

## 임계값 부트스트랩을 사용한 시물레이션 입력 시나리오의 생성\*

김윤배\*\* · 김재범\*\*

### Generation of Simulation Input Stream using Threshold Bootstrap\*

Yun Bae Kim\*\* · Jae Bum Kim\*\*

#### Abstract

The bootstrap is a method of computational inference that simulates the creation of new data by resampling from a single data set. We propose a new job for the bootstrap: generating inputs from one historical trace using Threshold Bootstrap. In this regard, the most important quality of bootstrap samples is that they be functionally indistinguishable from independent samples of the same stochastic process.

We describe a quantitative measure of difference between two time series, and demonstrate the sensitivity of this measure for discriminating between two data generating processes. Utilizing this distance measure for the task of generating inputs, we show a way of tuning the bootstrap using a single observed trace.

This application of the threshold bootstrap will be a powerful tool for Monte Carlo simulation. Monte Carlo simulation analysis relies on built-in input generators. These generators make unrealistic assumptions about independence and marginal distributions. The alternative source of inputs, historical trace data, though realistic by definition, provides only a single input stream for simulation. One benefit of our method would be expanding the number of inputs achieving reality by driving system models with actual historical input series. Another benefit might be the automatic generation of lifelike scenarios for the field of finance.

Keyword : Threshold Bootstrap, Input Generator, Monte Carlo Simulation

논문접수일 : 2004년 7월 1일      논문게재확정일 : 2005년 2월 21일

\* 이 논문은 한국학술진흥재단의 연구지원으로 이루어짐.

\*\* 성균관대학교 시스템경영공학과

## 1. 서 론

시뮬레이션 관련 연구 중 입력체계(input process)에 관련 분야는 최근 활발하게 연구가 진행되고 있는데, 본 연구에서는 이산 사건 시뮬레이션의 입력체계에 사용될 수 있는 시나리오 생성 방법에 대한 연구 결과로서 부트스트랩을 이용한 새로운 입력 생성 방법을 제시한다.

본 연구의 입력 체계 관련 연구는 대상 시스템의 단일 실측 입력데이터(single historical input data)에서 유도되는 Trace driven 시뮬레이션에 중점을 둔다. 본 연구에서 개발한 방법을 사용하여 Trace-driven 시뮬레이션을 수행 할 때 데이터가 제한되어 있는 경우에도 다수의 입력 데이터를 생성할 수 있다.

또한 복잡한 의사결정 과정을 요하는 작업과정의 인력을 훈련시키는데 있어서도 실세계 그대로의 시나리오를 자동적으로 생성할 수 있다.

Trace-driven 시뮬레이션은 오직 하나의 과거 실측 데이터만으로 입력을 제공하기 때문에 시뮬레이션으로 분석하고자 하는 시스템에 다양한 입력을 제공하는데 한계가 있다[12]. 기존의 시뮬레이션 입력 데이터 생성에 관한 연구는 과거 데이터를 바탕으로 모수적인(parametric) 모형을 개발하는데 초점을 두고 있다. 그러나 입력이 매우 복잡한 형태를 가지면 모수적인 모형을 찾는 것이 불가능해지므로 부트스트랩을 사용한 비모수적인(non-parametric) 접근 방법으로 다양한 입력을 제공하는 방법이 요구된다. 예를 들어 재무 분야의 투자 분석을 위한 시뮬레이션 수행 시, 입력으로 제공되는 데이터들은 과거의 단일 실측 데이터며 상관관계(즉, 각 데이터값 간에 종속적인 영향을 미치게 되어 독립적이지 않은 상태)를 가지고 있어 대부분의 경우 모수적 모형을 찾는 것이 어렵다. 또 다른 예로, 인터넷 트래픽 모델의 시뮬레이션 수행 시, 입력으로 제공되는 단위 시간당 요구되는 웹 페이지의 수 같은 경우 데이터들 간에 종속관계가 매우 심하고 복잡하여 모수적 모형을 세우는데 어려움

이 있다.

