한국경영과학회/대한산업공학회 2003 춘계공동학술대회
2003 년 5 월 16 일-17 일 한동대학교(포항)

임계값 붓스트랩을 사용한 입력 시나리오의 생성

김윤배[1] • 김재범[1] • 고종석[1]

[1]성균관대학교 시스템경영공학부

Generation of Simulation Input Data Using Threshold Bootstrap

Yun-Bae Kim[1] • Jae-Bum Kim[1] • Jong-Suk Ko[1]

Abstract

시뮬레이션 상의 입력모델에 대한 기존의 연구는 과거의 자료를 바탕으로 선형의 모수적인 (parametric) 모형을 개발하는데 초점을 두고 있다. 그러나 이 경우에는 입력이 매우 복잡한 형태를 가지면 모수적인 모형을 찾는 것이 불가능해지므로 비모수적인(non-parametric) 접근방법이 절실한 실정이다. 예로 인터넷 트래픽 모델의 시뮬레이션 수행시 입력으로 제공되는 단위 시간당 요구되는 웹 페이지의 수 같은 경우 데이터들 간에 종속관계가 매우 심하고 복잡하여 모수적 모형을 세우는데 어려움이 있다. 이러한 시스템들을 시뮬레이션 방법으로 분석하고자 할 때, 기존의 trace-driven 시뮬레이션 방법이나 모수적 모형을 찾아 다수의 사실적인 시뮬레이션 입력 자료를 확보하는 것은 현실적으로 어려움이 있다. 따라서, 비모수적인 방법으로 다수의 사실적인 시뮬레이션 입력 자료를 생성하는 것이 필요하다. 이러한 비모수적인 방법에 대한 평가기준 설정은 시뮬레이션 상의 입력 모델에 대한 타당성을 제시한다는 점에서 또한 매우 중요하다. 본 논문에서는 붓트스트랩의 방법중의 하나인 임계값 붓트스트랩을 이용하여 시뮬레이션 입력 자료 생성 방법을 개발하였고 Turing test를 통해 붓스트랩으로 생성한 입력 시나리오를 검증하였다.

1. 서론

붓트스트랩은 원시의 자료군에서 재 추출한 유사 자료군으로 원시 자료에 대한 추론을 계산하는 비모수적인 기법이다. 원래는 iid가 만족되는 자료에 대한 분석 기법으로 출발하였으나 최근에는 시계열 자료같이 종속적인 관계를 갖는 자료군에도 적용이 확장되고 있다.(Künsch, 1989; Politis와 Romano, 1994; Park과 Willemain, 1999). 이러한 종속적인 자료에의 적용은 주로 시뮬레이션 출력분석에 응용을 예로 들고 있고 (Kim, 1993 ; Park, 2001), 시뮬레이션 입력 자료 생성에의 적용은 최근에 발표되었다.(Demirel과 Willemain, 2001).

이러한 단일 실측 자료로부터 붓트스트랩 반복을 생성함에 있어서 붓트스트랩 방법이 얼마나 원시 실측자료를 유사하게 재생해내는지를 평가하는 기준이 필요하다.

붓트스트랩은 모든 통계적 분야(Efron 1979; Bickel과 Freedman 1981; Gotze와 Künch 1993; Hall과 Jing 1996)에서 기존의 직접 추론의 문제에 중점을 둔다. 즉, 반복으로 추출 되어진 자료 그 자체보다는 붓트스트랩 반복으로부터 계산되어진 통계치의 특성을 설명하는 것이다.

따라서, 고전의 통계 분야에서처럼 붓트스트랩 추정량의 특성; 불편성, 일치성, 효율성, 점근적인 정상성에 주안점을 둔다. 가령, 붓트스트랩 신뢰구간이 알맞은 범위 특성을 갖는다면, 그 간격에 기초한 붓트스트랩 반복열은 적당하다고 판단될 것이다. 실제 필요로 하는 것은 다양성과 충실도의 정확한 배합이라고 할 수 있다. 즉 반복을 통한 다양성, 원시 자료의 필수적인 요소로서 충실도가 제공되어야 한다.

