2018년 추계우수논문

A Study on the Test and Visualization of Change in Trends associated with the Occurrence of Non-stationary of Long-term Time Series Data based on Unit Root Test

ABSTRACT

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. 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 to visually check the detection of short-term structural change based on the unit root test method in a given long-term time series.

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

Unit Root Test를 기반으로 한 장기 시계열 데이터의 non-stationary 발생에 따른 추세 변화 검정 및 시각화 연구

요 약

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 적정한 대응을 할 필요가 있기 때문이다. 본 연구에서는 장기 시계열이 주어진 상황에서, 단위근 검정법을 기반으로 단기적인 구조 변화의 감지를 시각적으로 파악하는 방법을 제시하고자 한다.

키워드 : 시계열, 비정상성, 단위근 검정, 시각화

1. 서 론

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

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

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

$$\{y\}_{t=1}^{T} = \{y_1, y_2, \cdots, y_t, \cdots, y_{T-1}, y_T\},\ (t=1, \cdots, T\in\mathbb{T})$$

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.

$$E[y_t] = \mu < \infty,\ \forall t\in\mathbb{T}$$

$$E[(y_t-\mu)(y_{t-j}-\mu)] = \gamma_j,\ \forall t, j\in\mathbb{T}$$

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

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

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

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 그림 1과 같이 비정상(non-stationary)의 특징을 지니게 된다.

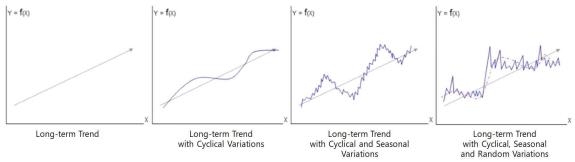

Fig 1. Changes due to variation elements of time series data

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

본 연구에서는 장기 시계열이 주어진 상황에서, 단위근 검정법을 기반으로 그림 2와 같은 단기적 구조 변화를 감지하여, 이러한 변화가 얼마나 지속할 것인지를 시각적으로 판단하는 방법을 제시하고자 한다.

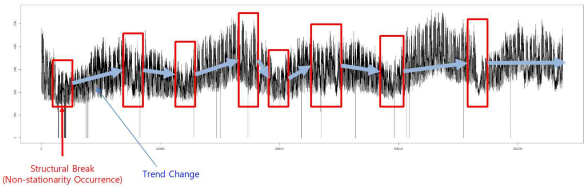

Fig 2. Short-term structural changes in long-term time series. Short-term trends before and after structural changes tend to change.

2. 단위근 프로세스

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

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

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

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

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

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

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

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

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

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

즉, 위 시계열은 평균과 분산이 무한히 커지면서, 만일 상수항 a가 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.

$$y_t = \beta\, y_{t-1} + \epsilon_t,\ -1 \le \beta \le 1$$

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

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

여기서 L은 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. 단위근 검정

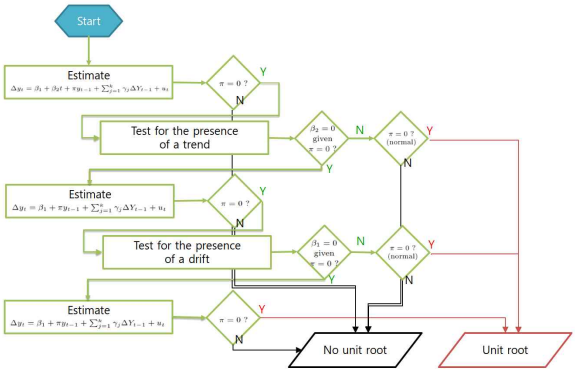

Fig 3. Procedure of Unit Root Test

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

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

여기서 $\rho = 1$이면 단위근을 가진다. 이때 y_t는 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_t가 비정상인가 여부를 다음과 같은 가설을 통해 판단할 수 있다.

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

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

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

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

$$\gamma \equiv \rho - 1$$

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

3.1 Adjusted Dickey-Fuller 검정(ADF 검정)

주어진 시계열 y_t의 비정상성 여부는, 이 y_t를 차분하여 이를 다시 y_{t-1}에 회귀하여 얻은 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. γ에 대한 t 값을 Dickey-Fuller (DF) 검정[1] 통계량이라고 하며, 이 통계량에 대한 임계값(critical value)를 Dickey and Fuller가 표로 제시

하였다. 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 상관이 있지 않다는 가정이 전제된다. 오차항에 자기 상관이 있을 경우를 고려하기 위해 다음과 같은 모형을 설정하여 이루어지는 단위 검정을 adjusted DF 검정이라고 한다.

