한국시뮬레이션학회 논문지
Vol. 19, No. 3, pp. 119-125 (2010. 9)

계절성 데이터의 부트스트랩 적용에 관한 연구

박진수[1] · 김윤배[1†]

A Study of Applying Bootstrap Method to Seasonal Data

Jinsoo Park · Yun Bae Kim

ABSTRACT

The moving block bootstrap, the stationary bootstrap, and the threshold bootstrap are methods of simulation output analysis, which are applicable to autocorrelated data. These bootstrap methods assume the stationarity of data. However, bootstrap methods cannot work if the stationary assumption is not guaranteed because of seasonality or trends in data. In the simulation output analysis, threshold bootstrap method is the best in describing the auto-correlation structure of original data set. The threshold bootstrap makes the cycle based on threshold value. If we apply the bootstrap to seasonality data, we can get similar accuracy of the results. In this paper, we verify the possibility of applying the bootstrap to seasonal data.

Key words : Bootstrap, Seasonality, Simulation output analysis

요 약

시뮬레이션 출력 분석 방법인 이동 블록 부트스트랩이나 정상 부트스트랩, 그리고 임계값 부트스트랩은 자기상관성이 존재하는 데이터에 적용 가능한 표본 재추출 방법론들이다. 이러한 부트스트랩 방법들은 데이터의 정상성을 가정하여 적용해 왔다. 그러나 실제 자료 또는 시뮬레이션 출력에 계절성이나 추세를 동반하여 그 정상성을 보장할 수 없는 경우에는 부트스트랩을 시뮬레이션 출력 분석에 적용하지 못하였다. 시뮬레이션 출력 분석 기법 중 자기상관성을 가장 잘 묘사하는 방법은 임계값 부트스트랩 방법이다. 임계값 부트스트랩은 자료의 임계값을 기준으로 주기를 형성하여 재추출하는 방법으로써 계절성이 존재하는 데이터에 부트스트랩을 적용한다면 임계값 부트스트랩과 유사한 정확도를 얻을 수 있다. 본 논문에서는 계절성이 존재하는 시계열 자료에 대한 부트스트랩 적용 가능성을 제시 및 검증해보고자 한다.

주요어 : 부트스트랩, 계절성, 시뮬레이션 출력 분석

1. 서 론

부트스트랩 방법[1,2]은 독립성이 보장되는 희귀한 데이터를 재추출(data resampling)하여 원하는 통계량의 추정에 이용한다. 시뮬레이션 출력이나 시계열등의 자료 간에 종속관계가 존재하는 경우에도 부트스트랩을 적용 시키려는 시도가 여러 가지 형태로 발전되어 왔다. 발전된 부트스트랩 방법은 일정 수만큼의 블록을 정의하여 재추출하는 이동 블록 부트스트랩(MBB: Moving Block Bootstrap)[4,5], 재생성점(regeneration point)을 기준으로 재추출하는 정상 부트스트랩(SB: Stationary Bootstrap)[8], 그리고 임계값을 기준으로 주기를 형성하여 재추출하는 임계값 부트스트랩(TB: Threshold Bootstrap)[6,7]이 있다. 이 중 임계값 부트스트랩은 데이터의 평균이나 중간값 등의 임계값을 이용하여 주기(cycle)를 형성한 후, 이 주기를 재추출 단위로 이용하는 방법으로서 그 계산 성능척도의 정확도가 다른 방법에 비해 매우 월등하다. 계절성(seasonality)은 주기적으로 변하는 추세를 의미하므로 결국 계절성 데이터를 주기 단위로 재추출하게 되면 임계값 부트스트랩과 비슷한 정확도의 성능척도를 얻을 수 있다.