붓트스트랩 재추출을 통하여 추구하는 것은 원시 자료의 다양한 독립적 실제 값을 갖게 되는 이상적 상황을 재창조하는 것이다. 다시 말하면 원시 자료와 그 원시 자료의 독립 반복 사이에서 발견할 수 있는 동일한 정도의 유사성과 차이를 갖는 붓트스트랩 반복을 구성하기를 원한다는 것을 의미한다. 그러한 독립 반복을 자매 자료열(sister series)로 생각할 수도 있다. 즉, 붓트스트랩 재추출 과정을 통해 생성된 각 반복이 자매 자료열로서의 특성을 갖는지 여부를 비교하는 것이 곧 붓트스트랩 입력 자료열의 신뢰성을 보장하는 것이 된다.

이를 위해 본 논문에서는 임계값 부트스트랩을 사용하여 입력자료를 생성하고 독립반복의 다양성과 유사성을 Turing 검정과 같은 정성적 방법을 통해 분석하여 방법의 타당성을 평가하도록 한다.

2. 임계값 붓트스트랩(Threshold Bootstrap)

Efron(1979)은 재추출 방법(Resampling)을 이용하여 통계적 추론을 수행할 수 있는 비 모수적인 접근 방법인 붓트스트랩을 고안하였다. 붓트스트랩 방법은 모수적인 추론이 어렵거나 해석적인 방법이 존재하지 않을 때 좋은 결과를 얻을 수 있다고 알려져 있다(Efron, Tibshirani 1993). Singh(1981)은 상관관계가 존재하는 데이터는 고전적인 붓트스트랩 방법을 적용하는데 문제점이 있음을 지적하였는데 이는 고전적인 붓트스트랩 방법에서의 재추출 단위가 하나의 데이터이기 때문에 데이터들간의 상관관계가 무시될 수 있다고 했다. 따라서 데이터들의 상관관계를 유지하기 위해서 붓트스트랩의 재추출 단위를 조정하는 방법이 고안되었는데 대표적인 방법이 임계값 붓트스트랩이다.

임계값 붓트스트랩(TB)은 시뮬레이션 출력분석을 위한 비모수적인 추론 방법으로 Kim(1993)에 의해 고안되었다. Kim(1993)은 상관관계가 존재하는 자료를 해석하기 위하여 고전적인 붓트스트랩의 재추출 단위인 하나의 데이터를 자료내에 존재하는 종속관계를 유지하도록 '주기' (Low Run 과 High Run)로 바꾸었다. Kim(1996)은 재추출 단위를 주기로 바꾼 상태에서 재추출된 붓트스트랩 자료가 원시 자료에 존재하는 종속관계를 유지하는 것을 보였다.

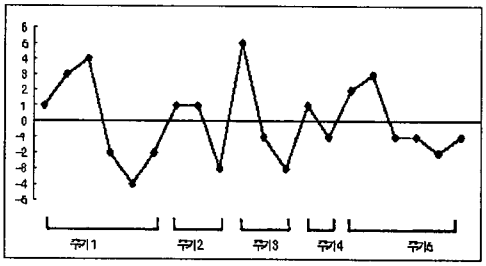
임계값을 시계열을 관통하는 수준(표본평균이나 표본의 중앙값 등)으로 정하면 시계열은 임계값 보다 높은 자료의 연속인 High Run과 낮은 자료의 연속인 Low Run으로 구분되어 진다. 재추출의 기본 단위를 중복되지 않고 연속되는 두 개의 Run(High Run 와 Low Run의 결합)으로 정하고, 이를 '주기'(cycle)라고 하였다. 임계값을 정하면 관측치 자체가 주기를 결정하게 되어, 재추출의 단위가 자동으로 정해지는 것이 TB의 장점이다. 또한 임계값 붓트스트랩은 일반적인 시뮬레이션 출력 분석에 사용되는 다수의 반복 시뮬레이션 실행 결과를 붓트스트랩 표본으로 대치할 수 있어서 시뮬레이션 실행시간을 단축할 수 있고, 동류의 단일실행 분석기법인 배치평균방법(Batch Mean)보다 사용하기 쉽다는 장점이 있다.

본 논문에서는 임계값 붓트스트랩을 사용하여 원시 자료의 붓트스트랩 표본을 생성하고 이를 시뮬레이션 입력 자료로 사용할 것이다.