현재까지 관련 연구에서는 입력체계를 매우 단 순화시킨 간단한 모수형 입력체계를 사용하여 그 제한적인 요소를 극복하고 있지만, 시뮬레이션 모델에 제공해야 될 입력데이터들이 복잡하고 종속적인 관계를 갖게 되면 단순화 과정은 실패하게 되므로 시뮬레이션 결과는 유효하지 못하다. 반면에 과거의 관측한 실측 데이터를 입력으로 사용하면 시뮬레이션 모델에 입력이 제한적으로 제공되고 시뮬레이션 출력 역시 제한되어 출력 분석이 어렵다. 이러한 문제는 시뮬레이션의 목적이 임의의 시스템의 운영자(Operator)를 훈련시키는 것일 때 더욱 심각해진다. 왜냐하면 단일 입력 데이터는 항상 같은 시뮬레이션 경로를 제공하므로 우수한 훈련 효과를 기대하기 어렵다. 따라서 비현실적인 다량의 입력 데이터를 사용할 것인가 혹은 현실적인 극히 적은 입력 데이터를 사용할 것인가에 대한 고민은 시뮬레이션을 사용하는 경영 과학도들이 겪는 고민이다. 바로 이점에서 우리는 부트스트랩을 사용해야 하는 당위성을 찾을 수 있다. 왜냐하면 부트스트랩은 현실적인 다수의 입력 데이터를 제공할 수 있기 때문이다.

부트스트랩은 원시의 데이터 군에서 재 추출한 유사 데이터 군으로 원시 데이터에 대한 추론을 계산하는 비모수적인 기법이다. 원래는 iid가 만족되는 데이터에 대한 분석 기법으로 출발하였으나 최근에는 시계열같이 종속적인 관계를 갖는 데이터 군에도 적용이 확장되고 있다[11, 18, 19]. 이러한 종속적인 데이터에의 적용은 주로 시뮬레이션 출력 분석에 적용되고 있고[8-10], 시뮬레이션 입력 데이터 생성에의 적용은 최근에야 발표되었다[16]. Demirel과 Willemain은 이동블록 부트스트랩(Moving Block Bootstrap, 이하 MBB)을 이용한 비모수적인 입력 데이터 생성 방법을 개발하였다. MBB는 재 추출 단위인 블록을 정하는데 어려움이 있어 많은 시행착오를 통해서 최적의 추출단위를 정한다. 하지만 임계값 부트스트랩은 임계값이 정해지면 곧 재추출 단위를 결정 할 수 있는 장점이 있어

사용하기가 매우 쉽다. 따라서, 본 논문에서는 부트스트랩 방법중의 하나인 임계값 부트스트랩(Threshold Bootstrap, 이하 TB)을 이용하여 시뮬레이션 입력 데이터 생성 방법을 개발하고자 한다. 여기서, 단일 실측 데이터로부터 부트스트랩 반복을 생성함에 있어서 그러한 부트스트랩 방법의 타당성을 평가하는 기준이 필요하다.

부트스트랩의 모든 통계적 분야[1, 3, 5, 7]에서는 기존의 직접 추론의 문제에 중점을 둔다. 즉, 반복으로 추출되어진 데이터 그 자체보다는 부트스트랩 반복으로부터 계산되어진 통계치의 특성을 설명하는 것으로서, 고전의 통계 분야에서처럼 부트스트랩 추정량의 특성 불편성, 일치성, 효율성, 점근적인 정상성에 주안점을 둔다. 가령, 부트스트랩 신뢰구간이 알맞은 범위 특성을 갖는다면, 그 간격에 기초한 부트스트랩 반복열은 적당하다고 판단될 것이다.

Demirel과 Willemain이 고안한 간접 추론의 방법은 부트스트랩 방법의 궁극적인 검사를 Turing 검정을 통하여 부트스트랩 반복과 원시 실측 데이터와 구분이 가능한지에 대한 정성적 평가에 초점을 두었다. 성공적인 부트스트랩 방법이라면 원시 데이터와 부트스트랩으로 구한 데이터 간의 구별은 어려워야 한다. 실제 필요로 하는 것은 다양성과 충실도의 정확한 배합이라고 할 수 있다. 즉 반복을 통한 다양성, 원시 데이터의 필수적인 요소로서 충실도가 제공되어야 한다. 부트스트랩 재추출을 통하여 추구하는 것은 원시 데이터와 그 원시 데이터의 독립 반복 사이에서 발견할 수 있는 동일한 정도의 유사성과 차이를 갖는 부트스트랩 반복을 구성하는 것이다. 가령, 독립 반복을 자매 데이터열(Sister series)로 생각할 수도 있는데, 자매들은 서로 닮았으나, 그들은 쌍둥이처럼 완전히 동일하거나, 먼 사촌사이처럼 완전히 다르지도 않다.