실제 데이터들은 주기성은 물론 추세(trend)를 포함하는 경우가 많다. 데이터가 증가 혹은 감소 등의 추세를 띠게 되면 주기를 형성한다 해도 부트스트랩을 이용하여 원

2010년 8월 6일 접수, 2010년 9월 20일 채택
[1] 성균관대학교 시스템경영공학과
주 저 자 : 박진수
교신저자 : 김윤배
E-mail; kimyb@skku.edu

본 데이터(original data)를 그대로 묘사하기는 힘들다. 가장 큰 이유는 정상성(stationarity)의 가정이 깨지기 때문이다. 물론 계절성 데이터도 정상성을 보장할 수 없지만 주기 단위로 묶으면 주기끼리는 정상성이 보장되는 경우가 많으므로 부트스트랩 적용이 가능하다.

자기상관(autocorrelation)이 존재하는 정상과정 데이터의 부트스트랩 방법에 있어 가장 많이 사용되는 것은 이동 블록 부트스트랩과 임계값 부트스트랩이다. 물론 정상 부트스트랩이 존재하지만 재생성점을 찾기 어려운 계절성 데이터에는 적합하지 않다. 따라서 본 논문에서는 이동 블록 부트스트랩과 임계값 부트스트랩 방법을 변형하여 적용함으로써 데이터에 존재하는 계절성과 자기상관성을 반영하는 방법을 제시하고 그 실용 가능성을 검증해 보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 세 가지 부트스트랩 방법론에 대해 간략히 설명하고, 3장에서는 계절성 데이터에 부트스트랩을 적용하는 방법을 제시한다. 4장에서는 계절성을 띠는 데이터에 부트스트랩을 직접 적용함으로써 검증을 수행하고 마지막 5장에서는 결론을 맺는다.

2. 종속적 자료를 위한 부트스트랩 방법론

종속적 부트스트랩 방법론은 크게 세 가지로 나눌 수 있다. 이동 블록 부트스트랩, 정상 부트스트랩, 임계값 부트스트랩이다. 이들은 각각 원본 데이터로부터 인접한 데이터들을 묶어서 한 단위로 취급하여 재추출을 수행하는데 이 묶음 단위에 따라 방법론이 달라진다. 본 장에서는 이 세 가지 부트스트랩 방법들에 대해 간략히 살펴보도록 한다.

2.1 이동 블록 부트스트랩

이동 블록 부트스트랩은 재추출 단위를 일정 개수만큼 지정하는 블록(block)으로 지정한다.

그림 1은 이동 블록 부트스트랩의 재추출 단위인 블록 구성방법을 보여준다. $X_1, X_2, \cdots, X_n$이 원본 데이터 시리즈라 하고 b를 한 블록의 크기라고 하면 k번째 블록은 $\boldsymbol{B}_k=\{X_k, X_{k+1}, \cdots, X_{k+b-1}\}$로 정의된다($k=1, 2, \cdots, n-b+1$). 이로부터 임의의 블록을 추출하는 작업을 n개의 데이터가 추출될 때까지 반복하며 n개를 넘기는 데이터는 버린다. 재추출되는 데이터의 수는 원본 데이터의 수와 같아야 하기 때문이다. 따라서 n은 b의 배수가 아닐 수 있기 때문에 식 (1)과 같이 표현되며 마지막 추출에서 원본의 수 n을 만족하는 $n-b\times\lfloor n/b \rfloor$ 개만큼만 취하면 된다.

$$n=b\times\lfloor n/b \rfloor+\{n-b\times\lfloor n/b \rfloor\} \tag{1}$$

이동 블록 부트스트랩에 있어서 가장 중요한 것은 블록의 크기를 결정하는 것이다. 본 논문에서는 평균제곱오차(MSE; mean squared error)를 최소로 하는 방법[3]을 사용하였다. 또한 이는 본 논문에서 제안하는 다른 하나의 방법인 임계값 부트스트랩 방법에도 적용이 가능하며 실험에 있어서도 이 방법을 이용하였다.

2.2 정상 부트스트랩

정상 부트스트랩은 재생성점을 기준으로 재추출 단위를 정의한다. 정상 부트스트랩의 재생성점 이해를 위해 그림 2를 보자.