확률 체계 P 에 의해 생성된 정상 시계열 $\{X_t,$

$X_2, \ldots, X_n$ } 이 약한 종속 구조를 갖고, 임의의 임계값이 R 개의 주기를 만든다고 가정하자. $C_i = \{ X_{i,1}, X_{i,2}, \ldots, X_{i,m} \}$ 를 크기가 n_i 인 i 번째 주기라면, $i = 1, 2, \ldots, R$ 이고 $\sum_{i=1}^{R} n_i = n$ 이다. 주기의 크기 n_i 와 주기의 수 R 은 둘 다 확률변수이다. 이 두 개의 확률 변수는 자료의 자기상관과 임계값에 매우 강한 종속관계를 갖는다. TB는 주기들의 집합 $\{ C_1, C_2, \ldots, C_R \}$ 에서 확률 1/R로 무작위로 추출하여 붓트스트랩 표본을 만든다. 미지의 P 는 재추출 구조에 근거해 P^* 로 추정 할 수 있다. 예로, 시계열 $\{ X_1, X_2, \ldots, X_n \}$ 의 표본 평균 $\overline{X_n}$ 의 편의(bias) 와 표준오차(standard error)를 추정하려 할 때, $\{ X^*_1, X^*_2, \ldots, X^*_n \}$ 가 P^* 에서 얻은 붓트스트랩 표본이라면, $P^* \rightarrow \{ X^*_1, X^*_2, \ldots, X^*_n \}$ 라고 표현하고 이때 표본평균 $\overline{X_n}$ 의 편의와 표준오차에 대한 붓트스트랩 추정치는 재추출 확률구조 P^* 에 의거한 $\overline{X_n}$ 의 편의와 표준오차와 일치하게 된다.

[그림 1]을 보고 TB를 사용하여 표본평균의 편의와 표준오차를 추정하는 방법을 이해해 보자.



주기1 = { 1,3,4,-2,-4,-2 }, 주기2 = { 1,1,-3 }, 주기3 = { 5,-1,-3 }

주기4 = { 1,-1 }, 주기5 = { 2,3,-1,-1,-2,-1 }

임계값 = 표본 평균 = 0

TB표본 1: {1,-1 1,1,-3 1,1,-3 1,3,4,-2,-4,-2 5,-1,-3 1,1,-3}

$\overline{X}^*_{1} = -0.1$

TB표본 2: {2,3,-1,-1,-2,-1 5,-1,-3 1,-1 1,1,-3 5,-1,-3 5,-1,-3}

$\overline{X}^*_{2} = 0.1$

TB표본 3: {1,3,4,-2,-4,-2 2,3,-1,-1,-2,-1 1,-1 5,-1,-3 1,1,-3}

$\overline{X}^*_{3} = 0.0$

표본평균의 편의 = (-0.1 + 0.1 + 0.0) / 3 - (표본평균) = 0.0

표본평균의 표준오차 = { -0.1, 0.1, 0.0 } 의 표준편차 = 0.1

[그림 1] 임계값 붓트스트랩의 재추출 방법

TB의 재추출 알고리즘을 정리해보면 다음과 같다.

Step 1 : n 개로 이루어진 자기상관이 존재하는 자료를 획득한다.

Step 2 : 임계값을 설정한다. (예를 들면, 표본평균이나 표본의 중간값 등)

Step 3 : 자료를 임계값을 중심으로 두 개의 자료군으로 분류한다. 자료군은 임계값보다 높은 자료의 연속인 High Run과 낮은 자료의 연속인 Low Run 으로 구분되어 진다. 중복되지 않고 연속되는 두 개의 Run(High Run과 Low Run의 결합)이 하나의 주기가 된다.

Step 4 : R 개의 주기에 추출될 확률을 각각 1/R로 할당하고, 균등분포(0,1)를 따르는 난수를 발생시켜 재추출할 주기를 선택한다. 복원을 허락하여 재추출한 주기들을 연결하여 붓트스트랩 표본을 원시 자료의 크기 n 만큼 생성한다.

Step 5 : 재 생성된 붓트스트랩 표본으로 관심 있는 통계량을 계산한다.

Step 6 : 단계 5와 단계 6을 총 B 번 반복한다.

Step 7 : B 번의 반복으로 계산된 통계량의 추정치를 계산한다.