본 논문은 임계값 부트스트랩으로부터 생성된 부트스트랩 표본 시계열에 대한 정성적(Qualitative)인 검증을 하기 위하여 Turing 검정을 추가적으로 실시했다. 단일 실측데이터는 정상성과 약한

자기상관관계를 가지고 있는 자기회귀/이동평균 시계열 모형(ARMA) 및 실측데이터로서 달러대엔화의 환율을 사용하여 실험을 수행했다. 실험결과 원시 실측 데이터의 독립 시계열과 임계값 부트스트랩으로부터 얻은 부트스트랩 표본 시계열을 구분하기가 어렵다는 결론을 얻었다. 즉, 다수의 사실적인 시뮬레이션 입력 데이터를 제공하는 방법으로 임계값 부트스트랩이 좋은 대안임을 보였다.

본 논문은 총 4장으로 구성되어 있다. 2장에서는 기존 연구에 대해 고찰하였다. 본 연구에서 사용될 부트스트랩 방법에 대한 기본적인 이론과 상관관계가 존재하는 데이터에 대한 부트스트랩 방법인 이동 블록 부트스트랩과 임계값 부트스트랩 방법에 대해서 설명하였다. 3장에서는 임계값 부트스트랩 방법을 이용해서 부트스트랩 표본 시계열을 생성하는 방법을 설명하였고, 생성된 부트스트랩표본 시계열을 검증하는 방법으로 Turing 검정법에 대한 방법과 실험 결과를 설명하였다. 4장에서는 결론 및 추후 연구 과제를 기술하였다.

## 2. 관련 연구

### 2.1 상관관계가 존재하는 데이터를 위한 부트스트랩 방법

Efron[3]은 재추출 방법(Resampling)을 이용하여 통계적 추론을 수행할 수 있는 비 모수적인 접근 방법인 부트스트랩을 고안하였다. 부트스트랩 방법은 모수적인 추론이 어렵거나 해석적인 방법이 존재하지 않을 때 좋은 결과를 얻을 수 있다고 알려져 있다[4]. Singh[21]은 상관관계가 존재하는 데이터는 고전적인 부트스트랩 방법을 적용하는데 문제점이 있음을 지적하였는데 이는 고전적인 부트스트랩 방법에서의 재추출 단위가 하나의 데이터이기 때문에 데이터들간의 상관관계가 무시될 수 있다고 했다. 따라서 데이터들의 상관관계를 유지하기 위해서 부트스트랩의 재추출 단위를 조정하는 방법이 고안되었는데 본 장에서는 대표적인

두 가지 방법인 이동블록 부트스트랩과 임계값 부트스트랩을 소개한다.

### 2.1.1 이동블록 부트스트랩(Moving Block Bootstrap)

이동블록 부트스트랩은 상관관계가 존재하는 데이터를 위한 비모수적인 부트스트랩 방법으로 Künsch[11], Liu and Singh[13]에 의해 독립적으로 제안되었다. 정상성을 갖는 시계열의 표본  $\{X_1, X_2, \dots, X_n\}$ 이 있을 때 부트스트랩 재추출 단위인 이동블록  $B_i$ (크기가  $b$ 인  $i$ 번째 이동블록)는 아래와 같이 정의된다.

