The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.5.1025
28(5), 1025–1034

Effects of Parameter Estimation in Phase I on Phase II Control Limits for Monitoring Autocorrelated Data

Sungim Lee[a,1]

[a]Department of Applied Statistics, Dankook University

(Received August 28, 2015; Revised September 9, 2015; Accepted September 9, 2015)

Abstract

Traditional Shewhart control charts assume that the observations are independent over time. Current progress in measurement and data collection technology lead to the presence of autocorrelated process data that may affect poor performance in statistical process control. One of the most popular charts for autocorrelated data is to model a correlative structure with an appropriate time series model and apply control chart to the sequence of residuals. Model parameters are estimated by an in-control Phase I reference sample since they are usually unknown in practice. This paper deals with the effects of parameter estimation on Phase II control limits to monitor autocorrelated data.

Keywords: autocorrelation, time series data, residual-based control chart, average run length

1. 서론

일반적으로 품질관리를 위한 Shewhart 관리도는 시간에 따라 관측한 품질특성치가 서로 독립임을 가정하게 된다. 그러나 최근들어 데이터 측정과 저장기술이 발전하면서 자기상관이 존재하는 공정 데이터가 많이 생산되고 있는 추세이다. 예를 들어, 센서에 의해 매우 짧은 간격으로 데이터 수집이 이루어진다면 데이터의 자기상관성 존재는 매우 자연스러운 현상이다. 자기상관이 있는 공정 데이터로부터 관리도를 작성하는 방법은 자기상관성을 어떻게 처리하는지에 따라 크게 두 가지로 나뉜다. 첫 번째는 자기상관성을 약화시키거나 제거할 수 있도록 데이터의 수집 간격을 넓히는 방법이고, 두 번째는 자기상관 구조에 대한 적절한 시계열 모형을 가정한 후 잔차를 구하여 기존의 Shewhart 관리도를 적용하는 방법이다 (Alwan과 Roberts, 1988; Montgomery 등, 1990; Box 등, 1994; Zhang, 1998). Montgomery (2012)에서 언급했듯이 데이터 수집의 간격을 넓혀 자기상관성을 제거하는 방법은 이용 가능한 수많은 데이터 정보를 무시하는 것으로 매우 비효율적인 방법이기 때문에, 좀 더 일반적으로 활용되는 것은 두 번째 방법이다. 그런데, 두 번째 방법을 적용하기 위해서는 데이터에 적절한 시계열 모형을 선택할 뿐 아니라 실제 문제에서 공정 데이터의 참 상관구조는 알려져 있지 않으므로, 관리도 작성을 위해 구한 잔차에는 모형에 대한 불확실성과 모수 추정에 따른 불확실성이 함께 존재한다. 본 논문에서는 자기상관을 설명하는 시계열 모형이 적절히 선택되었다는 가정 하에 잔차기반 Shewhart 관리도 작성에 있어 모수 추정에

This Research was supported by the Dankook University Research Grants 2014.
[1]Department of Applied Statistics, Dankook University, 126 Jukjeon-Dong, Suji-Gu, Gyeonggi-do 448-701, Korea. E-mail: silee@dankook.ac.kr

따른 불확실성에 대해 알아보고자 한다.

Quesenberry (1993), Chen (1997), 그리고 Nedumaran과 Pigniatiello (1999) 등은 Shewhart $\bar{X}$ 또는 X 관리도를 작성할 때, 관리상태하의 모수값 추정이 관리도의 성능을 표현하는 평균 런길이(Average run length; ARL)에 어떠한 변화가 있는지 연구하였는데, 관리상태하에서 얻어진 과거의 표본(즉, 일단계 표본)크기가 작을 때는 기대했던 평균 런길이가 커지고 분산도 매우 커짐을 보여주었다. 또한 모의실험을 바탕으로 일단계 표본의 크기가 적어도 300개 이상이 필요하다는 시뮬레이션 연구결과를 보여주었다. Woodall과 Montgomery (1999) 또한 관리한계선을 추정하기 위한 표본의 개수가 충분히 커야한다고 주장하였다. 자기상관이 존재하는 공정 데이터에 대하여 잔차기반한 Shewhart 관리도를 적용했을 때의 성능을 알아보기 위해 Harris와 Ross (1991), Wardell 등 (1994), 그리고 Lin과 Adams (1996) 등은 시계열 모형의 모수값이 알려져 있다고 가정하였다. 만약 시계열 모형의 참 모수값을 설정할 수 있다면, 잔차는 백색잡음과 유사한 성질을 갖게 되어 일반적인 Shewhart 관리도의 성능과 같을 것이다. 그러나 앞서 언급한대로 이러한 가정은 실제 문제에 적용하기 어려운 것으로 대부분 관리상태하의 일단계 표본으로부터 시계열 모형을 추정하게 되는데, 본 논문에서는 Shewhart 관리도 연구에서와 마찬가지로 시계열 모형에서 모수의 참값이 아닌 추정값을 사용함으로써, 이단계 공정모니터링의 관리 성능에 어떠한 영향이 미치는지 살펴보고자 한다. 실제 문제에서 공정 모니터링을 위해 관리도가 적용될 때 관리상태하에서의 런길이가 짧은 경우에는 가짜 알람률(false alarm rate)이 큰 것으로 공정의 흐름에 방해가 될 수 있고, 이상상태하에서의 런길이가 길어진다면 이상변화에 대한 탐지가 그만큼 늦어진다는 것으로 관리 성능은 매우 중요한 문제이다.

