-1}\sum_{s=1}^{\ell} \sum_{t=s+1}^T \hat{\epsilon}_t \hat{\epsilon}_{t-s} \right]$$

5) Kwiatkowski-Phillips-Schmidt-Shin 검정(KPSS 검정)

앞서 소개한 검정법들에서는 귀무가설이 단위근 과정이었지만 KPSS 검정[22]은 귀무가설이 정상 과정(stationary process)이다. 따라서 KPSS 검정을 한 결과 귀무가설을 기각하면, 시계열이 단위근을 가지고 있다는 결론을 내리게 된다. KPSS 검정에서는 다음과 같은 모델을 고려한다.

$$y_t = \xi t + r_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_u^2), \quad r_t = r_{t-1} + u_t$$

여기서 r_t 는 랜덤워크이다. $\xi = 0$ 이면 이 모델은 결정론적 회귀 변수만 남아 상수로 간주한다. 귀무가설 하에서 ϵ_t 는 고정되어 있으므로, y_t 는 추세가 고정된 경우, 즉 $\xi = 0$ 수준 하에서 고정된 경우가 된다.

먼저 레벨이나 추세를 테스트할지 여부에 따라 상수 또는 추세에 대해 y_t 의 회귀 식을 세운다. 그다음 이 식으로부터 잔차의 부분 합을 다음과 같이 계산한다.

$$S_t = \sum_{i=1}^t \hat{\epsilon}_i, \quad t = 1, 2, \dots, T$$

그러면 검정 통계량은 다음과 같이 구할 수 있다.

$$LM = \sum_{t=1}^T S_t^2 / \hat{\sigma}_{\epsilon}^2$$

여기서 σ_{ϵ}^2 는 위 단계에서 구한 오차 분산의 추정치이다.

3.4 추세 검정(Cox-Stuart 검정)

Cox-Stuart 검정[23]은 부호검정(sign test)의 변형된 기법으로, 추세(trend)가 존재하는지 여부를 파악하는 데 이용된다. 이때 추세는 처음 수치들보다 나중 수치들이 갈수록 커지는 상향 추세(upward trend)와, 갈수록 작아지는 하향 추세(downward trend)로 크게 구분된다. 추세를 파악하는 방법은, 전반부의 수치와 후반부의 수치를 쌍으로 형성한 다음 부호검정을 수행하는 과정으로 요약할 수 있다. 이때 만일 추세가 존재한다면, 쌍들 간에 상향 또는 하향의 경향을 발견할 수 있을 것이며, 그렇지 않다면 관측치들은 서로 독립이고 동일한 분포를 따른다고 말할 수 있게 될 것이다.

Cox-Stuart 검정은 추세가 존재하지 않는다는 귀무가설 하에서, “+” 혹은 “-” 부호는 파라미터 “n = ‘+’와 ‘-’의 독립인 관측치의 총수” 및 $p = 0.5$ 인 이항 분포를 갖는다는 사실에 기반하고 있다.

일련의 확률표본 y_1, y_2, \dots, y_{n^*} 를 n^* 이 짝수이면 $c = n^*/2$, 홀수이면 확률표본의 중앙값을 생략한 후 $c = (n^*+1)/2$ 로 간주한 후, $(y_1, y_{1+c}), (y_2, y_{2+c}), \dots, (y_{n^*-c}, y_{n^*})$ 와 같은 방법으로 쌍을 형성한다. 그리고 임의의 쌍 (y_i, y_{i+c}) 은 만일 $y_i < y_{i+c}$ 이면 “+”를, $y_i > y_{i+c}$ 이면 “-”를, $y_i = y_{i+c}$ 이면 “0”인 동점 값을 가지므로, 총 관측 수 n^* 로부터 동점 수를 제거한 후, 실질적으로 이용하는 관측치는 “n = ‘+’와 ‘-’의 총수”라고 정의한다.

$$H_0 : P(+) = P(-), H_1 : P(+) \neq P(-)$$

검정통계량 T 는 확률표본 (y_i, y_{i+c}) 의 쌍에서 $y_i < y_{i+c}$ 인 쌍의 수, 또는 “+” 쌍의 수로 정의한다. 이제 n^* 및 $p = 0.5$ 의 이항분포로부터 유의수준 α 가 주어졌을 때, $\alpha/2$ 에 해당하는 분위수를 각각 $x_{\alpha/2}$ 라 하면, $T < x_{\alpha/2}$ 또는 $T > n^* - x_{\alpha/2}$ 이면 H_0 를 기각한다.

3.5 단기적 구조 변화 감지를 위한 시각화

장기 시계열이 주어진 상황에서, 단기적으로 구조 변화를 감지하기 위해서는, 우선 주어진 시계열 데이터를 Fig. 6과 같이 분할한 후 검정을 진행해야 한다. 그러나 이 길이와 간격을 어떻게 정하여 분할해야 하는지에 대한 명확한 기준은 없다. 물론 정상(stationary) 시계열이라면 검정력(power)을 기준으로 표본 수를 산출하는 방법으로 분할할 수 있지만, 실존하는 대부분의 시계열은 비정상(non-stationary)이며, 도메인과 데이터 특성에 따라 달라지는 등의 원인으로 하여금 모든 데이터에 통용되는 방법은 존재하지 않는다. 따라서 partial data length와 step size에 대해 최적화를 진행하거나 사용자의 선택에 맡겨야 한다. 사용자의 선택에 맡기는 방법의 하나로는 상호작용적 시각화(interactive visualization)로 하여금 사용자의 정보에 대한 인지적, 지각적 요인을 활용하도록 할 수 있다. 주의할 점은 partial data(부분 데이터)

length는 step size보다 크거나 같게 선택하도록 해야 한다. 그렇지 않으면 “[partial data length - step size] × partial data 개수”만큼의 분석하지 않는 부분이 발생하게 되기 때문이다.

