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

유재성*, 주재걸*
*고려대학교 컴퓨터학과
e-mail:lv999@korea.ac.kr

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

Jaeseong Yoo*, Jaegul Choo*
*Dept of Computer Science and Engineering, Korea University

요 약

비정상(non-stationary) 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 수준의 대응을 할 필요가 있기 때문이다. 본 연구에서는 장기 시계열이 주어진 상황에서, 단위근 검정법을 기반으로 단기적으로 구조변화를 감지하여, 이러한 변화가 얼마나 지속될 것인지를 시각적으로 판단할 수 있는 방법을 제시하고자 한다.

1. 서론

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

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

여기서 $t \in \mathbb{T}$, $y(\cdot,t)$는 확률공간 Ω 상의 확률변수이고, 확률과정의 실현은 시간 $t \in \mathbb{T}$에 관한 각 $s \in \Omega$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.

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

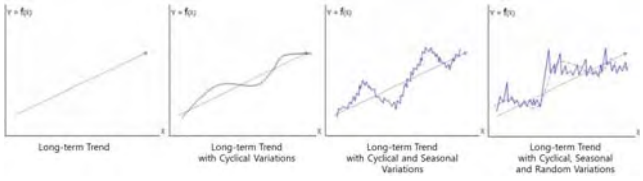
$$E[(y_t - \mu)(y_{t-j} - \mu)] = \gamma_j \ , \ \forall t, j \in \mathbb{T}$$

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.



(그림 1) 시계열 데이터의 변동 요소에 따른 변화

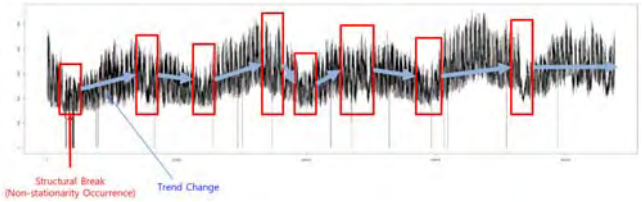
시계열 자료의 변수는 장기간에 걸쳐 추세변동(trend variation), 순환변동(cyclical variation), 계절변동(seasonal variation), 그리고 불규칙변동(irregular variation)이 동시에 일어나기 때문에, 시계열이 장기적일수록 구조 추론이 점점 어려워지는 현상이 발생한다. 단기 시계열의 경우 안정적인 시계열로 간주하고 분석해도 큰 무리가 없는 반면, 장기 시계열의 경우 이러한 변동 요인들로 인해 그림 1과 같이 비정상(non-stationary)의 특징을 지니게 된다.

이러한 비정상 장기 시계열 안에서도, 단기적으로 추세의 변화가 일시적인 것인지, 아니면 구조적으로 변한 것인

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

지를 적시에 판단하는 것은 중요하다. 이는 시계열 추세의 변화를 상시 감지하여, 변화에 맞는 적정한 수준의 대응을 할 필요가 있기 때문이다.

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



(그림 2) 장기 시계열 안에서 발생하는 단기 구조변화. 구조변화 전과 후의 단기 추세가 달라지는 경향이 있다.

2. 단위근 프로세스

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

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

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

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

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.

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

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

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

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.

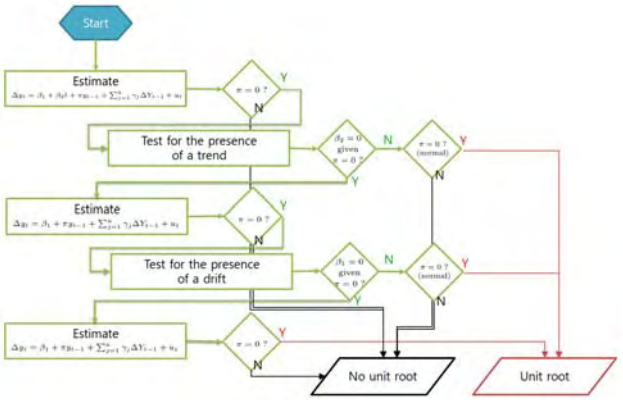
$$(1 - \beta L) y_t = a + \epsilon_t$$

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



(그림 3) 단위근 검정 절차

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

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

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.

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

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

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

이 AR(1)에서 y_t를 차분한 $\Delta y_t = y_t - y_{t-1}$ 은 ϵ_t, 즉 백색 잡음이므로, 안정적(stationary)이게 된다.

3.1. Adjusted Dickey-Fuller 검정(ADF검정)

주어진 시계열 y_t의 비정상성 여부는, 이 y_t를 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. γ에 대한 t 값을 Dickey-Fuller (DF) 검정[1] 통계량이라고 하며, 이 통계량에 대한 임계값(critical value)를 Dickey and Fuller가 표로 제시하였다. 만일 귀무가설이 기각될 경우, 정상 시계열이라고 볼 수 있으며, 통상적인 t 검정을 적용

할 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 자기상관이 있지 않다는 가정이 전제된다. 오차항에 자기상관이 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 자기상관의 문제를 고려하고 있다.