본 논문의 구성은 다음과 같다. 2절에서는 일반적으로 가정되는 시계열 모형과 이들 모형으로부터 잔차를 구하고, 3절에서는 잔차기반한 관리도 작성과 관리한계선을 구하는 방법을 소개하도록 한다. 4절에서는 시뮬레이션을 통해 잔차기반 관리도의 관리 성능에 대해 알아보고 5절에서는 결과요약과 함께 앞으로의 연구방향에 대해 요약하기로 한다.

2. 시계열 모형

자기상관이 있는 데이터 X_t에 대하여 잔차기반의 관리도를 작성하기 위해서는, 먼저 이들 데이터를 적절히 설명할 수 있는 시계열 모형을 적합해야한다. 잘 적합된 시계열 모형을 통해 데이터의 자기상관성을 제거한다면, 잔차는 관리상태하에서 평균이 0이고 분산이 σ^2인 $i.i.d$ 확률변수로 기대할 수 있다. 정상 시계열 모형 중 가장 널리 사용되는 모형은 다음의 자기회귀이동평균(autoregressive moving average; **ARMA**(p,q)) 모형이다.

$$X_t = \frac{(1-\Theta_q(B))}{(1-\Phi_p(B))}\epsilon_t \tag{2.1}$$

로 나타낼 수 있으며 $\Phi_p(B) = \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p$로 차수가 p인 자기회귀 연산자, $B^k X_t = X_{t-k}$는 후진 연산자(Backshift operator), $\Theta_q(B) = \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$는 차수가 q인 이동평균 연산자, 그리고 백색잡음 ϵ_t는 서로 독립이고 동일한 분포 $N(0,\sigma^2)$을 따른다고 가정한다. 관리상태하에서 수집된 과거 데이터(즉, 일단계 표본)로부터 자기상관을 설명하는 적절한 모형이 선택되면, 모수 추정값 $\hat{\phi}_1, \ldots, \hat{\phi}_p$, $\hat{\theta}_1, \ldots, \hat{\theta}_q$로부터

$$\hat{\Phi}_p(B) = \hat{\phi}_1 B + \hat{\phi}_2 B^2 + \cdots + \hat{\phi}_p B^p,$$
$$\hat{\Theta}_p(B) = \hat{\theta}_1 B + \hat{\theta}_2 B^2 + \cdots + \hat{\theta}_q B^q$$

를 추정하고, 잔차는 다음과 같이 구한다.

$$e_t = \frac{\left(1 - \hat{\Phi}_p(B)\right)}{\left(1 - \hat{\Theta}_q(B)\right)} X_t. \tag{2.2}$$

모형이 잘 맞는다면 이들 잔차는 오차항과 비슷한 성질을 보일 것으로 생각할 수 있다. 이와 같이 잔차에 근거한 관리도를 작성할 때 관리도의 성능은 잔차의 분산추정에 의존하게 되므로 다음 절에서는 **ARMA**(p, q) 모형의 특별한 경우인 **AR**(1) 모형과 **MA**(1) 모형의 잔차를 구하고, 이에 대한 기댓값과 분산을 구하기로 한다.

2.1. **AR**(1) (또는 **ARMA**$(1, 0)$) 모형

먼저 자기상관이 있는 시계열 X_t가 **AR**(1) 모형을 따른다고 가정할 때, X_t는 다음과 같이 정의될 수 있다.

$$X_t = \phi_1 X_{t-1} + \epsilon_t, \quad t = 1, 2, \ldots, n. \tag{2.3}$$

이 때, $|\phi_1| < 1$을 가정하고 예측값 $\hat{X}_t$는

$$\hat{X}_t = \hat{\phi}_1 X_{t-1} \tag{2.4}$$

로 t 시점에서의 잔차는 다음과 같이 쓸 수 있다.

$$\begin{aligned} e_t &= X_t - \hat{X}_t \\ &= \phi_1 X_{t-1} + \epsilon_t - \hat{\phi}_1 X_{t-1} \\ &= (\phi_1 - \hat{\phi}_1) X_{t-1} + \epsilon_t \\ &= (\phi_1 - \hat{\phi}_1) (\phi_1 X_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= (\phi_1 - \hat{\phi}_1) \left(\phi_1 \cdot \frac{e_{t-1} - \epsilon_{t-1}}{(\phi_1 - \hat{\phi}_1)} + \epsilon_{t-1} \right) + \epsilon_t \\ &= \phi_1 e_{t-1} - \hat{\phi}_1 \epsilon_{t-1} + \epsilon_t. \end{aligned} \tag{2.5}$$