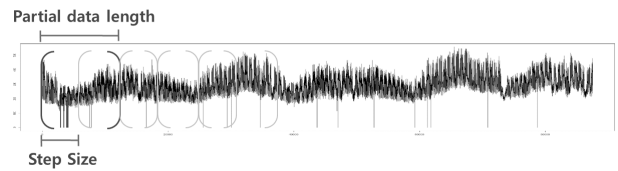


Fig. 6. Segmentation using Partial Data Length and Step Size of Time Series Data

부분 데이터를 추출하고 나서 각 데이터마다 단위근 검정을 한다. 만일 특정 partial data의 단위근 검정 결과 비정상 시계열일 가능성이 클 경우, Fig. 7과 같이 해당 부분을 강조한다.

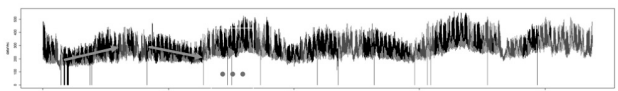


Fig. 7. Based on the Results of the unit Root Test, some Partial Data are Emphasized

강조된 부분의 직전과 직후의 강조되지 않은 두 부분을 서로 비교하여, 추세를 비교한다. 만일 추세가 변화한 것으로 판단되면, 장기 시계열 안의 구조 변화(structural break) 주기와 그 전후의 추세 패턴 변화를 발견한 것으로 볼 수 있다. 추세 변화를 확인하기 위해서는 Cox-Stuart 검정, Mann-Kendall 검정과 같은 추세 검정 방법을 사용할 수 있다.

4. 시각화 및 성능 평가 실험

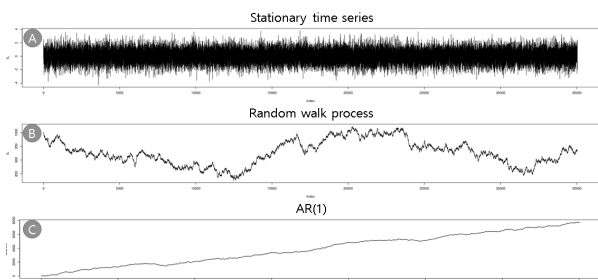
4.1 Prerequisite

1) 임의 데이터(synthetic data) 생성을 위한 기본 모형
임의 데이터 생성을 위한 기본 모형으로 stationary and random time series, random walk process, first order auto-regression(AR)(1) process 다음 3가지 타입을 고려하였다. 이를 각각 시각화하면 Fig. 8과 같이 나타난다.

a) Type 1: Stationary and random time series

$$y_t = \mu + \epsilon_t, \epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Type 1은 관측치가 일정한 평균 주위에서 변동하고, 연속적인 분산과 확률적으로 독립인 시계열이다. 관측치가 위 또는 아래로 향하는 추세가 없으며, 시간이 지남에 따라 분산이


 Fig. 8. Synthetic Data using the Base Model ($n = 35,040$)

- (A) Type 1. $\mu = 0, \sigma^2 = 1,$
 (B) Type 2. $y_0 = 1000, \mu = 0, \sigma^2 = 1$
 (C) Type 3. $\alpha = 1, \mu \sim Unif(0, 3), \sigma^2 = 5$

증가 또는 감소하지도 않는다.

b) Type 2: Random walk process

Type 2는 일정한 평균과 분산을 가지는 랜덤 프로세스이다.

$$y_t = y_{t-1} + z_t = y_0 + \sum_{i=1}^t z_i, \quad z_t \sim N(\mu, \sigma_z^2)$$

c) Type 3: First order auto-regression(AR)(1) process

Type 3는 시간의 흐름에 따른 변화를 고려한 AR(1) 프로세스이다.

$$y_t = \alpha y_{t-1} + z_t, \quad z_t \sim N(\mu, \sigma_z^2)$$

2) 이상치(Outlier) 모형

앞서 소개한 기본 모형은 정상(stationary) 시계열이므로 이를 그대로 분석에 활용하는 것은 본 논문에서 의미가 없다. 따라서 인위적으로 비정상성(non-stationary)을 부여할 필요가 있다.

ARIMA(p, d, q) 모형이 다음과 같다고 하자.

$$\begin{aligned} \Phi_p(L)\Delta^d y_t &= \Theta_q(L)\epsilon_t \\ \Leftrightarrow \Phi_p(L) \cdot \alpha_d(L)y_t &= \Theta_q(L)\epsilon_t \\ \Leftrightarrow y_t &= \frac{\Theta_q(L)}{\alpha_d(L)\Phi_p(L)}\epsilon(L) \equiv L_{p,d,q}(L)\epsilon_t \end{aligned}$$

여기서 $\Delta^d = (1-L)^d = \alpha_d(L)$

여기서 시간 $t = t_0$ 에서 이상치가 발생한 경우, 일반적인 이상치 모형은 다음과 같다.

$$z_t = y_t + \omega_0 \frac{\Phi_q(L)}{\alpha_d(L)\Theta_p(L)} I_t(t_0) = y_t + \omega_0 K_{p,d,q}(L) I_t(t_0)$$

여기서 ω_0 는 y_{t_0} 로부터의 편차, $I_t(t_0)$ 는 지시함수로

Table 1. Definition of 5 Types Outliers

Additive outliers (AO)
If $K_{p,d,q}(L) = 1,$ $z_t = \begin{cases} y_t + \omega_0 I_t(t_0) & m = 1 \\ y_t + \sum_{i=1}^m \omega_i I_t(t_i) & m > 1 \end{cases}, \quad \omega_i, i = 1, 2, \dots, m$
Innovational outliers (IO)
If $K_{p,d,q}(L) = \frac{\Phi_q(L)}{\alpha_d(L)\Theta_p(L)},$ $\begin{aligned} z_t &= y_t + \omega_0 K_{p,d,q}(L) I_t(t_0) \\ &= K_{p,d,q}(L)\epsilon_t + \omega_0 K_{p,d,q}(L) I_t(t_0) \\ &= K_{p,d,q}(L) [\epsilon_t + \omega_0 I_t(t_0)] \end{aligned}$
Level shift outliers (LSO)
If $K_{p,d,q}(L) = \frac{1}{1-L},$ $z_t = y_t + \omega_0 \frac{1}{1-L} I_t(t_0)$ $= y_t + \omega_0 I_t([t_0, \infty))$ $I_t([d, \infty)) = \begin{cases} 1 & t \in [d, \infty) \\ 0 & t \notin [d, \infty) \end{cases} = \begin{cases} 1 & t \geq d \\ 0 & t < d \end{cases}$
Temporary change outliers (TCO)
If $K_{p,d,q}(L) = \frac{1}{1-\delta L},$ $z_t = y_t + \frac{1}{1-\delta L} \omega_0 I_t(t_0), \delta \in (0, 1)$
Variance change (VC)
$\sigma_\omega^2 \rightarrow (1-\omega_0)\sigma_\omega^3 I_t(t_0)$