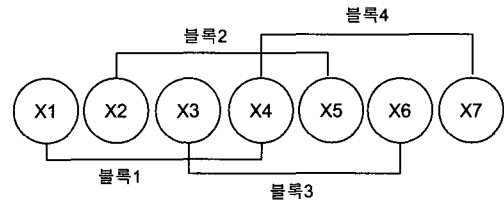
$$B_i = \{X_i, X_{i+1}, \dots, X_{i+b-1}\},$$

$$i = 1, 2, \dots, n - b + 1$$

즉,  $i$ 번째 이동블록  $B_i$ 는  $X_i$ 로 부터  $b$ 개의 연속적인 데이터들로 구성된다. 원시 시계열이 존재할 때,  $\{B_1, B_2, \dots, B_{n-b+1}\}$ 의 집합에서 복원을 허락하며 무작위로 재추출(Random Resampling)을 수행함으로써 다수의 새로운 부트스트랩 표본을 만들고 통계적인 추론을 수행할 수 있다. <그림 1>에 이동블록 부트스트랩의 재추출 방법을 나타내었다.

MBB에서 원시 시계열을 크기  $b$ 의 블록으로 묶는 이유는 블록 안에 있는 데이터들 간의 상관관계를 유지하고 그 상관관계가 부트스트랩 표본에도 반영될 수 있도록 하기 위해서 이다. MBB에서의 중요한 가정은 원시 시계열 내의 임의의 데이터와 거리  $b$ 만큼 떨어진 데이터는 상관관계가 존재하지 않는다는 것이다. MBB에서 블록의 크기를 1로 했을 때 고전적인 부트스트랩과 일치하게 된다. MBB에서의 가장 중요한 문제는 데이터들 간의 거리  $p$ 에서의 자기상관구조가 무시될 수 있도록 블록의 크기  $b$ 를  $p$ 보다 크게 잡아야 하는 동시에 블록의 크기  $b$ 보다 작은 거리에서의 데이터들 간의 자기상관구조를 유지하여야 한다는 것이다. 블록의 크기에 따라 이동블록 부트스트랩의 성능이 많은 영

향을 받게 되지만, 아직까지 최적의 블록크기를 결정하는 것에 대한 이론적인 방법은 존재하지 않으며 데이터들의 상관관계와 데이터의 길이에 따라 경험적으로 결정되어야 한다는 문제가 발생하게 된다. Hall[6]은 경험적으로 MBB의 최적의 블록크기를 결정하는 방법을 제시하였다.



<그림 1> 이동블록 부트스트랩의 재추출 구조 및 방법

### 2.1.2 임계값 부트스트랩

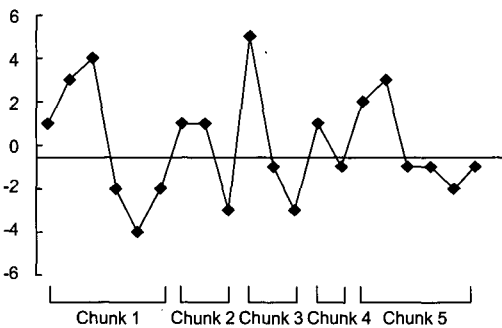
#### (Threshold Bootstrap)

임계값 부트스트랩(TB)은 시뮬레이션 출력 분석을 위한 비모수적인 추론 방법으로 Kim(1993b)에 의해 고안되었다. Kim(1993b)은 상관관계가 존재하는 데이터를 해석하기 위하여 고전적인 부트스트랩의 재추출 단위인 하나의 데이터를 데이터 내에 존재하는 종속관계를 유지하도록 'Chunk'로 바꾸었다. 시계열은 임계값을 시계열을 판독하는 수준(표본평균이나 표본의 중앙값 등)으로 정하면 임계값 보다 높은 데이터의 연속인 High Run과 낮은 데이터의 연속인 Low Run으로 구분되어진다. 재추출의 기본 단위를 중복되지 않고 연속되는 두 개의 Run(High Run과 Low Run의 결합)으로 정하고, 이를 'Chunk'(cycle)라고 정의한다. 임계값이 정해지면, 관측치를 통해 Chunk가 자체적으로 결정되어, 재추출의 단위가 자동으로 정해지는 것이 임계값 부트스트랩의 장점이다. 또한, 임계값 부트스트랩은 일반적인 시뮬레이션 출력 분석에 사용되는 다수의 반복 시뮬레이션 실행 결과를 부트스트랩 표본으로 대치할 수 있어서 시뮬레이션 실행 시간을 단축할 수 있고, 동류의 단일실행 분석기법인 배치평균방법(Batch Mean)보다 사용하기 쉽다

는 장점이 있다.