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

$$\Delta y_t = a_0 + \pi y_{t-1} + \sum_{j=1}^{k} \gamma_j \Delta y_{t-j} + u_{2t}$$

$$\Delta y_t = a_0 + a_1 t + \pi y_{t-1} + \sum_{j=1}^{k} \gamma_j \Delta y_{t-j} + u_{3t}$$

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

3.2 Phillips-Perron 검정(PP 검정)

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

PP 검정[5]은, 시차를 갖는 차분 값의 포함 없이 자기 상관을 고려하는 방법을 제시하였다. PP 검정은 다음 두 가지 회귀분석을 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 식 i)에 대해서 검정 통계량은 다음과 같다.

$$Z(\hat{\alpha}) = T(\hat{\alpha} - 1) - \hat{\lambda} / \overline{m}_{yy}$$

$$Z(\tau_{\hat{\alpha}}) = \left(\hat{s} / \hat{\sigma}_{T\ell}\right) t_{\hat{\alpha}} - \hat{\lambda}' {}_{T\ell} / \overline{m}_{yy}^{1/2}$$

$$Z(\tau_{\hat{\mu}}) = \left(\hat{s} / \hat{\sigma}_{T\ell}\right) t_{\hat{\mu}} - (\hat{\lambda}' {}_{T\ell} m_y) / (\overline{m}_{yy}^{1/2} m_{yy}^{1/2})$$

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

또한, 식 ii)에 대해서 검정 통계량은 다음과 같다.

$$Z(\tilde{\alpha}) = T(\tilde{\alpha} - 1) - \tilde{\lambda} / M$$

$$Z(t_{\tilde{\alpha}}) = \left(\tilde{s} / \tilde{\sigma}_{T\ell}\right) t_{\tilde{\alpha}} - \tilde{\lambda}' {}_{T\ell} / M^{1/2}$$

$$Z(t_{\tilde{\mu}}) = \left(\tilde{s} / \tilde{\sigma}_{T\ell}\right) t_{\tilde{\mu}} - (\widetilde{\lambda' {}_{T\ell}} m_y) / \left[M^{1/2} (M + m_y^2)\right]^{1/2}$$

$$Z(t_{\tilde{\beta}}) = \left(\tilde{s} / \tilde{\sigma}_{T\ell}\right) t_{\tilde{\beta}} - \left[\tilde{\lambda}' \tilde{\sigma}_{T\ell} \left(\frac{1}{2} m_y - m_{ty}\right)\right] / \left[(M/12)^{1/2} \overline{m}_{yy}^{1/2}\right]$$

$$M = (1 - T^{-2}) m_{yy} - 12 m_{ty}^2 + 12(1 + T^{-1}) m_{ty} m_y - (4 + 6T^{-1} + 2T^{-2}) m_y^2$$

여기서 $m_{ty} = T^{-5/2} \sum t y_t$이다.

3.3 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[2]은 Dickey-Fuller 단위근 검정을 변형시켜 검정력을 향상하려는 방법이다. 이 방법의 기각역은 Elliott, Rothenberg and Stock이 표로 제시하였다.

3.4 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[4]는 귀무가설과 대립가설하에서 동일한 일련의 불필요한 매개 변수를 정의하는 Lagrange multiplier(LM) 검정 방법을 제안했다. 또한, 이 검정 방법에서는 선형 추세보다 높은 다항식을 고려한다.

$$y_t = \alpha + Z_t \delta + x_t,\ Z_t = (t, t^2, \cdots, t^p)$$

$$x_t = \pi x_{t-1} + \epsilon_t,\ \epsilon_t \sim N(0, \sigma^2)$$

검정 통계량은 위 회귀 식을 실행함으로써 구성된다.

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

우선 $\tilde{\psi}_x = y_1 - Z_1 \tilde{\delta}$ (여기서 $\tilde{\delta}$는 δ의 추정량)를 계산한다. 다음으로 $\tilde{S}_t = y_t - \tilde{\psi}_x - Z_t \tilde{\delta}$를 정의한다. 마지막으로 검정을 위한 회귀 식을 다음과 같이 고려한다.