따라서, 이단계 표본에서 구한 잔차 e_t의 기댓값과 분산은 다음과 같다:

$$E(e_t) = 0, \tag{2.6}$$

$$\mathrm{Var}(e_t) = \left(\frac{1 - 2\phi_1\hat{\phi}_1 + \hat{\phi}_1^2}{(1 - \phi_1^2)} \right) \sigma^2. \tag{2.7}$$

따라서, $\hat{\phi}_1$이 ϕ_1을 잘 추정한다면 잔차 또한 서로 독립이고 분산이 σ^2으로 동일한 정규분포를 따른다는 것을 알 수 있다.

2.2. **MA**(1) (또는 **ARMA**$(0, 1)$) 모형

이 절에서는 시계열 X_t가 **MA**(1) 모형을 따를 때의 잔차를 구해보기로 한다. 시계열 X_t는 다음과 같이 정의된다.

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1}, \quad t = 1, 2, \ldots, n. \tag{2.8}$$

예측값 $\hat{X}_t$는

$$\hat{X}_t = \hat{\theta}_1 e_{t-1} \tag{2.9}$$

이 되고, t 시점에서의 잔차는 다음과 같다.

$$\begin{aligned} e_t &= X_t - \hat{X}_t \\ &= X_t - \hat{\theta}_1(X_{t-1} - \hat{X}_{t-1}) \\ &= X_t - \hat{\theta}_1 X_{t-1} + \hat{\theta}_1^2 X_{t-2} - \cdots\cdots + (-1)^{t-1}\hat{\theta}_1^{t-1} X_1 \\ &= (\epsilon_t + \theta_1\epsilon_{t-1}) - \hat{\theta}_1(\epsilon_{t-1} + \theta_1\epsilon_{t-2}) + \cdots\cdots + (-1)^{t-1}\hat{\theta}_1^{t-1}\epsilon_1 \\ &= \epsilon_t - (\hat{\theta}_1 - \theta_1)\epsilon_{t-1} + \hat{\theta}_1(\hat{\theta}_1 - \theta_1)\epsilon_{t-2} - \cdots\cdots + (-1)^{t-1}\hat{\theta}_1^{t-2}(\hat{\theta}_1 - \theta_1)\epsilon_1. \end{aligned} \tag{2.10}$$

따라서, 이단계 표본에서 구한 잔차 e_t의 기댓값과 분산은 다음과 같다.

$$E(e_t) = 0, \tag{2.11}$$

$$\mathrm{Var}(e_t) = \sigma^2 + (\theta_1 - \hat{\theta}_1)\sigma^2\left(\frac{1 - \hat{\theta}_1^{2(t-1)}}{1 - \hat{\theta}_1^2}\right), \tag{2.12}$$

여기서 주목할 것은 **AR**(1) 모형에서와 달리 **MA**(1) 모형의 경우에는

$$\mathrm{Var}(e_t) \geq \sigma^2$$

으로, 소표본에서는 잔차의 분산이 오차의 분산보다 커질 수 있으며, 특히 θ_1가 1에 가까울수록 잔차의 분산이 더 커질 수 있음을 알 수 있다.

3. 잔차기반 관리도와 관리한계선

앞 절에서 **AR**(1) 모형과 **MA**(1) 모형에 대한 잔차를 구하고 각각의 기댓값과 분산을 구하였다. 이 절에서는 이러한 시계열 데이터의 평균 변화를 모니터링하기 위해 잔차에 기반한 X 관리도를 적용하는 절차에 관하여 설명하기로 한다. 먼저 관리한계선을 설정해야 하는데, 식 (2.7)와 식 (2.12)에서의 σ^2이 알려져 있지 않으므로, 관리상태하에 수집된 n개의 일단계 표본으로 시계열 모형을 적합한 후 구한 잔차 $(e_1, e_2, \ldots, e_n,\ i = 1, 2, \ldots, n)$들로부터 관리하한(LCL)과 관리상한(UCL)은 다음과 같이 설정한다.

$$\begin{aligned} \mathrm{LCL} &= \bar{e} - 3\hat{\sigma}_e, \\ \mathrm{UCL} &= \bar{e} + 3\hat{\sigma}_e, \end{aligned} \tag{3.1}$$