$I_t(t_0) = \begin{cases} 1 & t = t_0 \\ 0 & otherwise \end{cases}$, 그리고 $\Phi_p(L), \Theta_q(L)$ 은 시차 다항식 (lag polynomial)이다.

$$\begin{aligned} \Phi_p(L) &= \phi_0 L^0 - \phi_1 L^1 - \dots - \phi_p L^p \\ \Theta_q(L) &= \theta_0 L^0 + \theta_1 L^1 + \dots + \theta_q L^q \\ &= \theta'_0 L^0 - \theta'_1 L^1 - \dots - \theta'_q L^q \end{aligned}$$

이를 바탕으로, ARIMA(p, d, q) 모형에 기반하여 5가지 형태의 이상치(outlier)를 부여하였으며, Table 1과 같이 정리하였다.

가법적 이상치(additional outliers; AO)는 시간 $t = t_0$ 에서 시계열 자료 중 한 개가 지나치게 크거나 작은 값을 갖는 관측치로, 기록이나 단위 조작 등 개인적 실수로 야기되는 이상치를 의미한다.

혁신적 이상치(innovational outliers; IO)는 $t = t_0$ 이후의 시계열이 전혀 다른 형태를 보이는 특징을 가지는 이상치이다.

수준 이동 이상치(level shift outlier; LSO)는 $t = t_0$ 이후의 시계열 전체가 위나 아래로 이동한 경우이다.

일시적 변화 이상치(temporary chance outliers; TCO)는

$t = t_0$ 에 일시적으로 시계열이 이동하였으나, 지수적으로 빠르게 원래의 상태로 돌아가는 시계열의 형태를 의미한다. 만약 $\delta = 0$ 이면, 이상치의 모형은 AO가 되고 $\delta = 1$ 이면 LSO가 된다.

분산 변화(variance change; VC)는 $t = t_0$ 이후에 분산이 일정하게 커지거나 작아진 경우를 의미한다. 물론 시간에 따라 분산이 점점 더 증가/감소하는 경우도 있을 것이다.

위 정의에 따르면, AO를 제외한 모든 이상치의 발생은 곧 structural break의 발생이라고 볼 수 있다.

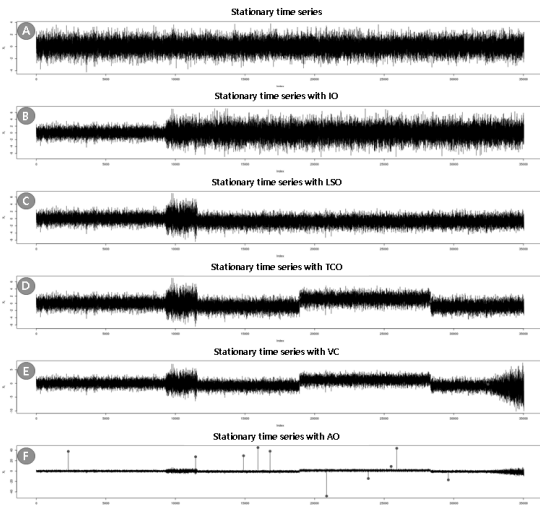


Fig. 9. An Example of Type 1 Synthetic Data ($n = 35,040, \xi = 1000$)

- (A) $y_t = \mu + \epsilon_t, \epsilon_t \sim N(0, \sigma^2), \mu = 0, \sigma^2 = 1$
- (B) (A) with IO. $\sigma^2 \sim Unif(0, 3)$ where $t \geq IDX_{IO}$
- (C) (B) with LSO. $\mu \sim Unif(-3, 3)$ where $t \geq IDX_{LSO}$
- (D) (C) with TCO. $y_t = y_t + Unif(-3, 3)$
where $IDX_{TCO_1} \leq t \leq IDX_{TCO_2}$
- (E) (D) with VC. $\sigma_t^2 = \sigma_t^2 + \frac{3-1}{n - IDX_{VC} + 1}$
where $t \geq IDX_{VC}$
- (F) (E) with AO. $y_t = y_t + 10 \times Unif(-5, 5)$
where $t = IDX_{AO}, \zeta = 10$

Fig. 9, Fig. 10은 각각 Type 1, Type 2에서 5가지 이상치가 차례대로 부여되는 모습을 나타내는 것이며, 각 이상치의 부여 시점(IDX)은 다음과 같은 방법으로 랜덤하게 차례대로 결정되도록 하였다. AO는 전체 구간 중 ζ 번의 랜덤한 위치마다 생성된다.

$$IDX = \left\{ \begin{array}{l} IDX^{(1)} = IDX_{IO}, \quad IDX^{(2)} = IDX_{LSO} \\ IDX^{(3)} = IDX_{TCO_1}, \quad IDX^{(4)} = IDX_{TCO_2} \\ IDX^{(5)} = IDX_{VC}, \quad IDX_{AO} \end{array} \right\}$$

$$IDX^{(k)} \sim \begin{cases} [Unif(0, n)] & k = 1 \\ [Unif(Idx^{(k-1)}+1, n-\xi)] & k > 1 \end{cases}$$