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

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.

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

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

앞서 소개한 검정법들에서는 귀무가설이 단위근 과정이었지만 KPSS 검정[3]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.

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

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

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

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

그러면 검정 통계량은 다음과 같이 구할 수 있다.

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

여기서 σ_ϵ^2은 위 단계에서 구한 오차 분산의 추정치이다.

4. 단기적 구조 변화 감지를 위한 시각화

장기 시계열이 주어진 상황에서, 단기적으로 구조 변화를 감지하기 위해서는, 우선 주어진 시계열 데이터를 분할하여 검정을 진행해야 한다. 그러나 이 길이와 간격을 어떻게 정하여 분할해야 하는지에 대한 명확한 기준은 없다. 이는 도메인과 데이터 특성에 따라 달라질 것이므로, 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.

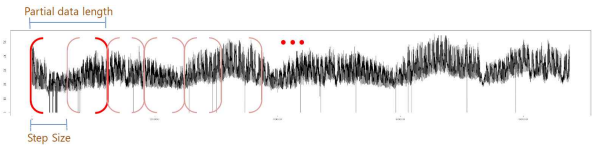

Fig 4. Segmentation using partial data length and step size of time series data

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

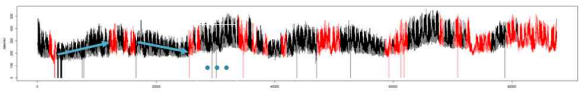

Fig 5. Based on the results of the unit root test, some partial data are emphasized

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.

5. 시각화 및 성능 평가 실험

5.1. 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.

a) Type 1: Stationary and random time series

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

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

b) Type 2: Random walk process

Type 2는 일정한 평균과 분산을 가지는 랜덤 프로세스이다.

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

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

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

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

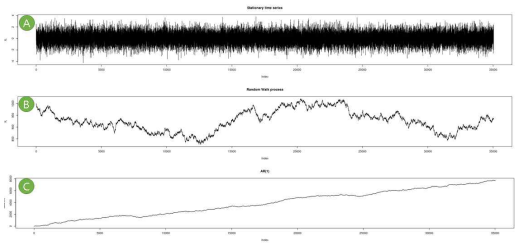

Fig 6. Synthetic Data using the base model($n = 35{,}040$)
(A) Type 1. $\mu = 0,\ \sigma^2 = 1,$
(B) Type 2. $y_0 = 1000,\ \mu = 0,\ \sigma^2 = 1$
(C) Type 3. $\alpha = 1,\ \mu \sim Unif(0,3), \sigma^2 = 5$

2) 이상치(Outlier) 모형

앞서 소개한 기본 모형은 정상(stationary) 시계열이므로 이를 그대로 분석에 활용하는 것은 본 논문에서는 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_0는 y_{t_0}로부터의 편차, $I_t(t_0)$는 지시함수로 $I_t(t_0) = \begin{cases} 1 & t = t_0 \\ 0 & otherwise \end{cases}$, 그리고 $\Phi_p(L)$, $\Theta_q(L)$은 시차 다항식(lag polynomial)이다.

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

이를 바탕으로, ARIMA(p, d, q) 모형에 기반하여 5가지 형태의 이상치(outlier)를 부여하였다.

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.

Table 1. Definition of 5 types outliers

Additive outliers(AO)
If $K_{p,d,q}(L) = 1$, $z_t = \begin{cases} y_t + \omega_0 I_t(t_0) & m = 1 \\ y_t + \sum_{i=1}^{m} \omega_i I_t(t_i) & m > 1 \end{cases}$, $\omega_i, i = 1, 2, \cdots, m$
Innovational outliers(IO)
If $K_{p,d,q}(L) = \frac{\Phi_q(L)}{\alpha_d(L)\Theta_p(L)}$, $z_t = y_t + \omega_0 K_{p,d,q}(L) I_t(t_0) = K_{p,d,q}(L)\epsilon_t + \omega_0 K_{p,d,q}(L) I_t(t_0) = K_{p,d,q}(L)[\epsilon_t + \omega_0 I_t(t_0)]$
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)
$\sigma_\omega^2 \rightarrow (1-\omega_0)\sigma_\omega^3 I_t(t_0)$