여기서 $\bar{e} = \sum_{i=1}^{n} e_i/n$, $\hat{\sigma}_e = s/c_4$, s는 일단계 표본에서 구한 잔차들의 표준편차이고, $c_4 \approx 4(n-1)/(4n-3)$이다. 일반적으로 X 관리도를 사용할 때, σ를 추정하는데는 이동평균(moving average)을 이용한 불편추정량이 많이 쓰이지만 서론에서 언급했듯이 기존 연구에서 일단계 표본 크기가 커야한다고 했는데, 이 경우 Cryer와 Ryan (1990)에 의해 표본 표준편차(s)를 사용하는 것이 좀 더 효율적이라는 사실이 알려져 있어 이를 사용하였다. 이때, 식 (3.1)의 잔차 $e_t\,(t = 1, \ldots, n)$는 모형의 참 모수값으로부터 계산되는 것이 아니라 일단계 표본에서 구한 모수의 추정값으로부터 계산되므로 이들 잔차는 $\hat{\phi}_1$ 또는 $\hat{\theta}_1$의 함수로 나타난다. 또한 관리한계선은 이들 잔차들의 표준편차로 추정되기 때문에 일단계 표본에서의 모수 추정이 관리도의 성능에 어떤 영향이 있는지 살펴볼 필요가 있다. 이를 위해 관리한계선을 작성한 후 이단계 표본 $Y_1, Y_2, \ldots$으로부터 식 (2.4)과 식 (2.9)을 이용하여 예측값과 잔차를 구

한 후 그 값이 관리한계선 내에 존재하는지 알아보았다. 모수의 추정값이 아니라 참값을 대입하여 오차 $\{\epsilon_t\}$에 대해 관리도를 작성한다면, 관리상태하에서 관리한계선을 처음 넘을 때까지 찍히는 잔차의 개수에 대한 확률분포는 모수가 $P(\epsilon_t < \text{LCL 또는 } \epsilon_t > \text{UCL})$인 기하분포를 따르는데, $\beta = P(\text{LCL} < \epsilon_t < \text{UCL})$라 한다면 이 개수에 대한 평균과 표준편차는 각각 다음과 같이 주어진다.

$$\text{ARL} = \frac{1}{1-\beta}, \qquad \text{SDRL} = \frac{\sqrt{\beta}}{1-\beta}. \tag{3.2}$$

3σ관리한계선을 사용하는 경우 ARL $\approx$ SDRL $\approx$ 370이 된다. 즉, 관리상태하에서는 평균 370개의 간격으로 관리상태를 벗어나는 잔차가 발생한다는 뜻이다. 그러나 이러한 관계는 실제 모수값이 알려져 있거나 관리통계량인 잔차가 서로 독립일때만 만족하는것이다. 따라서 시계열 모형과 일단계 표본의 크기에 따라 잔차에 근거한 관리도가 어떠한 성능을 나타내는지 모의실험을 통해 자세히 살펴보기로 한다.

4. 모의실험

먼저 잔차기반 X 관리도의 성능을 알아보기 위한 모의실험 절차를 설명하면 다음과 같다.

(1) 관리상태하에서의 일단계 표본으로 **AR**(1) 또는 **MA**(1) 모형을 따르는 랜덤표본 $X_1, X_2, \ldots, X_n$을 R 패키지 "forecast"의 "arima.sim" 함수를 이용하여 생성한다.

(2) (1)에서 생성된 데이터에 각 시계열 모형을 적합하고, 잔차 $e_1, e_2, \ldots, e_n$을 구한다.

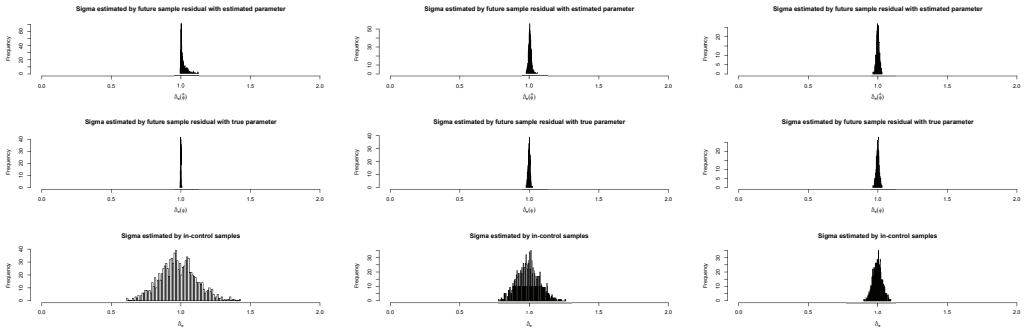
(3) 식 (3.1)을 이용하여, 이단계 모니터링을 위한 관리한계선을 추정한다.

(4) 이단계 모니터링을 위해 시계열 $Y_t \equiv X_t - \mu$ ($\mu = 0, 1, 2, 3$)가 **AR**(1) 또는 **MA**(1) 모형을 따르도록 랜덤표본을 추출하여, (2)에서 구한 모형에 대입하여 새로운 잔차를 계산한 후 (3)에서 설정한 관리한계선을 처음 넘어갈 때까지 걸리는 런 길이(run length)를 기록한다.