$k = 1, 2, \dots, 5$

ξ 는 IDX 가 n 과 같아지는 것을 막기 위한 양의 정수

$$IDX_{AO} = \{ IDX_{AO}^{(1)}, IDX_{AO}^{(2)}, \dots, IDX_{AO}^{(\zeta)} \}$$

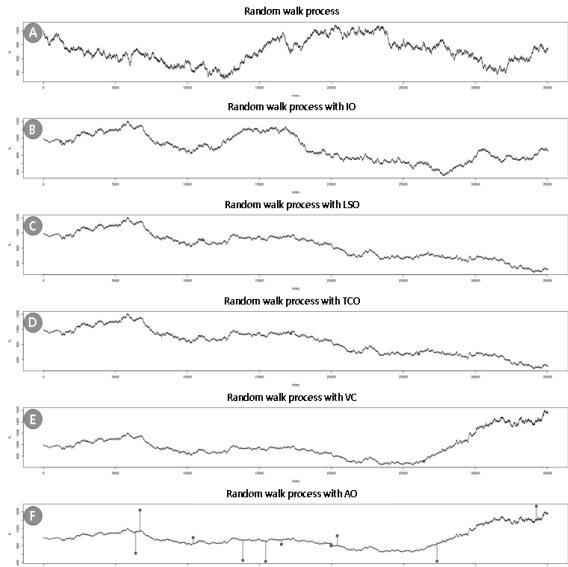


Fig. 10. An Example of Type 2 Synthetic Data ($n = 35,040, \xi = 1000$)

- (A) $y_0 = 1000, \mu = 0, \sigma^2 = 1$
- (B) (A) with IO. $\sigma^2 \sim Unif(0, 3)$ where $t \geq IDX_{IO}$
- (C) (B) with LSO. $\mu \sim Unif(-3, 3)$ where $t \geq IDX_{LSO}$
- (D) (C) with TCO. $y_t = y_t + Unif(-3, 3)$
where $IDX_{TCO_1} \leq t \leq IDX_{TCO_2}$
- (E) (D) with VC. $\sigma_t^2 = \sigma_t^2 + \frac{3-1}{n - IDX_{VC} + 1}$
where $t \geq IDX_{VC}$
- (F) (E) with AO. $y_t = y_t + 10 \times Unif(-5, 5)$
where $t = IDX_{AO}, \zeta = 10$

$$IDX_{AO}^{(\ell)} \sim [Unif(0, n)], \ell = 1, 2, \dots, \zeta$$

3) 추세(trend)의 변화

Type 1은 추세가 없고 Type 2는 추세 변화가 랜덤이므로, 추세 검정 평가를 하는 데 적절하지 않다. 그러나 Type 3의 경우 μ 가 양수일 경우 우상향, 음수일 경우 우하향하는 추세가 만들어진다. 따라서 μ 를 변경하는 방법으로 추세가 변화하는 구조 변화(structural break) 시점을 ζ 번 생성할 수 있다. Fig. 11은 구조 변화가 9번 발생한다는 가정하에서 생성한 것이다.

Table 2. Parameters for analysis of Type 1^(F), Type 2^(F) and Type 3^(B) (CV is critical value)

Partial data length				Step size			
96×14=1,344				96×3.5=336			
Parameters of Unit Root Test for Type 1 ^(F)							
		CV	Lag			CV	Lag
ADF	Trend	-3.13	96	ERS DG	Trend	-2.57	48
	Drift	-2.57	96		Const.	-1.62	96
PP	Trend	-3.14	4	ERS P	Trend	6.89	12
	Const.	-2.57	4		Const.	4.48	12
SP	Tau	.0001	12	KPSS	Tau	0.18	12
	Rho	.0001	12				
Parameters of Unit Root Test for Type 2 ^(F)							
		CV	Lag			CV	Lag
ADF	Trend	-3.13	96	ERS DG	Trend	-3.48	12
	Drift	-2.57	96		Const.	-2.57	12
PP	Trend	-3.99	12	ERS P	Trend	3.96	12
	Const.	-3.45	12		Const.	1.99	12
SP	Tau	.0001	12	KPSS	Tau	0.12	12
	Rho	.0001	12				
Parameters of Unit Root Test for Type 3 ^(B)							
		CV	Lag			CV	Lag
ADF	Trend	-3.42	96	ERS DG	Trend	-2.89	12
	Drift	-2.87	96		Const.	-1.94	12
PP	Trend	-3.42	12	ERS P	Trend	5.62	12
	Const.	-2.87	12		Const.	3.26	12
SP	Tau	.0001	12	KPSS	Tau	0.15	12
	Rho	.0001	12				
Significance Probability of Cox-Stuart Trend Test							
.001							

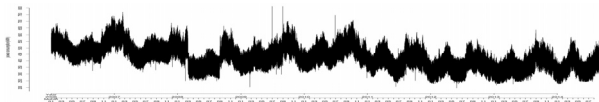


Fig. 11. Power Consumption Data of Montenegro by ENTSO-E

$$IDX_{TC}^{(c)} \sim \begin{cases} [Unif(0, c)] & c = 1 \\ [Unif(IDX_{TC}^{(c-1)} + 1, n - \xi)] & c > 1 \end{cases}$$

$$c = 1, 2, \dots, \zeta$$

ξ는 $IDX_{TC}^{(c)}$ 가 n과 같아지는 것을 막기 위한 양의 정수

4) 평가(evaluation)

사용자는 partial data length, step size, 단위근 검정에 대한 critical value(significance probability), lags를 결정한 후, 구조 변화 시점이 부분 데이터에 포함될 경우, 해당 부분 데이터에 대한 단위근 검정 결과 비정상성(non-stationary) 시계열이라는 결론을 내리길 기대할 것이다. 즉, 실제 구조 변화가 나타난 시점을 IDX , 단위근 검정을 통해 비정상 시계열이라고 판단된 부분 데이터의 시점을 \widehat{IDX} 라고 할 때, 다음과 같이 정의되는 IDX Gap이 0이 된다면 최적의 파라미터를 결정할 것이라고 할 수 있다.

$$IDX \text{ Gap: } \min_{\max(IDX)} \sum [|IDX - \widehat{IDX}|]$$

다만 이러한 평가 측정은 구조 변화 시점이 명확한 경우에만 가능하다. 따라서 임의 데이터를 이용하여 분석할 경우 인위적으로 구조 변화 시점을 지정해주었으므로 평가 측정이 수월하나, 실제 데이터를 이용할 경우 평가 측정이 어려울 수 있다는 한계가 있다.

4.2. 임의 데이터를 이용한 실험

1) 시각화

임의 데이터 Type 1(F)(Fig. 9. (F)), Type 2(F)(Fig. 10. (F)), Type 3(B)(Fig. 11. (B))에 대하여 파라미터를 Table 2와 같이 설정한 후 시각화를 진행하였으며, 그 결과 Fig. 12와 같이 나타났다.