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

Fig 7., Fig 8.은 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 ζ번의 랜덤한 위치마다 생성된다.

$$IDX = \begin{Bmatrix} IDX^{(1)} = IDX_{IO}, & IDX^{(2)} = IDX_{LSO}, \\ IDX^{(3)} = IDX_{TCO_1}, & IDX^{(4)} = IDX_{TCO_2}, \\ IDX^{(5)} = IDX_{VC}, & IDX_{AO} \end{Bmatrix}$$

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

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

ξ는 IDX가 n과 같아지는 것을 막기 위한 양의 정수

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

$$IDX_{AO}^{(\ell)} \sim \lceil Unif(0, n) \rceil, \ \ell = 1, 2, \cdots, \zeta$$

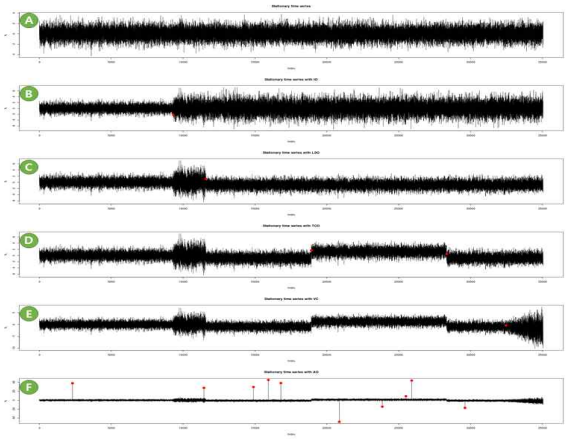

Fig 7. An example of Type 1 Synthetic data ($n = 35{,}040$, $\xi = 1000$)

(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. $\sigma_t^2 = \sigma_t^2 + \frac{3-1}{n - IDX_{VC} + 1}$, $\zeta = 10$ where $t \geq IDX_{VC}$

(F) (E) with AO. $y_t = y_t + 10 \times Unif(-5, 5)$ where $t = IDX_{AO}$

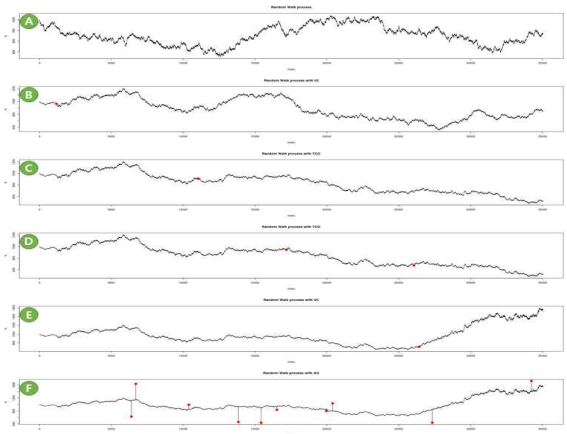

Fig 8. An example of Type 2 Synthetic data ($n = 35{,}040$, $\xi = 1000$)

(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. $\sigma_t^2 = \sigma_t^2 + \frac{3-1}{n - IDX_{VC} + 1}$, $\xi = 10$ where $t \geq IDX_{VC}$

(F) (E) with AO. $y_t = y_t + 10 \times Unif(-5, 5)$ where $t = IDX_{AO}$

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 μ가 양수일 경우 우상향, 음수일 경우 우하향하는 추세가 만들어진다. 따라서 μ를 변경하는 방법으로 추세가 변화하는 structural break 시점을 ζ번 생성할 수 있다.

$$IDX_{TC}^{(c)} \sim \begin{cases} \lceil Unif(0,c) \rceil & c=1 \\ \lceil Unif\left(IDX_{TC}^{(c-1)}+1, n-\xi\right) \rceil & c>1 \end{cases}$$
$$c = 1, 2, \cdots, \zeta$$