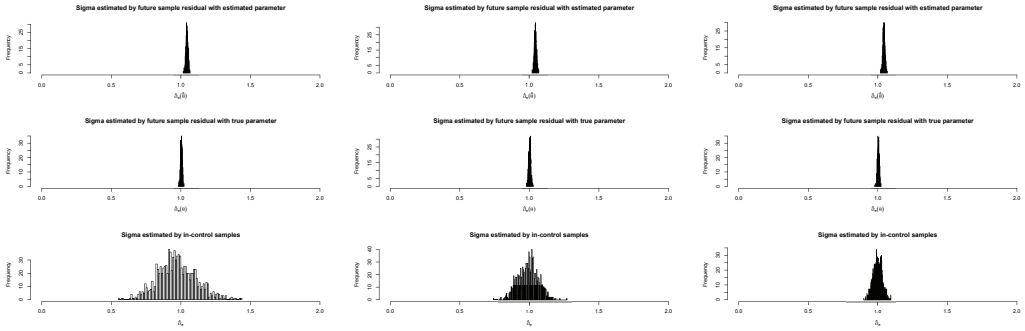
(5) (1)–(4)까지의 단계를 1,000번 반복한다.

이처럼 모의실험 절차는 관리상태하의 일단계 표본으로부터 관리한계선을 추정하고 이단계로 미래 표본에 대한 모니터링을 하고 있는데, 이것은 실제 문제에 관리도를 적용하는 절차와 같다. 따라서, 시계열 모형에 대한 참값이 아닌 추정값을 사용함에 따라 관리도의 성능에 어떠한 영향력이 있는지 살펴봄으로써, 관리도를 적용할 때 주의해야 할 문제가 무엇인지 알아보기로 한다. 본 논문에서는 **AR**(1)과 **MA**(1) 모형에 대하여 일단계 표본의 크기 n을 변화시키며, 관리상태하에서의 런길이의 평균(ARL)과 표준편차(SDRL)를 알아보았다. 이 때, 모형 (2.3)과 (2.8)에서 $\phi_1 = 0.3$과 $\theta_1 = 0.3$으로 하고, ϵ은 $N(0, 1)$을 따르는 $i.i.d$ 랜덤표본으로하여, $X_1, X_2, \ldots, X_n$을 추출하고, 각각 시계열 모형을 적합한 후 잔차를 구하였다. 또한 $\mu = 0$인 경우는 관리상태하에서의 표본을 나타내고, $\mu = 1, 2, 3$인 경우는 이단계 표본에서 시계열 X_t의 모평균이 변화한 것으로 이상상태가 발생한 것을 나타낸다. 일단계 표본의 크기는 $n = 30, 100, 300, 500$로 가정하였다.

관리도의 성능에서 제일 중요한 것은 관리한계선의 추정인데, 여기서는 결국 잔차들에 대한 분산추정과 밀접한 관련이 있다. Figure 4.1은 일단계 표본크기에 따라, 각 그림의 첫 번째 줄에는 이단계 표본 $\{Y_t\}$으로부터 생성된 잔차($= Y_t - \hat{\phi} Y_{t-1}$)들의 표본 표준편차($\hat{\sigma}_e(\hat{\phi})$), 두 번째 줄에는 이단계 표본에서 참 모수값을 이용한 잔차($= Y_t - \phi Y_{t-1}$)들의 표본 표준편차($\hat{\sigma}_e(\phi)$), 마지막으로 세 번째 줄에는 일단계 표본으로 부터 구한 잔차($= X_t - \hat{\phi} X_{t-1}$)들의 표본 표준편차($\hat{\sigma}_e$)의 표본분포를 나타낸다. 일반적으로 세 번째 줄의 표준편차($\hat{\sigma}_e$)로부터 이단계 모니터링을 위한 관리한계선이 설정되는데, 먼저 **AR**(1) 모형을 살펴보면 Figure 4.1 (a)–(c)에서 살펴보듯이 일단계 표본크기 n이 작은 경우에는 $\hat{\sigma}_e$의 산포가 매



(a) $\mathbf{AR}(1) : n = 30$ (b) $\mathbf{AR}(1) : n = 100$ (c) $\mathbf{AR}(1) : n = 500$



(d) $\mathbf{MA}(1) : n = 30$ (e) $\mathbf{MA}(1) : n = 100$ (f) $\mathbf{MA}(1) : n = 500$

Figure 4.1. Distribution of $\hat{\sigma}$ based on phase II sample with estimated parameter $\hat{\theta}$ and with true parameter θ, and phase I sample according to sample size n when data follows $\mathbf{AR}(1)$ and $\mathbf{MA}(1)$ processes.