본 논문에서는 임계값 부트스트랩을 사용하여 원시 데이터의 부트스트랩 표본을 생성하고 이를 시뮬레이션 입력 데이터로 사용할 것이다.

확률 체계  $P$ 에 의해 생성된 정상 시계열  $\{X_1, X_2, \dots, X_n\}$ 이 약한 종속 구조를 갖고, 임의의 임계값이  $R$ 개의 Chunk를 만든다고 가정하자.



Chunk 1 = {1, 3, 4, -2, -4, -2}, Chunk 2 = {1, 1, -3},  
 Chunk 3 = {5, -1, -3}  
 Chunk 4 = {1, -1}, Chunk 5 = {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}  
 $X_1^* = -0.1$   
 TB표본 2 : {2, 3, -1, -1, -2, -1, 5, -1, -3, 1, -1, 1, 1, -3, 5, -1, -3, 5, -1, -3}  
 $X_2^* = 0.1$   
 TB표본 3 : {1, 3, 4, -2, -4, -2, 2, 3, -1, -1, -2, -1, 1, -1, 5, -1, -3, 1, 1, -3}  
 $X_3^* = 0.0$   
 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 2> 임계값 부트스트랩의 재추출 방법

$$C_i = \{X_{i,1}, X_{i,2}, \dots, X_{i,n_i}\} \text{를 크기가 } n_i \text{인 } i \text{번}$$

째 Chunk라면,  $i = 1, 2, \dots, R$ 이고  $\sum_{i=1}^R n_i = n$ 이다.

Chunk의 크기  $n_i$ 와 Chunk의 수  $R$ 은 둘 다 확률 변수이다. 이 두 개의 확률 변수는 데이터의 자기상관과 임계값에 매우 강한 종속관계를 갖는다.

TB는 Chunk들의 집합  $\{C_1, C_2, \dots, C_R\}$ 에서 확률  $1/R$ 로 무작위로 추출하여 부트스트랩 표본을 만든다. 미지의  $P$ 는 재추출 구조에 근거해  $P^*$ 로 추정할 수 있다. 예로, 시계열  $\{X_1, X_2, \dots, X_n\}$ 의 표본평균  $\bar{X}_n$ 의 편의(bias)와 표준오차(standard error)를 추정하려 할 때,  $\{X_1^*, X_2^*, \dots, X_n^*\}$ 가  $P^*$ 에서 얻은 부트스트랩 표본이라면,  $P^* \rightarrow \{X_1^*, X_2^*, \dots, X_n^*\}$ 라고 표현하고 이때 표본평균  $\bar{X}_n^*$ 의 편의와 표준오차에 대한 부트스트랩 추정치는 재추출 확률구조  $P^*$ 에 의거한  $\bar{X}_n$ 의 편의와 표준오차와 일치하게 된다.

TB를 사용하여 표본평균의 편의와 표준오차를 추정하는 방법의 이해를 돕기 위하여 <그림 2>를 보였다.

TB의 재추출 알고리즘을 정리해보면 다음과 같다.

- Step 1 :  $n$ 개로 이루어진 자기상관이 존재하는 데이터를 획득한다.
- Step 2 : 임계값을 설정한다(예를 들면, 표본평균이나 표본의 중간값 등).
- Step 3 : 데이터를 임계값을 중심으로 두 개의 데이터군으로 분류한다. 데이터군은 임계값보다 높은 데이터의 연속인 High Run과 낮은 데이터의 연속인 Low Run으로 구분되어 진다. 중복되지 않고 연속되는 두 개의 Run(High Run과 Low Run의 결합)이 하나의 Chunk가 된다.
- Step 4 :  $R$ 개의 Chunk에 추출될 확률을 각각  $1/R$ 로 할당하고, 균등분포(0, 1)를 따르는 난수를 발생시켜 재추출할 Chunk를 선택한다. 복원을 허락하여 재추출한 Chunk들을 연결하여 부트스트랩 표본을 원시 데이터의 크기  $n$ 만큼 생성한다.
- Step 5 : 재 생성된 부트스트랩 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 데이터가 임계값 수준을 교차하는 횟수와 매우 밀접한 관계가 있다. 또한 강한 양의 상관관계를 가진 원시 데이터에서는 Chunk의 수