그림 2가 $M/G/1$ 대기행렬시스템의 고객수 그래프라고 하고 T개의 재생성점으로 이루어져 있다고 하자. $X_1, X_2, \cdots, X_n$이 고객의 대기시간과 같은 연속된 원본 데이터라고 하면 재생성점을 기준으로 다음과 같은 재추출 단위가 생성된다.

$$\boldsymbol{R}_k=\{X_{S_{k-1}+1}, X_{S_{k-1}+2}, \cdots, X_{S_k}\} \quad k=1,2,\cdots,T$$

정상 부트스트랩도 이동 블록 부트스트랩과 마찬가지로 임의의 $\boldsymbol{R}_k$를 추출하는 작업을 n개의 데이터가 추출될 때까지 반복하며 n개를 넘기는 데이터는 버린다.

정상 부트스트랩은 시뮬레이션 출력분석 방법인 재생성 방법(regeneration method)과 부트스트랩 방법의 조합

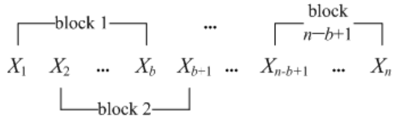

그림 1. MBB의 재추출 단위

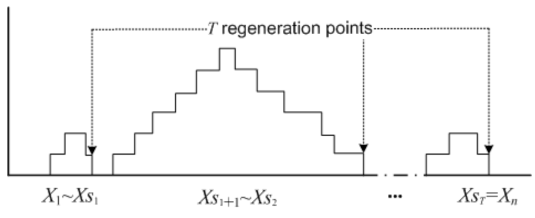

그림 2. SB의 재추출 단위

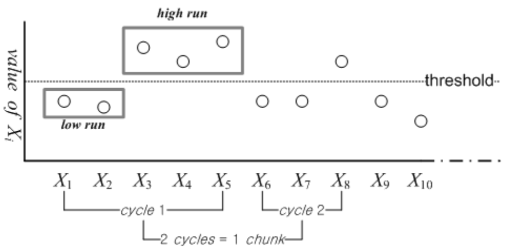

그림 3. TB의 재추출 단위

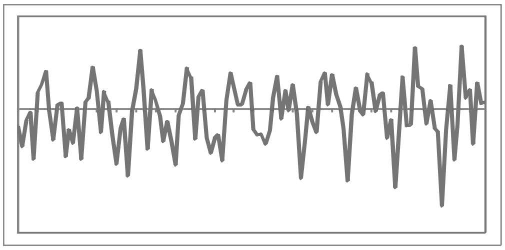

그림 4. 계절성 시계열 데이터의 예

된 방법이라 할 수 있으므로 비교적 좋은 성능을 보이지만 재생성점이 존재하지 않는 데이터에는 전혀 사용할 수 없다는 단점이 있다. 따라서 본 논문에서 사용하는 계절성 데이터에는 적합하지 않으므로 일반적인 적용이 불가능하다.

2.3 임계값 부트스트랩

임계값 부트스트랩은 원본 데이터의 평균 또는 중간값(median)과 같은 임계값을 기준으로 low run과 high run을 한 쌍으로 사이클(*cycle*)을 생성한다. 임계값 부트스트랩은 이 사이클을 일정 개수만큼 묶어 하나의 청크(*chunk*)로 정의하고 이를 재추출 단위로 삼는 것이다.

그림 3은 임계값 부트스트랩의 주기와 청크를 보여주는 예로써 두 개의 사이클이 하나의 청크를 형성하고 있다. $\boldsymbol{X}=\{X_1, X_2, \cdots, X_n\}$를 원본 데이터라고 하면 임계값에 의해 R개의 청크가 형성된다고 하자. 이로부터 $\boldsymbol{C}_i=\{X_{i,1}, X_{i,2}, \cdots, X_{i,n_i}\}$를 크기 n_i인 i번째 청크($i=1, \cdots, R$, $\Sigma_{i=1}^{R} n_i=n$)라고 하고 임의의 $\boldsymbol{C}_i$를 추출하는 과정을 반복한다. 물론 앞의 두 방법과 마찬가지로 재추출된 데이터가 n개를 넘어가게 되면 이후 데이터는 버린다.