우 커서 관리한계선이 작거나 크게 추정될 수 있는데, 이 경우 공정 모니터링을 위한 이단계 표본의 잔차를 타점하게 되면 관리한계선을 쉽게 벗어나거나 또는 벗어나기 어려워져 런길이의 산포가 커지게 된다. 그러므로 이단계 표본에서 구한 ARL과 SDRL이 매우 크게 나타날 수 있다. 이러한 현상은 n이 커지면서 줄어드는 것을 알 수 있는데, 이것은 n이 커지면서 관리한계선에서 추정한 표준편차 ($\hat{\sigma}_e$)와 이단계 표본에서 구한 표준편차($\hat{\sigma}_e(\hat{\phi})$)의 산포가 거의 일치하고 있어 이러한 사실을 확인할 수 있다. 따라서 관리한계선 추정을 위한 잔차들의 표준편차를 추정할 경우에는 일단계 표본의 크기가 커야함을 알 수 있다. Figure 4.1 (d)–(e)에서 $\mathbf{MA}(1)$ 모형에서의 여러 잔차의 표본 표준편차를 살펴보면, $\mathbf{AR}(1)$ 모형에서와 마찬가지로 일단계 표본크기가 작은 경우에는 세 번째 줄에있는 $\hat{\sigma}_e$의 산포가 매우 커져 관리한계선이 실제보다 작거나 크게 추정되는 경우로 인해 런길이의 산포가 다소 커지는 현상이 발생했고, n이 커지면서 일단계 표본의 $\hat{\sigma}_e$ 산포가 줄어드는 것을 볼 수 있다. 그런데, $\mathbf{AR}(1)$ 모형과 다른점도 있었는데, 이것은 이단계 표본에서 구한 잔차들의 표준편차($\hat{\sigma}_e(\hat{\theta})$)의 분포가 1보다 항상 커지는 현상이다. 이런 경우 일단계 표본 잔차들의 표준편차($= \hat{\sigma}_e$)로 추정한 관리한계선으로 이단계 모니터링을 할 경우 관리한계선을 넘기기 쉬워져 정상상태에서도 런길이가 짧아질 수 있음을 의미하고, 이것은 표본크기가 큰 n에 대해서도 마찬가지였다. 식 (2.12)의 결과와 비교할 때, 이러한 차이는 $|\hat{\theta}_1|$의 값이 커질수록 분산이 커질 수 있음을 알 수 있다.

이단계 표본에 대하여 관리도의 공정성능을 평가하기 위해 공정이 정상상태와 이상상태일때의 ARL을

Table 4.1. Average run lengths(ARL) and standard average run lengths(SDRL) for residual based individual measurements control chart for monitoring time-series process (with standard errors of ARL's)

	n	Mean shift				Parameter shift	
		0	1	2	3	$\phi = 0.6$	$\phi = 0.9$
AR(1)	30	781.9 (87.6)	200.0 (28.3)	27.3 (1.5)	10.1 (0.7)	418.3 (60.8)	39.8 (2.2)
		2770.6	893.5	47.2	22.4	1921.3	68.9
	100	433.2 (21.4)	110.6 (5.0)	23.1 (1.0)	6.9 (0.2)	236.0 (26.0)	33.4 (1.2)
		678.2	159.4	32.8	6.5	333.4	38.4
	300	388.0 (14.6)	98.0 (3.8)	19.9 (0.6)	7.1 (0.2)	210.1 (7.5)	30.5 (1.1)
		461.0	118.8	18.9	6.4	237.8	33.2
	500	369.3 (12.5)	96.9 (3.2)	20.2 (0.6)	6.6 (0.2)	218.7 (7.5)	29.1 (0.9)
		394.6	99.9	20.0	5.1	236.6	28.8
		0	1	2	3	$\theta = 0.6$	$\theta = 0.9$
MA(1)	30	404.7 (39.0)	57.5 (5.0)	8.6 (0.4)	2.6 (0.1)	168.8 (11.7)	53.1 (2.8)
		1232.6	156.6	12.8	2.7	370.2	87.2
	100	283.8 (12.3)	41.4 (1.6)	7.1 (0.2)	2.5 (0.1)	116.2 (4.7)	45.0 (1.6)
		389.7	51.4	7.1	2.1	149.5	50.8
	300	261.2 (9.3)	41.4 (1.4)	7.4 (0.2)	2.5 (0.1)	120.0 (3.7)	46.6 (1.6)
		293.7	44.4	7.3	2.3	117.4	49.2
	500	245.8 (9.1)	38.3 (1.2)	7.0 (0.2)	2.4 (0.1)	106.6 (3.4)	46.5 (1.5)
		288.3	37.6	6.1	2.0	106.5	47.7