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

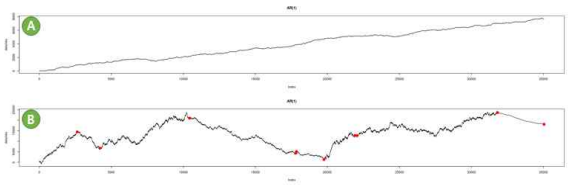

Fig 9. An example of Type 3 Synthetic data
($n = 35{,}040$, $\xi = 1000$)
(A) $\alpha = 1$, $y_0 = 10$, $\mu \sim Unif(0,3)$, $\sigma^2 = 5$
(B) $\mu_\tau \sim Unif(-5,5)$, $\sigma_\tau^2 \sim Unif(0,100)$, $\zeta = 9$

$$\text{where } \begin{cases} IDX_{TC}^{(c)} \le \tau < IDX_{TC}^{(c+1)} & c < \xi \\ IDX_{TC}^{(c)} \le \tau \le n & c = \xi \end{cases}$$

4) 평가(evaluation)

사용자는 partial data length, step size, 단위근 검정에 대한 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 파라메터를 결정한 것이라고 할 수 있다.

$$\text{IDX Gap: } \min \sum_{\max(IDX)} \left[\left| IDX - \widehat{IDX} \right| \right]$$

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.

5.2. 임의 데이터를 이용한 실험

1) 시각화

임의 데이터 Fig 7. (F), Fig 8. (F), Fig 9. (B)에 대하여 파라메터를 Table 2와 같이 설정한 후 시각화를 진행하였다.

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. 추세 검정은 Cox-Stuart trend 검정을 사용하였다.

Table 2. Parameters for analysis of Fig 7. (F), Fig 8. (F), Fig 9. (B) (CV is critical value)

Partial data length				Step size			
96 × 14 = 1,344				96 × 3.5 = 336			
Parameters of Unit Root Test for Fig 7. (F)							
		CV	**Lag**			**CV**	**Lag**
ADF	Trend	-3.13	96	**ERS**	Trend	-2.57	48
	Drift	-2.57	96	**DG**	Const.	-1.62	96
PP	Trend	-3.14	4	**ERS**	Trend	6.89	12
	Const.	-2.57	4	**P**	Const.	4.48	12
SP	Tau	.0001	12	**KPSS**	Tau	0.18	12
	Rho	.0001	12				
Parameters of Unit Root Test for Fig 8. (F)							
		CV	**Lag**			**CV**	**Lag**
ADF	Trend	-3.13	96	**ERS**	Trend	-3.48	12
	Drift	-2.57	96	**DG**	Const.	-2.57	12
PP	Trend	-3.99	12	**ERS**	Trend	3.96	12
	Const.	-3.45	12	**P**	Const.	1.99	12
SP	Tau	.0001	12	**KPSS**	Tau	0.12	12
	Rho	.0001	12				
Parameters of Unit Root Test for Fig 9. (B)							
		CV	**Lag**			**CV**	**Lag**
ADF	Trend	-3.42	96	**ERS**	Trend	-2.89	12
	Drift	-2.87	96	**DG**	Const.	-1.94	12
PP	Trend	-3.42	12	**ERS**	Trend	5.62	12
	Const.	-2.87	12	**P**	Const.	3.26	12
SP	Tau	.0001	12	**KPSS**	Tau	0.15	12
	Rho	.0001	12				
Significance Probability of Cox-Stuart Trend Test							
.001							

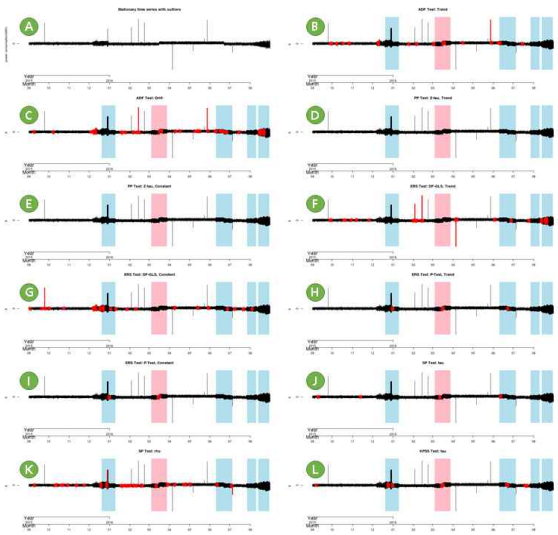

Fig 10. Visualization of analysis of Fig 7. (F)
(A) Line Graph (B) (A) with ADF Trend
(C) (A) with ADF Drift (D) (A) with PP Trend
(E) (A) with PP Const. (F) (A) with ERS DF-GLS Trend
(G) (A) with ERS DF-GLS Const. (H) (A) with ERS P Trend
(I) (A) with ERS P Const. (J) (A) with SP Tau
(K) (A) with SP Rho (L) (A) with KPSS Tau