임계값 부트스트랩 방법은 데이터의 자기상관성을 가장 잘 묘사하는 재추출 방법론이며, 이동 블록 부트스트랩이나 정상 부트스트랩 방법에 비해 우수한 성능척도를 구할 수 있는 것으로 보고되었다[7].

3. 계절성 데이터 부트스트랩

계절성이 존재하는 시계열 데이터는 정해진 주기를 기준으로 일정 패턴(pattern)을 반복한다. 가장 많이 나타나는 주기는 월별 데이터의 경우 분기별로 4의 배수로 나타나고 월별로 12의 배수로 나타난다. 일별 데이터는 1주일인 7 또는 한 달인 30의 배수로 나타나기도 한다.

그림 4는 계절성이 반영된 시계열 데이터이다. 그림 4의 데이터는 Seasonal AR(1) 모형으로써 다음과 같은 시계열 식의 시뮬레이션 결과이다.

$$z_t=0.9z_{t-12}+\varepsilon_t \tag{2}$$

여기서 ε_t는 백색잡음(white noise)을 의미하며 $N(0,\sigma^2)$를 따른다. 그림 4의 데이터에서는 일반적으로 시계열 분석에 많이 사용되는 표준정규분포를 사용하였다.

시계열 모형에 계절성을 부여한 모형들이 다수 있으나 실제 데이터들은 모형에 적용가능하지 않은 경우도 많다. 모델 적용이 되지 않는 경우 혹은 적용이 되는 경우라도 통계적 분석이 어려운 경우가 많은데 이러한 경우에 재추출 방법은 매우 유용한 도구이다. 본 논문에서는 이러한 주기를 갖는 계절성 데이터에 대한 두 가지 부트스트랩 방법론을 제시한다.

3.1 주기를 이용한 이동 블록 부트스트랩의 적용

계절성 데이터의 재추출 방법으로서 간단한 방법은 다음과 같다. 계절성 데이터는 주기마다 유사한 패턴이 반복되게 된다. 따라서 부트스트랩 적용을 위하여 재추출 최소 단위를 주기로 설정하여 사용한다. 데이터에 내재된 상관관계 구조(correlation structure)를 정확히 반영하기 위해서는 한 주기만을 재추출 단위로 사용하는 것이 아니라 주기의 묶음 단위로 재추출을 시행해야 한다. $\boldsymbol{X}=\{X_1, X_2, \cdots, X_n\}$를 원본 데이터라 하고 주기를 c라 하면 데이터는 $q=\lfloor n/c \rfloor$ $(n=cq+r)$ 개의 주기가 형성되며 주기 데이터의 특성상 나머지 r개에 대한 부분은 버리게 되는데 이 때문에 원본데이터를 모두 반영할 수 없다는 단점이 있다. i번째 주기 데이터 집합 $\boldsymbol{Y}_i$는 다음과 같이 정의된다.

$$\boldsymbol{Y}_i=\{X_{c(i-1)+1}, X_{c(i-1)+2}, \cdots, X_{ci}\},\ i=1,2,\cdots,q \tag{3}$$

결국 수정된 원본데이터는 $\boldsymbol{X}'=\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_q\}$로 구성된다. 이로부터 이동 블록 부트스트랩에서의 방법과 동일한 방법을 적용하여 재추출을 시행하게 된다. 물론 $\boldsymbol{Y}_i$는 하나의 데이터 값이 아닌 데이터 집합이므로 이동 블록 부트스트랩과는 차이점이 있으나 방법론 자체는 대동소이하다. 이 방법을 사용하면 데이터의 계절성과 상관관계 구조를 반영할 수 있게 된다. 이 방법은 주기별 상관관계는 물론 주기 내에 잠재된 상관관계까지 반영할 수 있게 된다. 다만 앞에서 언급했듯이 원본데이터를 모두 재사용하지 않는다는 단점이 있다.