단위근 검정 결과는 시계열 선 그래프 안에서 빨간색으로 강조되도록 하였으며, 추세 검정 결과는 증가 추세로 판단될 경우 붉은 배경색, 하락 추세로 판단될 경우 푸른 배경색으로 표시되도록 하였다.

각 결과에 대한 IDX Gap을 측정한 결과는 Table 3의 [A]와 같다. Fig. 13과 함께 살펴보면, Type 1(F), Type 2(F), Type 3(B) 모두에서 공통적으로 SP 검정, KPSS는 IO, LSO, TCO, 혹은 VC에 따른 구조변화를 가리지 않고 잘 탐지하는 모습을 보였으며, ADF 검정과 PP 검정은 LSO와 TCO, ERS DG 검정은 VC에 따른 구조 변화를 비교적 잘 탐지하는 모습을 볼 수 있었다. 다만 Type 1(F)의 경우 ADF 검정과 ERS DF-GLS trend 검정은 구조변화라고 볼 수 없는 AO까지 구조변화로 판단하는 모습을 볼 수 있으며, Type 2(F)의 경우 대부분의 검정 방법이 AO를 구조 변화로 판단하는 모습을 볼 수 있었다. AO에 대해 Type 1(F)과 Type 2(F)의 구조변화 탐지 성능이 서로 다른 것은, Type 1(F)의 경우 추세가 없으므로 이상치 발생 이후의 구조 변화를 탐지하는 것이 더 수월하기 때문으로 여겨진다.

2) 파라미터의 평가

같은 조건으로 데이터를 생성하되, 난수 생성 시드를 달리

Table 3. [A] IDX Gap of Results by Analysis and [B] Sum of IDX Gap of Results by 100 Times Analysis (right) for Type 1^(F), Type 2^(F) and Type 3^(B)

Unit Root Test		[A] IDX Gap					[B] Sum of IDX Gap of results by 100 times												
		IO	LSO	TCO ⁽¹⁾	TCO ⁽²⁾	VC	IO	LSO	TCO ⁽¹⁾	TCO ⁽²⁾	VC								
Type 1 ^(F)																			
ADF	Trend	1	20	0	10	19	1490	1000	1217	1266	1201								
	Drift	0	19	0	0	9	616	453	473	554	680								
PP	Trend	0	0	0	0	0	0	0	3	2	1								
	Const.	0	0	0	0	0	1490	1000	1217	1266	1201								
ERS DG	Trend	10	31	18	4	7	0	0	3	1	1								
	Const.	0	2	22	4	0	1670	1424	1597	1351	1346								
ERS P	Trend	0	0	0	0	0	0	16	23	10	1								
	Const.	0	0	0	0	0	0	13	20	17	1								
SP	Tau	0	0	0	0	0	7	2	6	5	4								
	Rho	0	0	0	0	0	7	2	6	5	4								
KPSS	Tau	0	0	0	0	0	0	19	26	20	2								
Type 2 ^(F)																			
ADF	Trend	2	1	11	0	0	1235	804	867	879	851								
	Drift	0	1	11	0	3	741	624	623	713	1056								
PP	Trend	32	2	4	3	0	1106	1089	1018	1161	1211								
	Const.	52	4	4	6	7	1235	804	867	879	851								
ERS DG	Trend	33	26	10	11	7	1347	991	1035	1281	1778								
	Const.	53	0	17	7	11	3258	2769	2940	3811	5659								
ERS P	Trend	4	29	11	63	59	3909	2005	2934	3249	4324								
	Const.	39	0	1	7	11	2092	1561	1983	2795	4019								
SP	Tau	3	0	3	4	0	591	626	644	735	821								
	Rho	3	0	3	4	0	592	625	639	735	821								
KPSS	Tau	18	0	2	1	5	745	655	714	703	963								
Type 3 ^(B)																			
c of $IDX_{TC}^{(c)}$		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
ADF	Trend	54	38	14	10	10	7	13	15	0	2217	2014	1614	1673	2069	1924	2164	2396	2268
	Drift	4	2	0	45	45	50	27	26	6	2424	1788	1671	1881	1999	2196	2065	2093	2464
PP	Trend	1	3	19	23	23	4	0	0	0	804	717	702	778	918	901	757	809	1052
	Const.	80	64	0	3	3	15	0	0	11	2217	2014	1614	1673	2069	1924	2164	2396	2268
ERS DG	Trend	15	32	7	3	3	7	0	0	0	1484	1107	1139	1184	1336	1177	1219	1195	1670
	Const.	19	3	2	36	36	21	0	0	11	3152	2634	2334	2524	2447	2150	2513	2582	3358
ERS P	Trend	15	3	5	3	3	7	0	0	0	1705	1259	1195	1127	1178	972	1098	1116	1088
	Const.	19	3	2	4	4	14	0	0	11	1727	1324	1247	1335	1214	1000	1092	1362	1796
SP	Tau	11	3	9	5	5	13	0	0	0	1396	1046	1274	1221	1377	1146	1043	1177	1473
	Rho	11	3	9	5	5	13	0	0	0	1416	1053	1281	1221	1380	1142	1046	1194	1475
KPSS	Tau	1	18	5	5	5	11	0	0	2	514	471	536	496	468	485	417	450	399