구한 결과가 Table 4.1과 같다. 시계열의 평균 모수 $\mu = 0$가 관리상태를 나타내고, $\mu = 1, 2, 3$인 경우가 이상상태를 나타낸다. 또한 두 모형 모두 모수가 각각 0.6과 0.9로 변화한 이상상태를 가정하였다. **AR**(1) 모형의 경우 관리상태하에서의 평균 런길이를 살펴보면 이전 Shewhart X 관리도에서의 결과와 마찬가지로 표본의 크기가 작은 경우에는 원래 3σ 관리한계선에서 의도한 370보다는 훨씬 큰 평균 런길이를 보인다는 것을 알 수 있다. 또한 SDRL도 매우 커서 기하분포와 잘 맞지 않는 다는 것을 알 수 있다. 그러나 일단계 표본의 크기가 500개만 되어도 관리상태하의 평균 런길이는 369.3으로 원래 설정한 값과 잘 맞는다는 것을 알 수 있다. 이러한 결과는 앞의 Figure 4.1 (c)에서 살펴보듯이 표본의 크기가 클 때에는 일단계 표본에서 구한 잔차들의 표본 표준편차나 이단계 표본에서 구한 잔차들의 표본 표준편차나 모두 원래 참 값에 수렴하는데 기인한다. 평균이 변화한 이상상태에 대하여 잔차기반 관리도는 민감한 결과는 보여주지 못했는데, 시계열 X_t의 평균이 1로 변화했을 때, $n = 30$인 경우에는 평균 런길이가 200으로, 이상상태 발생 후 200번의 런 후에 그 변화를 감지한 것으로 해석된다. $n = 500$인 경우에도 평균 변화 후 약 97번만에 그 변화를 탐지한 것으로 평균변화에 대해 만족할만한 수준은 아니라고 할 수 있다. 평균이 3으로 많이 변화한 경우에도 $n = 500$으로 대표본임에도 불구하고 평균적으로 7번의 런 후에 그 변화를 감지한 것으로 나타났다. 또한 식 (4.1)에서 보이는 것처럼 **AR**(1) 모형의 모수값이 $\phi = 0.3$에서 $\phi = 0.6$과 $\phi = 0.9$로 자기상관이 크게 증가한 경우에는 평균의 커다란 변화를 감지하는 수준으로 그 변화를 탐지하는 것을 알 수 있다.

MA(1) 모형의 경우에는 일단계 표본크기가 작은 경우에 관리상태하의 ARL이 다소 증가하긴 했지만 **AR**(1) 모형과 비교하면 그 차이는 크지 않다. 뿐만 아니라 표본크기가 커진 경우에는 ARL이 3σ 관리한계선에서 기대하는 ARL보다 오히려 짧아지고 있는 현상을 보이고 있다. 예를 들어 $n = 500$으로 큰 경우에도, 평균적으로 246번 중 1번 꼴로 가짜 알람이 나타나게 된다. 이러한 이유는 앞의 Figure 4.1 (d)–(e)의 첫 번째 줄에서 살펴보듯이 이단계 표본에서 생성된 잔차들의 표본 표준편차가 원래 오차항의 산포보다 크게 추정되는 것에 기인한다.

평균이 변화한 이상상태에 대해서는 관리상태하에서의 ARL이 짧아진 만큼, 이상상태에서의 ARL도 **AR**(1) 모형보다 짧아져 좀 더 평균변화의 탐지가 빠르다고 할 수 있다. 그러나 모수가 $\theta = 0.3$에서 $\theta = 0.9$로 변화한 경우를 살펴보면 **AR**(1) 모형보다 ARL이 증가했는데, 이것은 $\theta = 0.9$인 경우 이단계 표본에 대한 잔차의 표준편차가 커진 현상에 기인한 것으로 생각된다.

5. 결론

본 논문은 실제 많은 응용분야에서 사용될 수 있는 시계열 데이터에 대한 잔차기반 관리도의 성능에 대해 알아보았다. Jensen 등 (2006)은 Shewhart X 관리도에서는 모수가 알려진 경우보다 추정되어 사용될 경우 관리상태 또는 이상상태하에서의 ARL과 SDRL이 모두 높게 나타난다고 보고하였다. 본 논문에서는 시계열 데이터에 대한 잔차기반의 X 관리도에 대한 모의실험 연구를 통해 시계열 모형에 따라 모수추정의 영향으로 ARL이 오히려 작아질 수도 있음을 보였다. Shewhart X 관리도에서는 이러한 모수 추정의 영향을 제거하기 위해 일단계 표본의 크기를 300개 이상으로 하라고 권하고 있지만, 앞의 모의실험 결과에서도 살펴보듯이 시계열 모형의 잔차 기반 X 관리도에서는 시계열 모형에 따라 그러한 권고안이 맞을 수도 있지만 그렇지 않을 수도 있음을 알 수 있다. **AR**(1) 모형의 경우에는 Shewhart X 관리도에서 마찬가지로 일단계 표본의 크기가 큰 경우 관리상태하의 ARL = 370으로 관리할 수 있지만, 시계열이 **MA**(1)이나 **ARMA**$(1,1)$ 모형에 대해서는 관리상태하에서의 평균 런길이가 370보다 짧아질 수 있어 가짜 알람률이 커질 수 있음에 유의해야 한다. 다시 말해, 자기상관이 있는 데이터에 대하여 잔차기반 관리도를 적용할 경우에는 시계열 모형에 따른 주의가 필요하다고 할 것이다.