3.2 임계값 부트스트랩의 적용

식 (2)로부터 알 수 있듯이 시점 t와 시점 $t-12$의 데이터는 AR(1)의 관계를 갖는다. 이로부터 일반적으로 주기가 c인 계절성 데이터들은 c의 간격을 두고 정상과정(stationary process)을 형성하고 있다고 할 수 있다. 따라서 데이터를 c개의 집합으로 재구성한다. 즉, $\boldsymbol{X}=\{X_1, X_2, \cdots, X_n\}$를 원본 데이터라 하면 다음과 같은 c개의 데이터 집합을 구성할 수 있다.

$$
\begin{aligned}
\boldsymbol{S_1} &= \{X_1\, X_{1+c} \cdots X_{1+(c-1)q}\, X_{1+cq}\} \\
\boldsymbol{S_2} &= \{X_2\, X_{2+c} \cdots X_{2+(c-1)q}\, X_{2+cq}\} \\
&\vdots \\
\boldsymbol{S_r} &= \{X_r\, X_{r+c} \cdots X_{r+(c-1)q}\, X_{r+cq}\} \\
&\vdots \\
\boldsymbol{S_c} &= \{X_c\, X_{2c} \cdots X_{cq}\ \}
\end{aligned}
$$

여기서 $n=cq+r\ (r<c)$ 이다. $\boldsymbol{S_i}$는 주기내의 i번째 데이터들로 구성된 데이터 시리즈임을 의미한다.

각 $\boldsymbol{S_i}$로부터 임계값 부트스트랩을 적용하여 재추출한 결과를 $\boldsymbol{S_i^*}$라 하면 $\boldsymbol{S_i^*}$들을 이용하여 새로운 n개의 데이터를 재구성할 수 있다. 이 방법을 이용하면 역시 계절성과 함께 주기마다 잠재되어있는 상관관계 구조를 반영할 수 있게 된다.

4. 시뮬레이션 및 결과 분석

본 절에서는 3절에서 제안한 방법론의 적용에 대한 타당성을 검증한다. 이를 위해 시뮬레이션을 수행하여 계절성 데이터를 생성한 후 이로부터 제안한 부트스트랩 방법의 결과를 분석하여 그 적용 가능성을 검증한다.

4.1 시뮬레이션 적용 모형

본 실험에서는 계절성 모형 중 SAR(seasonal autoregressive) 모형과 SMA(seasonal moving average) 모형을 다음과 같이 구성하여 사용한다.

$$\text{SAR}: \ z_t = 0.9z_{t-4} + \varepsilon_t \ \ (\phi = 0.9) \tag{4}$$

$$\text{SMA}: \ z_t = \varepsilon_t + 0.9\varepsilon_{t-7} \ \ (\theta = -0.9) \tag{5}$$

첫 번째 모형은 $SAR(1)_4$ 모형으로서 분기별 모형이라고 할 수 있으며 두 번째는 $SMA(1)_7$ 모형으로서 요일별 모형의 예라 할 수 있다. 본 실험에서 사용된 백색잡음 ε_t는 표준정규분포를 따르도록 하였다. 물론 여러 가지 실험을 통해 σ^2의 값이 다르더라도 동일한 결과를 갖는 것을 확인하였지만 일반적으로 가장 많이 사용되는 표준정규분포를 사용한 결과를 명시하였다.

4.2 재추출 단위 분석

이동 블록 부트스트랩과 임계값 부트스트랩을 적용하기 위해서는 이동 블록 부트스트랩의 최적 블록 크기(optimal block size)와 임계값 부트스트랩의 최적 청크 크기(optimal chunk size)를 결정하는 것이 가장 중요하다. 본 논문에서는 이를 구하기 위해 3,000개의 원본데이터를 생성한 후, 이로부터 각 블록 크기와 청크에 대해 300번의 부트스트랩을 반복하여 평균을 구한다. 이러한 과정을 각각의 블록 크기와 청크 크기에 대해 1,000번 수행한 결과로부터 평균제곱오차를 계산하여 그 값이 최소가 되는 최적 블록 크기와 최적 청크 크기를 결정한다[7].