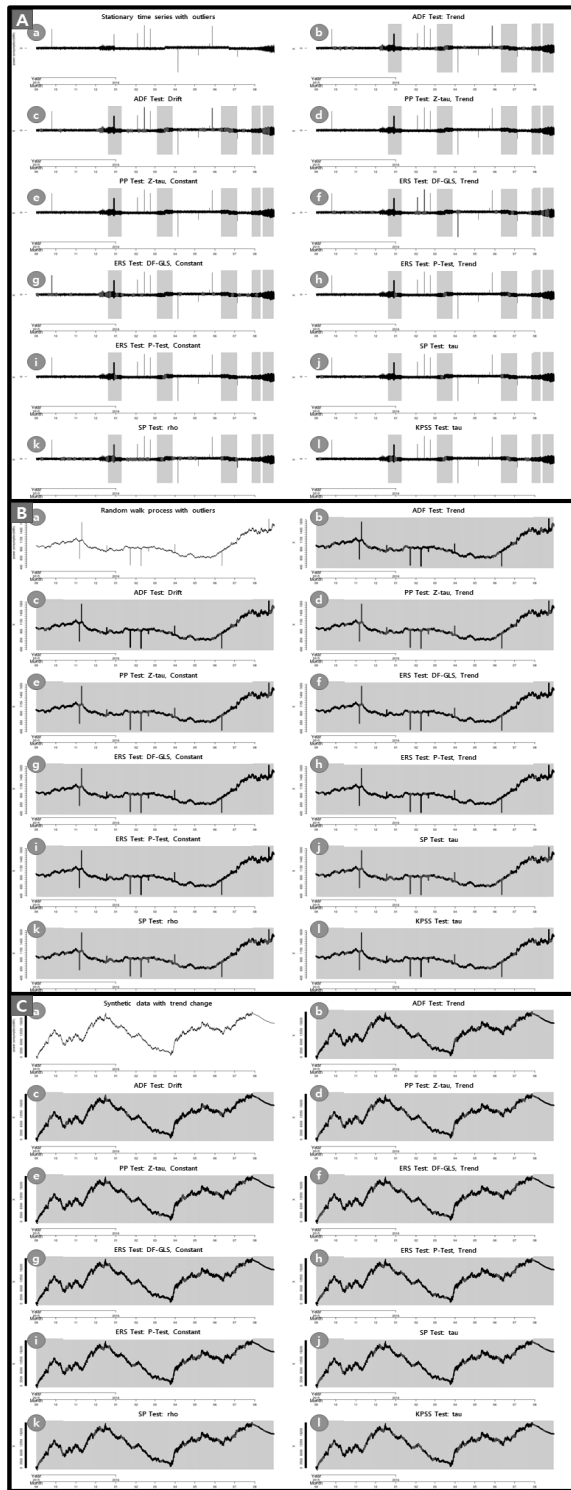


Fig. 12. Visualizations of Analysis of [A] Type 1(F), [B] Type 2(F) and [C] Type 3(B)

- (a) Line Graph (b) (a) with ADF Trend
- (c) (a) with ADF Drift (d) (a) with PP Trend
- (e) (a) with PP Const. (f) (a) with ERS DF-GLS Trend
- (g) (a) with ERS DF-GLS Const. (h) (a) with ERS P Trend
- (i) (a) with ERS P Const. (j) (a) with SP Tau
- (k) (a) with SP Rho (l) (a) with KPSS Tau

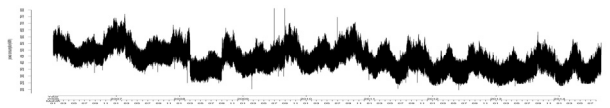


Fig. 13. Power Consumption Data of Montenegro by ENTSO-E

하여 동일한 분석을 100회 반복하여 각각의 IDX Gap을 구한 뒤 그의 합을 Table 3의 [B]와 같이 정리하였으며, 위 1)에서 얻은 결론을 대부분 뒷받침하는 결과를 얻을 수 있었다.

각 검정 방법마다 전제된 가정이 다른 만큼 비정상성 여부를 판단하는 결과도 달라질 것이고, 이에 시계열 데이터의 도메인이나 특성에 따라 적용해야 하는 단위근 검정 방법이 달라질 것이므로, 한두 가지 검정 방법만 집중하도록 하는 것은 무리가 있을 수 있다. 따라서 이처럼 각종 단위근 검정을 모두 수행하여 결과를 비교하거나, 혹은 어떠한 변동 요소에 주목할 것인지를 선택하여 검정을 수행할 수 있다. 사용자는 파라미터를 바꿔가며 각 검정 결과에 따른 시각화 모습을 인터랙티브하게 살펴봄으로써 검정 방법 선택을 더욱 쉽게 진행할 수 있다.

임의 데이터를 이용한 실험을 통해, 데이터의 분포 형태 및 구조 변화 유형에 따른 각 검정 방법의 특성을 파악할 수 있었다. 그러나 실제 데이터에 바로 적용할 때는 각 검정 방법이 어떠한 특징을 보일지 미리 가늠하기가 쉽지 않다. 이에 사용자는 구조 변화 시점을 명확하게 알 수 없는 실제 데이터를 분석하기에 앞서서, 데이터의 형태와 유사한 임의 데이터를 먼저 생성, 위와 같은 반복 분석을 통해 적절한 파라미터를 선정한 후, 실제 데이터의 분석에 활용하는 것 또한 방법이 될 수 있다.

Table 4. Parameters for Power Consumption of Montenegro by ENTSO-E (CV is critical value)

Partial data length		Step size					
24 × 7 = 268		24					
Parameters of Unit Root Test							
		CV	Lag		CV	Lag	
ADF	Trend	-3.43	24	ERS DG	Trend	-3.46	4
	Drift	-3.46	24		Const.	-2.58	4
PP	Trend	-4.02	24	ERS P	Trend	4.05	24
	Const.	-3.47	24		Const.	1.91	24
SP	Tau	.0001	12	KPSS	Tau	0.12	24
	Rho	.0001	12				
Significance Probability of Cox-Stuart Trend Test							
.001							

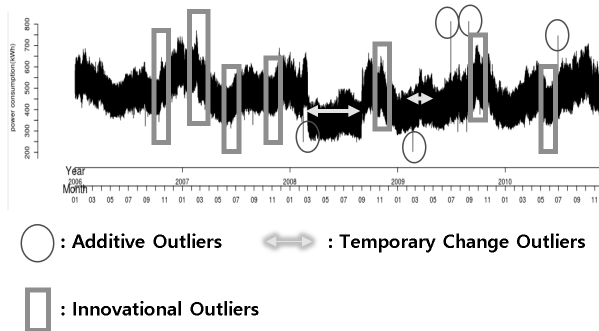


Fig. 14. Outliers and Structure Breaks Intuitively Confirmed from Power Consumption(kWh) Data of Montenegro by ENTSO-E

4.3 ENTSO-E의 몬테네그로 통신원 전력 소비 데이터

ENTSO-E는 유럽 연합의 에너지 정책 이행을 지원하고, 유럽의 에너지 및 기후 정책 목표를 달성하기 위해 유럽 TSO(transmission system operators)가 참여하고 있는 협의 기구이다. 유럽 연합 내의 각기 다른 사업체들로 구성된 TSO는 주요 고전압 전력망에서 대량의 전력 송신을 담당한다. 2018년 12월 기준으로 총 38개의 통신원(correspondant)이 협업하고 있으며, 매월 전력 소비/생산 데이터, 매년 순 생산량(net generating capacity), 전송 내역 데이터를 ENTSO 홈페이지를 통해 공개하고 있다. 본 실험에서는 “몬테네그로” 통신원에서 제공한 2007년 1월 1일부터 2015년 12월 31일까지 1시간 단위 전력 사용량(consumption) 데이터를 사용하였다. 모든 값을 선 그래프로 그리면 Fig. 13과 같은 모습을 확인할 수 있다.