TB는 원시 자료가 임계값 수준을 교차하는 횟수와 매우 밀접한 관계가 있다. 또한 강한 양의 상관관계를 가진 원시 자료에서는 주기의 숫자가 적고 주기의 길이가 긴 반면 음의 상관

관계의 자료에서는 주기의 숫자가 많고 그 길이가 짧다. 경험적으로 AR(1) 시계열의 평균주기의 길이와 1차 자기회귀계수와 관계를 -0.9에서 +0.9로 변화시켜 보면 평균주기의 길이가 증가한다. 따라서 일반적으로 교차횟수가 많을수록, 재추출 단위가 짧아지므로 재추출 횟수가 많아지는데 이런 현상은 대표적으로 강한 음의 상관관계가 존재할 때 나타난다. 반면에 시계열이 완만하고 변화의 속도가 느리면 재추출하는 횟수는 적어진다. 문제는 강한 음의 상관관계가 존재하는 자료에서 발생하는데, 많은 횟수의 재추출을 시행하면 주기간의 접합점(주기의 양 끝점)이 많아져서 접합점 주변의 자료에 상관관계가 무너지는 경우가 발생한다.

따라서 음의 상관관계가 존재하는 자료의 주기의 길이에 대한 의문점이 발생한다. 즉 상관관계는 증가하지만 평균주기의 길이는 줄어든다. 그런데 TB의 주목적은 원시 자료에 존재하는 상관관계를 재추출한 시계열에도 보존하는 것이므로 임계값을 교차하는 횟수가 많아지면 재추출 단위에 대한 조정이 필요하다.

Park(1997)은 재추출 단위를 하나 혹은 그 이상의 주기로 이루어진 토막(Chunk)으로 조정하였다. 재추출의 기본 단위인 Chunk 크기는 원시 자료의 크기와 자기상관 정도에 연관되어 있어 단순하게 결정할 수 없다. 최적의 Chunk 크기를 구하기 위한 해석적인 방법은 계속 연구되고 있으며 Park(1997)은 실험을 통하여 평균제곱오차(Mean Squared Error)를 최소로 하는 최적의 Chunk 크기를 구하는 경험적인 방법을 제시하였다.

3. Turing 검정

Melamed (1992)는 시계열 자료의 유사성을 검증하는 정성적인(qualitative) 방법으로 Turing 검정(Turing 1950)을 제안하였다. 시뮬레이션 모델에 의해 생성된 입력 데이터와 원시 데이터를 사람에게 보여주고, 두 데이터를 구별할 수 있는지 물어본다. 만약 대부분의 사람들이 두 데이터를 구별하지 못한다면 그 시뮬레이션 모델은 실제 프로세스를 표현하는데 성공한다고 말할 수 있다. Melamed는 이러한 정성적인 검증이 정량적인 검증과 함께 이루어져야 한다고 생각했다.

시뮬레이션 모델의 유효성 검증(model validation) 분야에서는 Carson (1986)과 Scruben(1980)이 Turing 검정을 사용하였다. 시뮬레이션 모델의 입력데이터와 출력데이터의 표본과 실측 데이터를 구분할 수 있는 지를 시스템에 대해서 잘 알고 있는 사람들에게 물어보고 시뮬레이션 모델의 유효성을 검증하였다.

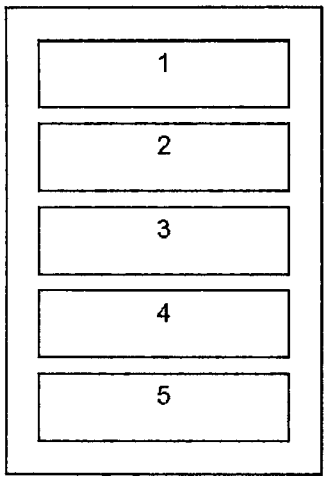
상기 기존 연구에서도 알 수 있듯이 Turing 검정을 이용한 붓트스트랩 표본 시계열의 시각적인 조사는 중요한 의미를 갖는다. 즉, 붓트스트랩 시계열 자체가 원시 시계열의 독립 반복 시계열 처럼 사용될 수 있는가를 평가하는 것이다.

3.1 Turing 검정에 사용된 데이터

실험에 사용된 시계열의 종류는 6개로 하나는 실세계의 실측 자료이고 나머지 5개는 앞서 사용했던 ARMA 모형의 자료이다. 실측자료는 일본 엔화의 미국달러에 대한 환율 데이터로서 시계열이 정상성을 갖도록 변환을 취하였다. 환율 데이터는 총6000개로 10개의 부분 시계열로 나누었다.