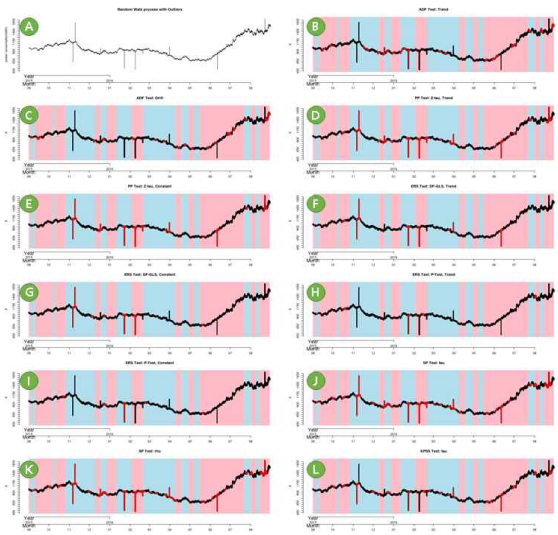

Fig 11. Visualization of analysis of Fig 8. (F)
(A) Line Graph (B) (A) with ADF Trend
(C) (A) with ADF Drift (D) (A) with PP Trend
(E) (A) with PP Const. (F) (A) with ERS DF-GLS Trend
(G) (A) with ERS DF-GLS Const. (H) (A) with ERS P Trend
(I) (A) with ERS P Const. (J) (A) with SP Tau
(K) (A) with SP Rho (L) (A) with KPSS Tau

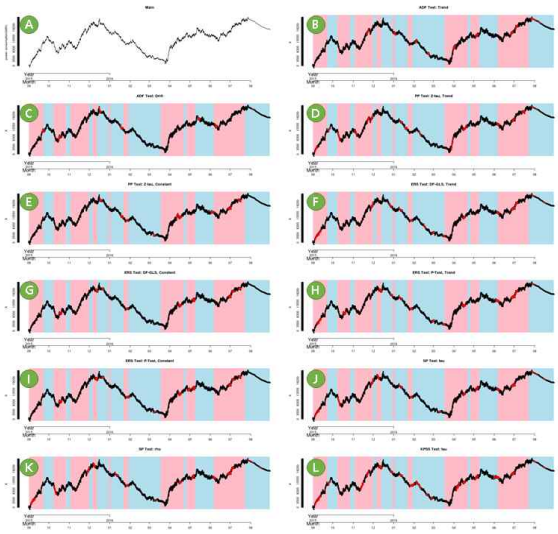

Fig 12. Visualization of analysis of Fig 9. (B)
(A) Line Graph (B) (A) with ADF Trend
(C) (A) with ADF Drift (D) (A) with PP Trend
(E) (A) with PP Const. (F) (A) with ERS DF-GLS Trend
(G) (A) with ERS DF-GLS Const. (H) (A) with ERS P Trend
(I) (A) with ERS P Const. (J) (A) with SP Tau
(K) (A) with SP Rho (L) (A) with KPSS Tau

각 결과에 대한 IDX Gap을 측정한 결과는 Table 3과 같다. Fig 7. (F)의 경우 추세가 없으므로 이상치 발생 이후의 구조 변화를 탐지하는 것이 더 수월한 것으로 여겨진다.

Table 3. IDX Cap of results by analysis for Fig 7. (F), Fig 8. (F), Fig 9. (B)