전력데이터의 특성상 계절에 따라 추세가 변화하는 특징이 보이며, 큰 흐름을 봤을 때 혁신적 이상치(IO)가 발생할 시 추세 전환이 일어나며, 상승 추세일 때는 대체로 점진적으로 분산이 커지는(VC), 하락 추세일 때는 대체로 점진적으로 분산이 작아지는 형태(VC)로 보인다.

분석을 위한 파라미터를 Table 4와 같이 설정하였으며, 시각화 결과는 Fig. 15와 같이 나타났다. Cox-Stuart 추세 검정 결과는 추세에 맞는 배경색이 대체로 잘 나타난 것으로 보이며, 2008년 2~4월, 2009년 7~8월, 그리고 2010년 7~8월 사이에서는 대부분의 단위근 검정 결과 비정상성(non-stationary)이 탐지되는 모습이 나타났다. 이는 2008년의 코소보 독립선언, 2009년 제3차 에너지 패키지(Directive 2009/72)를 통한 전력시장 자유화, 2010년의 폭우 등이 전력 소비량에도 반영된 것으로 보이며, 이에 따른 비정상성이 탐지된 것으로 보인다.

다만 SP tau 검정, KPSS 검정의 경우 다른 검정에 비해 매우 빈번하게 비정상성을 탐지하는 모습이 보였다. 이는 데이터를 자세히 살펴보면 Fig. 14와 같이 여러 가지 이상치가 혼재한 것을 볼 수 있으며, 앞서 임의 데이터 실험에서 이 두 검정이 구조 변화의 유형을 가리지 않고 모두 탐지하던 특징이 실제 데이터에서도 나타난 것으로 보인다.

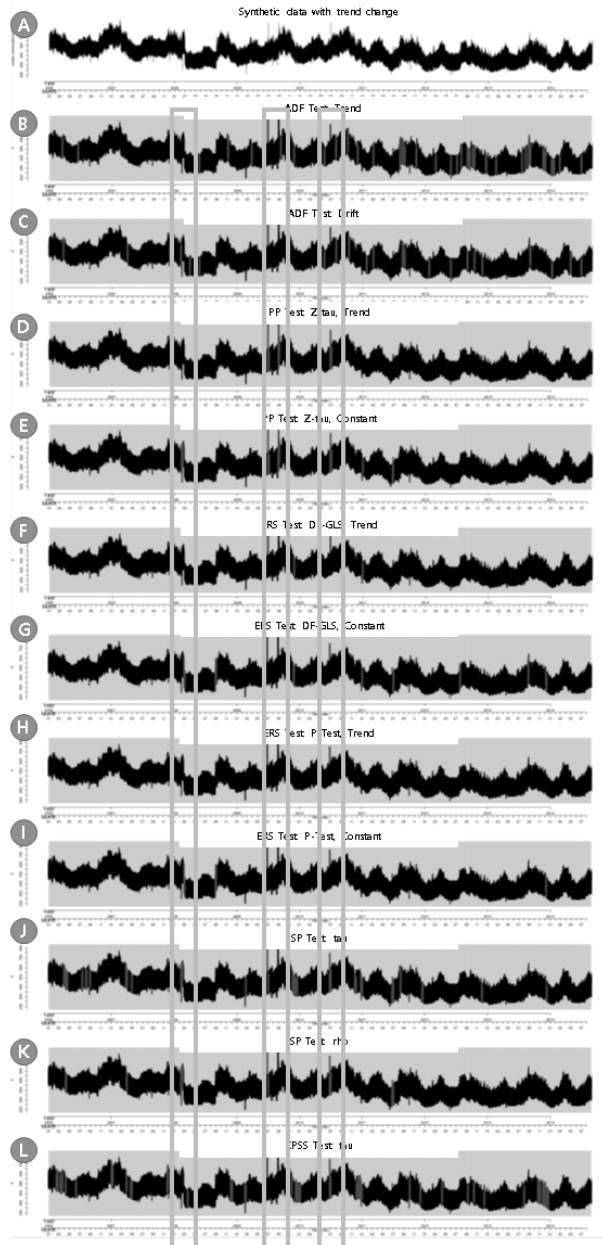


Fig. 15. Visualizations of Analysis of Power Consumption(kWh) Data of Montenegro by ENTSO-E

- (A) Line Graph (B) (A) with ADF Trend
- (C) (A) with ADF Drift (D) (A) with PP Trend
- (E) (A) with PP Const (F) (A) with ERS DF-GLS Trend
- (G) (A) with ERS DF-GLS Const (H) (A) with ERS P Trend
- (I) (A) with ERS P Const (J) (A) with SP Tau
- (K) (A) with SP Rho (L) (A) with KPSS Tau

5. 결론

본 연구에서는 장기 시계열 데이터를 분리하여 각 부분 데이터마다 단위근 검정을 수행, 그 결과를 시각적으로 표현하여 단기 구조 변화를 파악하는 방법을 제시하였다. 단위근 검

정 결과 비정상(non-stationary)적으로 나타난 시점의 전후 기간에 대한 추세 검정을 한 결과, 비정상성이 나타난 시점을 기준으로 추세 변화가 나타나는 모습을 시각화를 통해 효과적으로 파악할 수 있었다. 사용자는 이러한 시각화를 바탕으로 단기적인 변화에 따른 앞으로의 추세 변화를 예측, 대응할 수 있을 것이다.