자가 적고 Chunk의 길이가 긴 반면 음의 상관관계의 데이터에서는 Chunk의 숫자가 많고 그 길이가 짧다. 경험적으로 AR(1) 시계열의 평균 Chunk의 길이가 1차 자기회귀계수와 관계를  $-0.9$  에서  $+0.9$  로 변화시켜 보면 평균 Chunk의 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 데이터에서 발생하는데, 많은 횟수의 재추출을 시행하면 Chunk 간의 접합점(Chunk의 양 끝점)이 많아져서 접합점 주변의 데이터에 상관관계가 무너지는 경우가 발생한다. 따라서 음의 상관관계가 존재하는 데이터의 Chunk의 길이에 대한 의문점이 발생한다. 즉 상관관계는 증가하지만 평균 Chunk의 길이는 줄어든다. 그런데 TB의 주목적은 원시 데이터에 존재하는 상관관계를 재추출한 시계열에도 보존하는 것이므로 임계값을 교차하는 횟수가 많아지면 재추출 단위에 대한 조정이 필요하다. Park[17]은 재추출 단위를 하나 혹은 그 이상의 Chunk로 이루어진 것으로 조정하였다. 재추출의 기본 단위인 Chunk 크기는 원시 데이터의 크기와 자기상관 정도에 연관되어 있어 단순히 결정할 수 없다. 최적의 Chunk 크기를 구하기 위한 해석적인 방법은 계속 연구되고 있으며, Park[17]은 실험을 통하여 평균제곱오차(Mean Squared Error)를 최소로 하는 최적의 Chunk 크기를 구하는 경험적인 방법을 제시하였다.

Park[17]의 방법을 적용하여 최적의 Chunk 크기를 실험을 통하여 발견하였는데, 시계열 AR(1)  $\phi=0.9$ , MA(4) ( $\theta = -1.0, -1.0, -1.0, -1.0$ ), ARMA(1, 1) ( $\phi=0.9, \theta=-0.9$ )의 경우는 최적의 Chunk 크기가 1로서 시계열의 길이가 증가하여도 변하지 않는 반면, 시계열 AR(1) ( $\phi=-0.9$ ), AR(5) ( $\theta=0.4, 0.2, 0.1, 0.1, 0.1$ )는 길이가 증가하면서 최적의 Chunk 크기가 증가하는 것을 보였다.

그러나 본 연구의 결과는 시계열의 종류와는 상관 없이 시계열의 길이가 증가할 때 일정한 비율로 최적의 Chunk 크기가 증가함을 알 수 있다. 이는 통계량의 분포를 이용하여 시계열의 유사성을 판단하는 방법이 충실도와 다양성을 모두 평가하기 때문이다. 즉, 시계열의 길이가 증가함에 따라 재추출되는 Chunk의 크기를 늘림으로서 부트스트랩 표본시계열이 길이에 맞는 다양성을 유지할 수 있는 것이다.

### 3. 부트스트랩 표본 시계열의 검정

Melamed[14]는 시계열 데이터의 유사성을 검증하는 정성적인(qualitative) 방법으로 Turing 검정[22]을 제안하였다. 시뮬레이션 모델에 의해 생성된 입력 데이터와 원시 데이터를 평가자들에게 보여주고, 두 데이터를 구별할 수 있는지 물어본다. 만약 대부분의 평가자들이 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 유효성검증(Model Validation)분야에서는 Carson[2]과 Schruben[20]이 Turing 검정을 사용하였다. 시뮬레이션 모델의 입력데이터와 출력데이터의 표본과 실측 데이터를 구분할 수 있는 지를 시스템에 대해서 잘 알고 있는 평가자들에게 물어보고 시뮬레이션 모델의 유효성을 검증하였다.