분석		IDX Gap				
		IO	LSO	TCO(1)	TCO(2)	VC
Fig 7. (F)						
ADF	Trend	1	20	0	10	19
	Drift	0	19	0	0	9
PP	Trend	0	0	0	0	0
	Const.	0	0	0	0	0
ERS DG	Trend	10	31	18	4	7
	Const.	0	2	22	4	0
ERS P	Trend	0	0	0	0	0
	Const.	0	0	0	0	0
SP	Tau	0	0	0	0	0
	Rho	0	0	0	0	0
KPSS	Tau	0	0	0	0	0
Fig 8. (F)						
ADF	Trend	2	1	11	0	0
	Drift	0	1	11	0	3
PP	Trend	32	2	4	3	0
	Const.	52	4	4	6	7
ERS DG	Trend	33	26	10	11	7
	Const.	53	0	17	7	11
ERS P	Trend	4	29	11	63	59
	Const.	39	0	1	7	11
SP	Tau	3	0	3	4	0
	Rho	3	0	3	4	0
KPSS	Tau	18	0	2	1	5
Fig 9. (B)						
c of $IDX_{TC}^{(c)}$		1	2	3	4	5	6	7	8	9
ADF	Trend	54	38	14	10	10	7	13	15	0
	Drift	4	2	0	45	45	50	27	26	6
PP	Trend	1	3	19	23	23	4	0	0	0
	Const.	80	64	0	3	3	15	0	0	11
ERS DG	Trend	15	32	7	3	3	7	0	0	0
	Const.	19	3	2	36	36	21	0	0	11
ERS P	Trend	15	3	5	3	3	7	0	0	0
	Const.	19	3	2	4	4	14	0	0	11
SP	Tau	11	3	9	5	5	13	0	0	0
	Rho	11	3	9	5	5	13	0	0	0
KPSS	Tau	1	18	5	5	5	11	0	0	2

2) 파라메터의 평가

같은 조건으로 데이터를 생성하되, 난수 생성 시드를 달리하여 동일한 분석을 100회 반복하여 각각의 IDX Gap을 구한 뒤 그의 합을 Table 4와 같이 정리하였다. 역시 Fig 7. (F)의 경우 추세가 없기 때문에 이상치 발생 이후의 구조 변화를 탐지하는 것이 수월한 것으로 해석된다.

각 검정 방법마다 전제한 가정이 다른 만큼 비정상성 여부를 판단하는 결과도 달라지므로, 모든 검정 결과에 따라 공통으로 강조되는 시점을 찾아야 하는 어려움이 있다. 다만 시계열 데이터의 도메인이나 특성에 따라 적용해야 하는 단위근 검정 방법이 달라질 것이므로, 한두 가지 검정 방법만 집중하도록 하는 것은 무리가 있을 수 있으며, 그럴 때는 어떠한 변동 요소에 주목할 것인지를 선택할 수도 있다.

구조 변화 시점을 명확하게 알 수 없는 실제 데이터를 분석하기에 앞서서, 데이터의 형태와 유사한 임의 데이터를 먼저 생성, 위와 같은 반복 분석을 통해 적정한 파라메터를 선정한 후, 실제 데이터의 분석에 활용하는 것도 방법이 될 수 있다.

Table 4. Sum of IDX Cap of results by 100 times analysis for Fig 7. (F), Fig 8. (F), Fig 9. (B)

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

5.3. 고려대학교 녹지캠퍼스 전력사용량 데이터

고려대학교 녹지캠퍼스의 2015년 9월 1일부터 2018년 2월 28일까지 15분 단위로 측정한 전력 사용량(kWh) 데이터를 사용하였다.

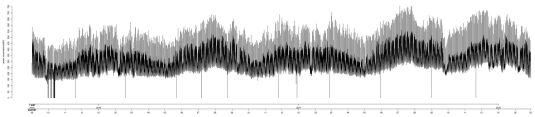

Fig 13. Power consumption data of Korea university(kWh)

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 대는 대체로 점진적으로 분산이 작아지는 형태(VC)로 보인다.

분석을 위해 파라메터를 Table 5.와 같이 설정하였으며, 시각화 결과는 Fig 14.와 같이 나타났다.

Table 5. Parameters for power consumption of Korea University (CV is critical value)

Partial data length				Step size			
96 × 14 = 1,344				96 × 3.5 = 336			
Parameters of Unit Root Test							
		CV	Lag			CV	Lag
ADF	Trend	-3.98	96	ERS DG	Trend	-2.57	96
	Drift	-3.44	96		Const.	-1.94	96
PP	Trend	-3.99	96	ERS P	Trend	3.96	96
	Const.	-3.45	96		Const.	1.99	96
SP	Tau	.0001	12	KPSS	Tau	0.12	96
	Rho	.0001	12				
Significance Probability of Cox-Stuart Trend Test							
.001							