다음 표 1은 이로부터 구한 최적 청크 크기와 최적 블록 크기를 나타낸다. 이 중 $SAR(1)_4$ 모형의 결과는 Park & Willemain이 발표한 논문[7]의 $AR(1)$, $\phi=0.9$ 모델의 결과와 거의 일치하는 것을 알 수 있다. 이는 실제로 $SAR(1)_4$ 모형의 계절성을 배제하면 일반 $AR(1)$모형이 되기 때문이다. 표 1에서 TB의 경우 최적의 평균 청크 크기를, MBB의 경우 최적의 블록크기를 의미한다.

표 1. 최적의 블록 크기와 청크 크기

	Optimal chunk/block size	
	TB	MBB
$SAR(1)_4$	14.06	40
$SMA(1)_7$	5.93	4

4.3 부트스트랩 결과 분석

이 절에서는 본 논문에서 제안한 부트스트랩 방법들의 적용 결과와 그 적용 가능성을 검토한다. 먼저 각각의 모형에 대해 시뮬레이션을 수행하여 3,000개씩의 데이터를 생성한다. 다음으로 이 데이터를 이용하여 본 논문에서 제안한 두 가지 부트스트랩 방법을 적용하여 그 결과를 분석해 보고자 한다.

그림 5와 그림 6은 앞의 두 모형에 대해 부트스트랩을 적용한 결과 중 임의의 5개 주기를 추출하여 타점한 결과이다. 원본 데이터가 기본 모형에 대한 시뮬레이션 수행 결과이므로 주기성이 뚜렷하게 나타나지는 않지만 관찰해 보면 그림 5는 주기 4를, 그림 6은 7의 주기를 띠고 있음을 확인할 수 있다. TB와 MBB를 응용하여 적용한 결과 역시 같은 주기를 띠고 있음을 알 수 있다.

이 그림만으로는 주기성 반영에 대한 결과는 알 수 있으나 모형 자체를 반영하는지는 확인 할 수 없다. 따라서 다음 단계로는 이에 대한 분석을 수행해보고자 한다.

모형의 자체를 정확히 반영하고 있는지 확인하기 위하여 몇 가지 분석을 수행해야만 한다. 첫 번째는 성능척도들을 분석하는 방법이다. 표 2는 성능척도들 중 평균과 분산의 계산 결과를 보여준다. 표 2에서 괄호안의 값은 원본 데이터에 대한 상대오차에 대한 백분율로서 단위는 %이며 나열된 모든 값은 소수점 셋째 자리에서 반올림한 결과이다.

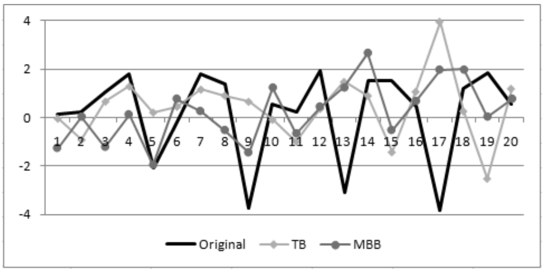

그림 5. $SAR(1)_4$ 데이터의 부트스트랩 적용 결과

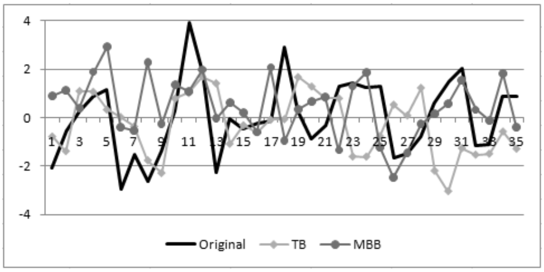

그림 6. $SMA(1)_7$ 데이터의 부트스트랩 적용 결과

평균에 대한 오차가 큰 것은 이론적 평균값 자체가 0이며 데이터의 양이 충분치 않으므로 의미 없는 결과라 할 수 있다. 만일 원본의 데이터 수가 모형의 평균 0을 반영할 만큼 충분하다면 그 결과도 보다 정확해 질 것이다. 평균에 반해 분산의 값은 비교적 정확한 값으로 나타나고 있다. 이는 데이터 분석에 있어서 매우 중요한 요소인 분산을 적은 양의 데이터로도 정확하게 계산할 수 있다는 것을 의미한다. 표 2의 결과를 보면 TB적용보다는 MBB의 적용이 조금 더 정확한 결과를 나타내고 있으나 데이터에 존재하는 상관관계 구조를 얼마나 잘 반영하는지는 확인할 수 없다.