다만 본 연구에서는 시계열의 변동 요인 중 계절 요인에 대해 고려를 하지 않았다는 한계가 있다. 이는 앞으로 발전시켜야 할 것이다.

IDX Gap을 최소화하는 파라미터를 사용자가 직접 선택해야 하는 어려움이 있다는 한계도 남아있다. 이는 앙상블 방법을 통해 결과를 종합하거나, 최적화를 통해 최적 파라미터를 제시하는 방식, 혹은 상호작용적 시각화(interactive visualization) 모듈을 구현하여 사용자가 파라미터를 자유롭게 조절할 수 있는 편리한 UI 제공 등을 통해 개선할 수 있을 것이다.

References

- [1] Yong Dae Cho and Phil Sang Lee, "A Study on Estimating the Timing of Structural Change in the Common Stock Returns," *Asian Review of Financial Research*, Vol.14, No.2, pp.131-160, 2001.
- [2] R. H. Hooker, "The Suspension of the Berlin Produce Exchange and Its Effect upon corn Prices," *Journal of the Royal Statistical Society*, Vol.64, No.4, pp.574-613, 1901.
- [3] J. Spencer, "On the Graduation of the Rates of Sickness and Mortality Presented by the Experience of the Manchester Unity of Odd fellows during the Period 1893-97," *Journal of the Institute of Actuaries*, Vol.38, No.4, pp.334-343, 1904.
- [4] V. O. Anderson and U. Nachmals, "The Elimination of Spurious Correlation due to Position in Time or Space," *Biometrika*, pp.269-279, 1914.
- [5] F. R. Macauley, "The smoothing of time series," *National Bureau of Economic Research*, pp.121-136, 1930.
- [6] M. T. Copeland, "Statistical Indices of Business Conditions," *The Quarterly Journal of Economics*, pp.522-562, 1915.
- [7] R. B. Cleveland, W. S. Cleveland, J. E. McRae and I. Terpenning, "STL: A Seasonal-Trend Decomposition Procedure Based on Loess," *Journal of Official Statistics*, Vol.6, No.1, pp.3-73, 1990.
- [8] Henry L. Gray, Chu-Ping C. Vijverberg and Wayne A. Woodward, "Nonstationary Data Analysis by Time Deformation," *Communications in Statistics-Theory and Methods*, Vol.34, No.1, pp.163-192, 2005.
- [9] G. C. Chow and A. Lin, "Best Linear Unbiased Interpolation, Distribution and Extrapolation of Time Series by Related Series," *Review of Economics and Statistics*, Vol.53, pp. 372-375, 1971.
- [10] F. T. Denton, "Adjustment of Monthly or Quarterly Series to Annual Totals: an Approach Based on Quadratic Minimization," *Journal of American Statistical Association*, Vol.66, pp.99-102, 1971.
- [11] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan and Q. Zhang, "Tiara: a Visual Exploratory Text Analytic System," *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.153-162, 2010.
- [12] N. Glance and M. H. Ruge, "Derviving Marketing Intelligence from Online Discussion," *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pp.419-428, 2005.
- [13] P. Sawon, C. Subhankar and M. Rajdeep, "StationPlot: A New Non-stationarity Quantification Tool for Detection of Epileptic Seizures," *IEEE Global Conference on Signal and Information Processing (IEEE GlobalSIP)*, 2018.
- [14] CK. Vincent, Cheung, D. Karthik, S. Giacomo, T. Andrea and B. Paolo, "Decomposing Time Series Data by a Non-negative Matrix Factorization Algorithm with Temporally Constrained Coefficients," *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2015, pp.3496-3499, 2015.
- [15] L. B. Godfrey and M. S. Gashler, "Neural Decomposition of Time-series Data for Effective Generalization," *IEEE Transactions on Neural Networks and Learning Systems*, Vol.29, No.7, pp.2973-2985, 2018.
- [16] Component of Time Series Data [Internet], <https://googl/jgbTZN>
- [17] Bernhard Pfaff, "Analysis of Integrated and Cointegrated Time Series with R," 2nd ed., Springer, pp.63, 2008.
- [18] D. A. Dickey and W. A. Fuller, "Distributions of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, Vol.74, pp.427-431, 1979.
- [19] P. C. B. Phillips and P. Perron, "Testing for a Unit Root in Time Series Regression," *Biometrika*, Vol.75, No.2, pp. 335-346, 1988.
- [20] Elliott G., T. J. Rothenberg, and J. H. Stock, "Efficient Tests for an Autoregressive unit Root," *Econometrica*, Vol.64, No.4, pp.813-836, 1996.
- [21] P. Schmidt and P. C. B. Phillips, "LM Tests for a unit Root in the Presence of Deterministic Trends," *Oxford Bulletin of Economics and Statistics*, Vol.54, No.3, pp.257-287, 1992.
- [22] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the Null Hypothesis of Stationarity Against the Alternative of a unit Root: How Sure are we that Economic Time Series have a unit Root?," *Journal of Econometrics*, Vol.54, pp.159-178, 1992.
- [23] D. R. Cox and A. Stuart, "Some Quick Sign Test for Trend in Location and Dispersion," *Biometrika*, Vol.42, pp.80-95, 1955.



유재성

<https://orcid.org/0000-0003-1682-4357>

e-mail : lv999@korea.ac.kr

2012년 충남대학교 경제학(학사),
경영학(학사), 컴퓨터공학(학사)
2015년 고려대학교 통계학과(석사)
2015년~현 재 고려대학교 컴퓨터학과
박사수료

관심분야: Stochastic Process 기반 분석(시계열, 생존분석),
그래프 이론, 통계적 검정 및 시각화



주재걸

<https://orcid.org/0000-0003-1071-4835>

e-mail : jchoo@korea.ac.kr

2001년 서울대학교 전기공학부(학사)
2009년 미국 Georgia Institute of
Technology 전기컴퓨터공학
(석사)

2013년 미국 Georgia Institute of Technology 컴퓨터공학(박사)

2015년~현 재 고려대학교 컴퓨터학과 조교수

관심분야: 컴퓨터비전, 자연어처리, 의료영상인식, 추천 시스템,
정보 시각화 및 시각화 분석 시스템, 텍스트 및 웹
데이터 마이닝