PP 검정[5]은, 시차를 갖는 차분 값의 포함 없이 자기상관을 고려하는 방법을 제시하였다. PP검정은 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 $m_y = T^{-3/2}\sum y_t$, $\overline{m}_{yy} = T^{-2}\sum (y_t - \overline{y})^2$, $m_{yy} = T^{-2}\sum y_t^2$, , 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) 과정인 경우 검정력이 낮아진다는 것이다. ERS 검정[2]은 Dickey-Fuller 단위근 검정을 변형시켜 검정력을 향상시키기 위한 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.

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

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

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.

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

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.

$$y_t = \xi t + r_t + \epsilon_t \ , \ \epsilon_t \sim i.i.d.N(0, \sigma_u^2)$$

$$r_t = r_{t-1} + u_t$$

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 식으로

부터 잔차의 부분 합을 다음과 같이 계산한다.

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

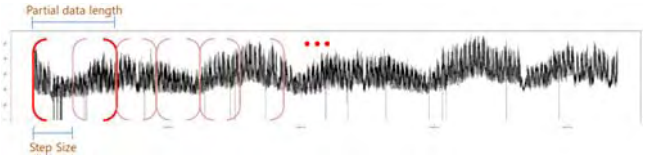
그러면 검정 통계량은 다음과 같이 구할 수 있다.

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

여기서 σ_ϵ^2은 위 단계에서 구한 오차 분산의 추정치이다.

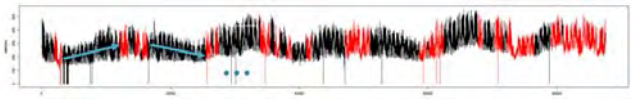
4. 단기적 구조변화 감지를 위한 시각화

장기 시계열이 주어진 상황에서, 단기적으로 구조변화를 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 length는 step size보다 크거나 같게 선택하도록 해야 한다.



(그림 4) 시계열의 partial data length와 step size를 이용한 분할

부분 데이터를 추출하면 데이터마다 단위근 검정을 한다. 만일 특정 부분 데이터의 단위근 검정 결과 비정상 시계열일 가능성이 클 경우, 해당 부분을 강조한다.



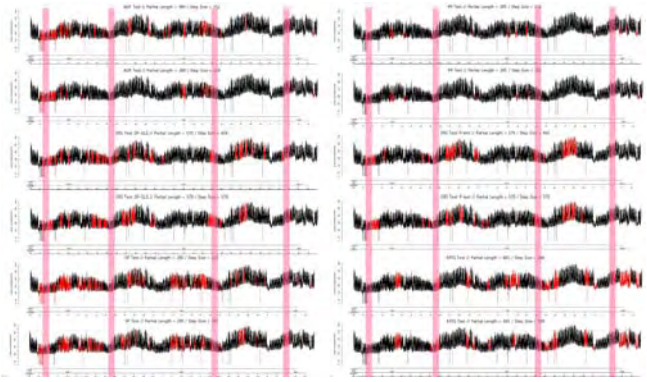
(그림 5) 단위근 검정 결과를 바탕으로 일부 부분 데이터를 강조한 모습

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 확인함에 있어서는 Cox-Stuart 검정, Mann-Kendall 검정과 같은 추세 검정 방법을 사용할 수 있다.

5. 실험

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

검정 방법마다 고려하는 가정이 다르므로 강조된 부분은 제각각이지만, 계절적 특성이 뚜렷해지는 11~12월, 4~5월 사이에 있어서는 검정 방법과 관계없이 일관되게 강조되는 모습을 볼 수 있다.



(그림 6) 고려대학교 전력데이터의 단기 구조변화 감지 결과

2016년 5월 말을 기준으로 직전 3개월(2016년 3월부터 5월까지)과 직후 3개월(2016년 6월부터 8월까지)을 Cox-Stuart 검정을 진행한 결과 전자는 하향 추세인 것으로(p-value = 2.721e-110), 후자는 상향 추세인 것으로(p-value = 8.8397e-206) 나타났다.

6. 결론

본 연구에서는 단위근 검정 결과를 시각적으로 표현하여 단기적 구조변화를 파악하는 방법을 제시하였다. 모든 검정 결과가 공통으로 비정상(non-stationary)적으로 나타난 시점의 전후 기간에 대한 추세 검정을 한 결과, 비정상성이 나타난 시점을 기준으로 추세 변화가 나타났으며, 이에 따라 단기적인 비정상성을 바탕으로 변화에 대한 대응을 할 수 있다.

다만 각 검정 방법마다 전제한 가정이 다른 만큼 비정상성 여부를 판단하는 결과도 달라지는 만큼, 모든 검정 결과에 따라 공통으로 강조되는 시점을 찾아야 하는 어려움이 있다. 그러나 시계열 데이터의 도메인이나 특성에 따라 적용해야 하는 단위근 검정 방법이 달라질 것이므로, 한두 가지 검정 방법만 집중하도록 하는 것은 무리가 있다. 이는 앙상블 방법을 통해 결과를 종합하거나, 공통으로 강조된 부분을 시각적으로 표시해주는 방식을 통해 개선될 수 있을 것이다.

참고문헌

[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.