데이터에 존재하는 상관관계 구조의 반영을 살펴보기 위해 자기상관 계수(autocorrelation coefficient)를 살펴보아야 한다. 실험에 사용된 두 모형은 이론적인 자기상관계수를 구할 수 있으므로 이에 대한 분석을 수행해 보고자 한다. 먼저 SAR 모형의 경우 4의 배수에 대한 lag의 이론적인 자기상관 계수를 살펴보아야 하며 SMA 모형의 경우 7의 배수에 대해 살펴보아야 한다. 이론적인 결과에 따르면 실험에 사용된 모형들의 lag $4k$ 자기상관 계수는 다음과 같다.

$$SAR(1)_4 \;:\; \rho_{4k} = \phi^k = 0.9^k, \qquad k = 1, 2, \cdots \tag{6}$$

$$SMA(1)_7 \;:\; \rho_{7k} = \begin{cases} \dfrac{-\theta}{1+\theta^2} = \dfrac{0.9}{1+0.9^2}, & k = 1, \\ 0, & k > 1. \end{cases} \tag{7}$$

자기상관 계수를 계산한 결과를 표 3에 나열하였다. 표 3에 나열된 값들은 모두 소수 다섯째 자리에서 반올림한 결과이다. 결과를 살펴보면 임계값 부트스트랩의 경우 자기상관 계수를 정확히 맞추는 반면 이동 블록 부트스트랩의 경우 과소추정(underestimate)되고 있음을 알 수 있다. 이는 앞서 설명했듯이 임계값 부트스트랩이 다른 방법들에 비해 상관관계 구조를 더 잘 반영한다는 것을 확인할 수 있는 결과이다. $SMA(1)_7$ 모형에서 lag 14와 21의 경

표 2. 부트스트랩 적용 데이터의 평균 및 분산

	$SAR(1)_4$			$SMA(1)_7$		
	원본	TB	MBB	원본	TB	MBB
평균	-0.31	-0.41 (32.02)	-0.23 (26.98)	0.05	0.02 (63.27)	0.09 (69.80)
분산	5.85	5.81 (0.68)	5.86 (0.26)	1.75	1.72 (2.12)	1.74 (1.07)

표 3. 부트스트랩 적용 데이터의 자기상관 계수

	lag	이론값	원본	TB	MBB
$SAR(1)_4$	4	0.9000	0.9117	0.9069	0.8918
	8	0.8100	0.8278	0.8182	0.7904
	12	0.7290	0.7437	0.7276	0.6978
	16	0.6561	0.6649	0.6478	0.6119
	20	0.5905	0.5941	0.5794	0.5388
$SMA(1)_7$	7	0.4972	0.4963	0.4776	0.3732
	14	0.0000	0.0015	-0.0066	0.0169
	21	0.0000	0.0245	-0.0285	-0.0282

우 이론값이 0이므로 이때 발생한 오차는 표본으로부터 발생하는 것으로 간주하여 무시할 수 있는 수치이다.