3.2 Turing 검정에 사용될 자료의 준비

총 6장의 설문지 중에서 각각의 한 장의 자료 구성은 다음과 같은 방법으로 이루어 졌고 [그림 2] 과 같은 구성을 하고 있다.



[그림 2] Turing 검정 설문지의 구성 형태

실험에 사용된 데이터의 길이는 600이다. 먼저 길이 600의 원시 독립 시계열 5개를 임의로 축출한다. 이중 4개의 시계열은 그대로 배치하고 나머지는 원시 시계열의 붓트스트랩 시계열로 대체하여 배치한다. 즉, 한 장의 설문지에는 4개의 독립 표본 시계열과 1개의 붓트스트랩 포본 시계열이 섞여서 배치된다. 실측자료의 경우는 원시 시계열의 길이가 6000개 이다. 이를 10개의 부분 시계열로 나누어 K-S 검정을 수행하여 최적의 Chunk 크기를 결정한다. 실험에 사용된 시계열들의 최적의 Chunk 크기는 [표 1] 과 같다.

시계열의 종류	최적의 Chunk 크기
실측 데이터	1
AR(1)+0.9	2
AR(1)-0.9	3
AR(5)	4
MA(4)	1
ARMA(1,1)	1

[표 1] 시계열들의 Chunk 크기

10개의 부분 시계열 중에서 임의로 5개를 선택하고, 이중에서 1개를 임의로 선택하여 선택된 시계열의 붓트스트랩 시계열을 배치하고 나머지 4개의 시계열은 그대로 배치한다. 실험에 사용된 붓트스트랩 표본의 위치와 시계열의 종류는 [표 2] 과 같다.

설문지 순서	종류	붓트스트랩 표본의 위치
1	실측 데이터	2
2	AR(1)+0.9	5
3	AR(1)-0.9	5
4	AR(5)	4
5	MA(4)	3
6	ARMA(1,1)	1

[표 2] 설문지에 사용된 자료의 순서와 붓트스트랩 표본의 위치

3.3 Turing 검정의 실시

실험 대상에 참여한 사람은 통계적인 지식이 있고 특히 시계열에 대한 지식을 가지고 있는 10명을 대상으로 선정하였다.
또한 가중 임계값 붓트스트랩에 대한 설명을 설문지에 실어 실험 전에 붓트스트랩에 대한 지식을 인식하고 있도록 하였다.

3.4 Turing 검정의 결과

Turing 검정의 결과는 [표 3] 과 같다.

대상	붓트스트랩 표본을 선택한 수
A	2
B	1
C	0

D	1
E	2
F	0
G	3
H	0
I	1
J	1

[표 3] Turing 검정의 결과

위에서 언급했듯이 실험은 10명을 대상으로 수행하였으며 한명 당 6장의 설문지가 나누어졌으므로 총 60개의 결과를 얻었다. 총 60개의 결과 중에서 11개의 붓트스트랩 표본을 찾아낼 수 있었다. 결과에 대한 통계적인 분석을 수행하기 위해서 가설검정을 수행하였다. 각 설문지에 5개의 시계열이 있고 이중 하나의 붓트스트랩 표본 시계열을 선택할 확률은 0.2 이다. 붓트스트랩 표본 시계열과 원시 시계열의 독립 표본 시계열들을 구분할 수 있는가? 에 대한 질문에 대해 다음과 같은 가설을 세웠다.

$$H_0 : p = 0.2$$

$$H_1 : p > 0.2$$

가설 검정 결과 $n = 60, p = 0.2, q = 0.8, X = 11$ 이라 하면 검정 통계량 Z는

$$Z = \frac{X - np}{\sqrt{npq}} = \frac{11-12}{\sqrt{9.6}} = -0.3227 < 1.645$$

을 만족하므로 유의수준 0.05 에서 귀무 가설을 기각하는데 실패했다. 결론적으로 붓트스트랩 시계열과 원시 시계열을 구별하기 어렵다는 결과를 얻었다.