기존의 임계값 부트스트랩은 재추출된 데이터로부터 계산된 통계치를 활용하는 통계적인 추론 방법이었다. 부트스트랩 시계열 자체가 원시 시계열의 독립 반복 시계열처럼 사용될 수 있는가를 평가하는 것이다. 따라서 Turing 검정을 이용한 부트스트랩 표본 시계열의 시각적인 조사는 중요한 의미를 갖는다. 이번 장에서는 Turing 검정을 사용하여 “원시 시계열의 독립 반복 시계열들과 원시 시계열의 부트스트랩 시계열을 구별할 수 있는가?”라는 중요한 질문에 답하고자 한다.

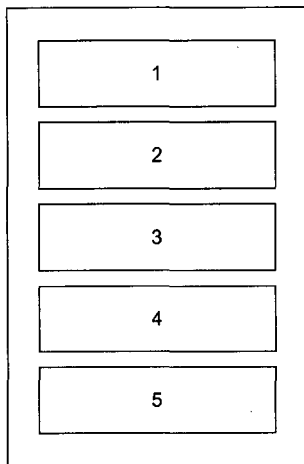
### 3.1 Turing 검정

#### 3.1.1 Turing 검정에 사용된 데이터

실험에 사용된 시계열의 종류는 6개로 하나는 실세계의 실측 데이터고 나머지 5개는 앞서 사용했던 ARMA 모형의 데이터다. 실측데이터는 일본 엔화의 미국달러에 대한 환율 데이터로서 시계열이 정상성을 갖도록 변환을 취하였다. 본 논문의 엔화의 미국달러 환율 자료는 <http://pacific.commerce.ubc.ca/xr/data.html>에서 1980년 10월부터 2002년 11월까지의 데이터를 사용하였다. 환율 데이터는 총6000개로 10개의 부분 시계열로 나누었다

#### 3.1.2 Turing 검정에 사용될 데이터의 준비

총 6장의 설문지 중에서 각각의 한 장의 데이터 구성은 다음과 같은 방법으로 이루어 졌고 <그림 2>와 같은 구성을 하고 있다.



<그림 2> Turing 검정 설문지의 구성 형태

실험에 사용된 데이터의 길이는 600이다. 먼저 길이 600의 원시 독립 시계열 5개를 임의로 추출한다. 이중 4개의 시계열은 그대로 배치하고 나머지는 원시 시계열의 부트스트랩 시계열로 대체하여 배치한다. 즉, 한 장의 설문지에는 4개의 독립 표본 시계열과 1개의 부트스트랩 표본 시계열이 섞여서

배치된다. 실측데이터의 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.

<표 1> 시계열들의 Chunk 크기

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

10개의 부분 시계열 중에서 임의로 5개를 선택하고, 이중에서 1개를 임의로 선택하여 선택된 시계열의 부트스트랩 시계열을 배치하고 나머지 4개의 시계열은 그대로 배치한다. 실험에 사용된 부트스트랩 표본의 위치와 시계열의 종류는 <표 2>와 같다. 부록에 설문지(1)~설문지(6)을 첨부하였다.

<표 2> 설문지에 사용된 데이터의 순서와 부트스트랩 표본의 위치

설문지 순서	종 류	부트스트랩 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

#### 3.1.3 Turing 검정의 실시 및 결과

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

Turing 검정의 결과는 <표 3>과 같다.

<표 3> Turing 검정의 결과

대상	부트스트랩 표본을 선택한 수
A	2
B	1
C	0
D	1
E	2
F	0
G	3
H	0
I	1
J	1

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

$$H_0: p = 0.2$$

$$H_1: p > 0.2$$

가설 검정 결과  $n=60$ ,  $p=0.2$ ,  $q=0.8$ ,  $X=11$ 이라면 검정 통계량  $Z$ 는

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

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

## 4. 결 론

본 논문에서는 과거의 단일 실측 데이터가 복잡

하고 종속적이어서 모수적 모형을 찾는데 실패할 때 이와 유사한 다수의 데이터를 생성하는 방법을 연구하였다. 이렇게 생성된 데이터들은 이산 사건 시뮬레이션의 입력으로 사용될 수 있다.