본 실험으로부터 계절성 데이터에 부트스트랩을 적용한 결과로써 다음과 같은 결론을 도출할 수 있다. 먼저 계절성 자체를 정확히 반영할 수 있다는 점이 그 첫 번째 이다. 다음으로 추정된 분산의 정확성을 들 수 있다. 데이터 분석에 있어서 가장 많이 사용되는 분산을 적은 양으로도 정확히 추정함으로써 적은 데이터를 사용하더라도 정확도 높은 분석을 수행할 수 있다는 점이 두 번째 장점이다. 마지막으로 상관관계 구조의 반영에 대한 정확성을 확인할 수 있다. 부트스트랩을 이용하는 가장 큰 이유가 바로 상관관계 구조의 반영에 있으므로 그 장점을 그대로 가지고 있다는 것이다. 즉, 계절성 데이터에 존재하는 상관관계 구조를 정확히 반영함을 확인함으로써 분석의 정확성을 더할 수 있다는 것이 세 번째이다. 이로부터 계절성 데이터에 부트스트랩 방법을 적용한 결과는 매우 뛰어난 성능을 보임을 알 수 있다.

5. 결론 및 추후연구과제

본 논문에서는 계절성 데이터에 이동 블록 부트스트랩과 임계값 부트스트랩 방법을 적용하고 활용 가능성을 검증해 보았다. 두 가지 부트스트랩 방법을 적용하여 재추출한 데이터는 원본 데이터의 계절성은 물론 내재된 상관관계 구조를 그대로 묘사할 수 있었다. 물론 기존 연구들에서 보이듯이 임계값 부트스트랩이 그 정확도 면에서 우월함도 확인하였다. 이로부터 계절성이 존재하는 시뮬레이션 출력 또는 계절성 데이터 분석에 있어 많은 시간과 비용을 줄일 수 있는 유용한 도구로써 부트스트랩을 활용할 수 있다는 것을 재확인할 수 있다.

실제 데이터들은 계절성은 물론 추세가 존재하기도 하므로 추세가 존재하는 데이터에 부트스트랩을 적용할 경우에 대해서도 방법론을 개발한다면 가치 있는 연구라 할 수 있겠다.

참 고 문 헌

1. Efron, B., “Bootstrap methods: another look at the jack-knife,” Annals of Statistics, vol. 7, pp. 1-26, 1979.
2. Efron, B. and Tibshirani, R., An introduction to the bootstrap, Chapman & Hall, New York, 1993.
3. Hall, P., Horowitz, J. and Jing, B., “On blocking rules for the bootstrap with dependent data,” Biometrika, vol. 82, pp. 561-574, 1995.
4. Kunsch, H., “The jackknife and the bootstrap for general stationary observations,” Annals of Statistics, vol. 17, pp. 1217-1241, 1989.
5. Liu, R. and Singh, K., “Moving blocks jackknife and bootstrap capture weak dependence,” In: LePage, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 225-248, 1992.
6. Park, D., Kim, Y.B., Shin, K.I. and Willemain T.R., “Simulation output analysis using the threshold bootstrap,” European Journal of Operational Research, vol. 134, pp. 17-28, 2001.
7. Park, D. and TR. Willemain, “The threshold bootstrap and threshold jackknife,” Computational Statistics & Data Analysis, vol. 31, pp. 187-202, 1999.
8. Politis, D.N. and Romano, J.P., “The stationary bootstrap,” Journal of American Statistics Association. 89, pp. 1303-1313, 1994.
9. Wei, W.W.S, Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley Publishing Company, Inc, 1990.

김 윤 배 (kimyb@skku.edu)

1982 성균관대학교 산업공학 학사
1986 University of Florida, Industrial and Systems Engineering 공학석사
1992 Rensselaer Polytechnic Institute Decision Science and Engineering Systems Ph. D.
1995～1998 성균관대학교 산업공학과 조교수
1998～2004 성균관대학교 산업공학과 부교수
2004～현재 성균관대학교 산업공학과 교수

관심분야 : 시뮬레이션 방법론, 기술시장 분석, 에너지 수요예측

박 진 수 (jsf001@skku.edu)

1998 성균관대학교 산업공학과 학사
2000 성균관대학교 산업공학과 석사
2008 성균관대학교 산업공학과 박사
2008～현재 성균관대학교 시스템경영공학과 박사후연구원

관심분야 : 시뮬레이션 출력분석, 시뮬레이션모델링, 대기행렬시스템