4. 결론 및 추후 연구

본 논문에서는 과거의 단일 실측 자료가 있을 때 이와 유사한 다수의 자료를 생성하는 방법 및 그 방법의 타당성 평가기준을 연구하였다. 원시 시계열과 유사한 의사 시계열을 생성하기 위한 방법으로 임계값 붓트스트랩이라는 재추출 방법을 사용하였다. K-S 검정법과 같은 정량적인 평가 외에도 Turing 검정법을 적용하여 붓트스트랩 포본 시계열에 대한 시각적인 조사를 수행함으로써 정성적인 평가를 수행하였다. 전문가들도 원시 시계열과 붓트스트랩 표본 시계열을 구분하는 것을 어려워 함을 알 수 있었다. 결론적으로, 임계값 붓트스트랩의 Chunk 크기를 조절하여 붓트스트랩 표본 시계열을 생성하면 원시 시계열의 독립 반복 시계열로 사용될 수 있는 다수의 의사 시계열을 생성할 수 있음을 알 수 있었다.

본 논문에서 원시 시계열에 대한 정상성과 약한 자기 상관성을 가정 하였다. 따라서 비정상성을 가지거나 강한 자기 상관성을 가지는 시계열들에 대해 붓트스트랩 방법을 적용하기 위한 방법이 연구되어야 할 것이다.

또한 부분 시계열 기법을 사용하여 얻은 부분 시계열의 최적의 Chunk 크기를 알고 있을 때 원시 시계열의 최적의 Chunk 크기를 찾을 수 있는 이론적인 방법이 연구되어야 할 것이다.

임계값 붓트스트랩은 원시 시계열이 길지 않으면 주기의 개수가 부족하여 의사 시계열을 생성하는데 실패하므로 원시 시계열의 길이가 충분치 않을 때 적용할 수 있는 또 다른 임계값 붓트스트랩 방법을 개발하여야 할 것이다.

예를 들어 각각의 Chunk를 겹쳐서(overlapping) 인식하면 선택될 수 있는 Chunk 의 수가 늘어나므로 더 나은 다양성을 제공할 수 있을 것이다. 또한 임계값 붓트스트랩 방법의 재추출의 최소 단위가 연속되는 High Run 과 Low Run으로 구성되는 주기이지만 재추출 단위를 Run 단위로 축소 시킨다면 원시 시계열의 길이가 충분치 못한 경우에 더 나은 다양성을 제공할 것이다.

참고문헌

[1] A.C. Davison and D.V. Hinkley "Bootstrap Methods and their Application", Cambridge Series in Statistical and Probabilistic Mathematics

[2] BANKS, J., CARSON, J. and NELSON, B. L. (1996) "Discrete Event System Simulation", 2nd ed., Prentice Hall: Upper Saddle River, NJ.

[3] CHENG, R.C.H. (1995) "Bootstrap Methods in Computer Simulation Experiments"., In Proceedings of the 1995 Winter Simulation Conference, C. Alexopouslos, K. Kang,W. Lilegdon and D. Goldsman eds., IEEE Press: Piscataway, NJ, 171-177.

[4] KEDEM, B. (1993) "Time Series Analysis by Higher Order Crossings". IEEE Press: Piscataway, NJ.

[5] KIM, Y., HADDOCK, J., and WILLEMAIN, T. R. (1993a) "The Binary Bootstrap: Inference with Autocorrelated Binary Data", Communications in Statistics: Simulation and Computation. 22, 205-216.

[6] KIM, Y., WILLEMAIN, T. R., HADDOCK, J. and RUNGER, G. (1993b) "Simulation Output Analysis Using the Threshold Bootstrap",Technical Report no. 37-93-378, Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, New York.

[7] KIM, Y., WILLEMAIN, T. R., HADDOCK, J. and RUNGER, G. (1993c) "The Threshold Bootstrap: A New Approach to Simulation Output Analysis", In Proceedings of the 1993 Winter Simulation Conference, G. Evans, M. Mollaghasemi, E. Russell and W. Biles, eds. IEEE Press: Piscataway, NJ. 498-502.

[8] LEEMIS, L. (1995) "Input Modeling for Discrete Event Simulation", In Proceedings of the 1995 Winter Simulation Conference, C. Alexopoulos, K.Kang, W. Lilegdon and D. Goldsman, eds. IEEE Press: Piscataway, NJ. 16-23.

[9] PARK, D. (1997) "The Threshold Bootstrap For Time Series Analysis", Unpublished Ph.D. Dissertation. Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute.

[10] SEILA, A. F. (1992) "Advanced Output Analysis for Simulation", In Proceedings of the 1992 Winter Simulation Conference, J. J. Swain, D. Goldsman, R. C. Crain and J. R. Wilson, eds. IEEE Press: Piscataway, NJ. 190-197.