본 논문에서는 시계열 데이터의 참 모형은 알려져 있다고 가정함으로써 모형의 불확실성은 고려하지 않고, 모수의 추정에 대한 불확실성만을 고려하여 관리도의 성능에 어떠한 영향이 있는지 살펴보았다. 앞으로 이와 관련하여 모형의 잘못된 설정에 대해서 관리도의 성능이 얼마나 강건한지 알아보는 것도 매우 중요한 문제일 것이다. 또한 X 관리도이외에 CUSUM 관리도나 EWMA 관리도에 미치는 영향도 함께 알아볼 수 있을 것이다.

감사의 글

바쁘 시간 속에서도 논문을 읽고 많은 조언을 해주신 임요한 교수님과 두 분 심사위원분들께 감사의 마음을 전합니다.

References

Alwan, L. C. and Roberts, H. V. (1988). Time-series modeling for statistical process control, *Journal of Business & Economic Statistics*, **6**, 87–95.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, Prentice-Hall, Englewood, NJ.

Chen, G. (1997). The mean and standard deviation of the run length distribution of $\bar{X}$ charts when control limits are estimated, *Statistica Sinica*, **7**, 789–798.

Cryer, J. D. and Ryan, T. P. (1990). The estimation of sigma for an X Chart, *Journal of Quality Technology*, **22**, 187–192.

Harris, T. J. and Ross, W. H. (1991). Statistical process control procedures for correlated observations, *The Canadian Journal of Chemical Engineering*, **69**, 48–57.

Jensen, W. A., Jones-Farmer, L. A., Champ, C. W. and Woodall, W. H. (2006). Effects of parameter estimation on control chart properties: A literature review, *Journal of Quality Technology*, **38**, 349–364.

Lin, W. S. W. and Adams, B. M. (1996). Combined control charts for forecast-based monitoring schemes, *Journal of Quality Technology*, **28**, 289–301.

Montgomery, D. C. (2012). *Statistical Quality Control: A Modern Introduction*, John Wiley & Sons, New York.

Montgomery, D. C., Johnson, L. A. and Gardiner, J. S. (1990). *Forecasting and Time Series Analysis*, McGraw-Hill, New York.

Nedumaran, G. and Pigniatiello, J. J. (1999). On constructing T^2 control charts for on-line process monitoring, *IIE Transactions*, **31**, 529–536.

Quesenberry, C. P. (1993). The effect of sample size on estimated limits for $\bar{X}$ and X control charts, *Journal of Quality Technology*, **25**, 237–247

Wardell, D. G., Moskowitz, H. and Plante, R. D. (1994). Run-length distributions of special-cause charts for correlated processes (with discussion), *Technometrics*, **36**, 3–27.

Woodall, W. H. and Montgomery, D. C. (1999). Research issues and ideas in statistical process control, *Journal of Quality Technology*, **31**, 376–385.

Zhang, N. F. (1998). A statistical control chart for stationary process data, *Technometrics*, **40**, 24–38.

자기상관 데이터 모니터링에서 일단계 모수 추정이 이단계 관리한계선에 미치는 영향 연구

이성임[a,1]

[a]단국대학교 응용통계학과

(2015년 8월 28일 접수, 2015년 9월 9일 수정, 2015년 9월 9일 채택)

요 약

1920년대에 소개되었던 Shewhart 관리도는 관측치가 서로 독립임을 가정했다. 오늘날은 데이터 측정과 자료수집 기술이 발전하면서 자기상관 공정 데이터가 많이 발생하고 있으며, 이것은 통계적 공정 관리의 성능에 부정적인 영향을 끼치게 된다. 자기상관이 존재하는 데이터에 대하여 가장 쉽게 접근할 수 있는 관리도는 먼저 자기상관구조를 모형화할 수 있는 적절한 시계열 모형을 가정한 다음 잔차를 구하여, 그 잔차에 기반한 Shewhart 관리도를 적용하는 것이다. 실제 문제에서 시계열 모형의 참 모수값은 알려져 있지 않으므로, 이 값은 일단계 표본(과거의 관리상태 표본)으로부터 추정된다. 본 논문에서는 이러한 모수추정이 이단계 표본을 모니터링하는데 어떠한 영향이 있는지 살펴보았다.

주요용어: 자기상관, 시계열 데이터, 잔차기반 관리도, 평균 런길이

이 연구는 2014년도 단국대학교 대학연구비의 지원으로 연구되었음.

[1](448-701) 경기도 용인시 수지구 죽전동, 단국대학교 응용통계학과. E-mail: silee@dankook.ac.kr