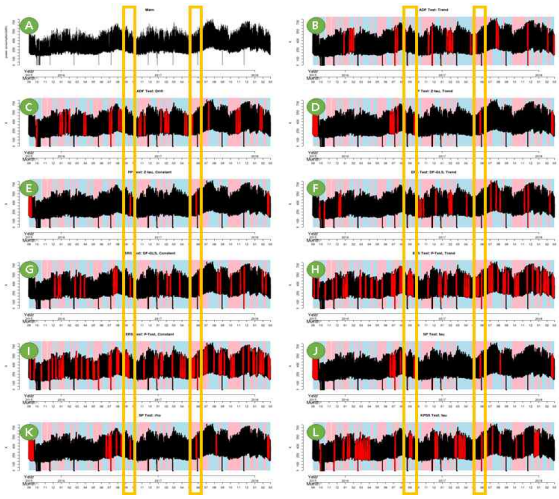

Fig 14. Visualization of analysis of power consumption(kWh) data of Korea university
(A) Line Graph (B) (A) with ADF Trend
(C) (A) with ADF Drift (D) (A) with PP Trend
(E) (A) with PP Const. (F) (A) with ERS DF-GLS Trend
(G) (A) with ERS DF-GLS Const. (H) (A) with ERS P Trend
(I) (A) with ERS P Const. (J) (A) with SP Tau
(K) (A) with SP Rho (L) (A) with KPSS Tau

Cox-Stuart 추세 검정 결과는 추세에 맞는 배경색이 대체로 잘 나타난 것으로 보이며, 2016년 9~10월과 2017년 5~6월 사이에서는 모든 단위근 검정 결과 비정상성(non-stationary)이 탐지되는 것으로 나타났다. 이는 계절의 변화에 따른 전력사용량 변화에 따른 것으로 보인다. ADF 검정, PP 검정, SP 검정, KPSS 검정에 따라 비정상성이 탐지된 경우, 그 직후에 추세가 변화된다는 특징이 있었다.

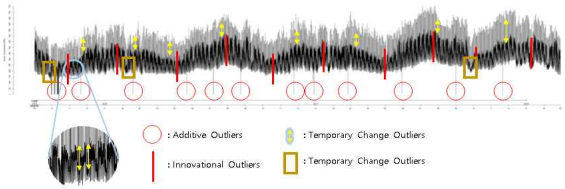

Fig 15. Outliers and structure breaks intuitively confirmed from power consumption(kWh) data of Korea university

그러나 ERS 검정의 경우 너무 빈번하게 비정상성으로 판단하는 모습이 보였다. 이는 데이터를 자세히 살펴보면 Fig 15.와 같이 여러 가지 이상치가 혼재한 것을 볼 수 있으며, 이를 전부 탐지한 결과로 보인다.

5. 결 론

본 연구에서는 장기 시계열 데이터를 분리하여 각 부분 데이터마다 단위근 검정을 수행, 그 결과를 시각적으로 표현하여 단기 구조 변화를 파악하는 방법을 제시하였다. 단위근 검정 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, 대응 할 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 파라메터를 사용자가 직접 선택해야 하는 어려움이 있다는 한계도 남아있다. 이는 앙상블 방법을 통해 결과를 종합하거나, 최적화를 통해 최적 파라메터를 제시하는 방식, 혹은 상호작용적 시각화(interactive visualization) 모듈을 구현하여 사용자가 파라메터를 자유롭게 조절할 수 있는 방식을 통해 개선할 수 있을 것이다.

참 고 문 헌

[1] Dickey, D. A. and Fuller, W. A., "Distributions of the estimators for autoregressive time series with a unit root", Journal of the American Statistical Association 74, 427 - 431. 1979.

[2] Elliott G., T.J. Rothenberg and J. H. Stock., "Efficient tests for an autoregressive unit root", Econometrica, 64-4, 813 - 836. 1996.

[3] Kwiatkowsk, D., Phillips P. C. B., Schmidt P. and Shin Y., "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 54, 159 - 178. 1992.

[4] Schmidt P. and Phillips P. C. B., "LM tests for a unit root in the presence of deterministic trends", Oxford Bulletin of Economics and Statistics 54-3, 257 - 287. 1992.

[5] Phillips P. C. B. and Perron P., "Testing for a unit root in time series regression", Biometrika 75-2, 335 - 346. 1988.