원시 시계열과 유사한 의사 시계열을 생성하기 위한 방법으로 임계값 부트스트랩이라는 재추출 방법을 사용하였다. 임계값 부트스트랩 방법과 부분 시계열 기법을 사용하여 다수의 부트스트랩 표본 시계열을 생성하는 알고리즘을 정리하였으며 부분 시계열의 최적의 Chunk 크기를 원시 시계열의 최적의 Chunk 크기에 적용하기 위해서 실험을 통한 경험적인 방법을 제시 하였다. 추가적으로 Turing 검정법을 적용하여 부트스트랩 표본 시계열에 대한 시각적인 조사를 수행함으로써 정성적인 평가를 수행하였다. 전문가들도 원시 시계열과 부트스트랩 표본 시계열을 구분하는 것이 어려웠다. 결론적으로, 임계값 부트스트랩의 Chunk 크기를 조절하여 부트스트랩 표본 시계열을 생성하면 원시 시계열의 독립 반복 시계열로 사용될 수 있는 다수의 의사 시계열을 생성할 수 있음을 알 수 있었다.

본 연구에서는 원시 시계열에 대한 정상성과 약한 자기 상관성을 가정하였다. 따라서 비정상성을 가지거나 강한 자기 상관성을 가지는 시계열들에 대해 부트스트랩 방법을 적용하기 위한 방법이 추후에 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 부트스트랩은 원시 시계열이 길지 않으면 Chunk의 개수가 부족하여 의사 시계열을 생성하는데 실패하므로 원시 시계열의 길이가 충분치 않을 때 적용할 수 있는 또 다른 임계값 부트스트랩 방법을 개발하여야 할 것이다. 예를 들어 각각의 Chunk를 겹쳐서(overlapping) 인식하면 선택될 수 있는 Chunk의 수가 늘어나므로 더 나은 다양성을 제공할 수 있을 것이다. 또한 임계값 부트스트



랩 방법의 재추출의 최소 단위가 연속되는 High Run과 Low Run으로 구성되는 Chunk이지만 재추출 단위를 Run 단위로 축소시킨다면 원시 시계열의 길이가 충분치 못한 경우에 더 나은 다양성을 제공할 것이다.

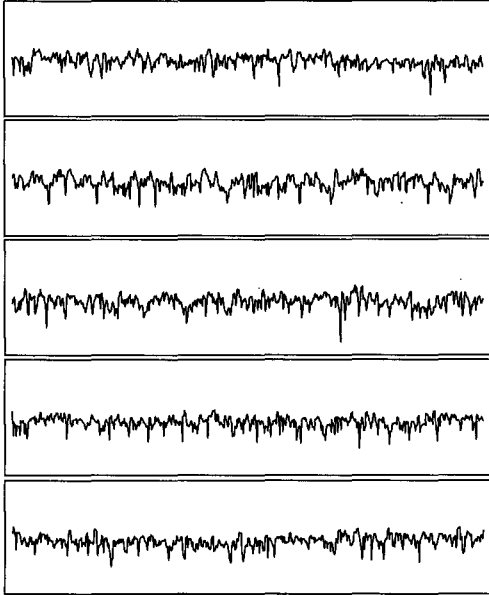
## 참 고 문 헌

- [1] Bickel, P. and D. Freedman, "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, Vol.9(1981), pp.1196-1217.
- [2] Carson, J.S., "Convincing User's of Model's Validity is Challenging Aspect of Modeler's Job," *Industrial Engineering*, Vol.18(1986), pp.74-85.
- [3] Efron, B., "Bootstrap Methods : Another Look at the Jackknife," *Annals of Statistics*, Vol.7(1979), pp.1-26.
- [4] Efron, B. and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall Inc., New York, 1993.
- [5] Gotze, F. and H.R. Künsch, *Blockwise Bootstrap for Dependent Observation : Higher Order Approximations for Studentized Statistics*, Technical Report, Univ. Bielefeld, Germany, 1993.
- [6] Hall, P., J.L. Horowitz and B. Jing, "On Blocking Rules for the Bootstrap with Dependent Data," *Biometrika*, Vol.82(1995), pp.561-74.
- [7] Hall, P. and B. Jing, "On Sample Reuse Methods for Dependent Data," *Journal of the Royal Statistical Society B*, Vol.58(1996), pp.727-737.
- [8] Kim, Y., J. Haddock and T.R. Willemain, "The Binary Bootstrap : Inference with Autocorrelated Binary Data," *Communications in Statistics : Simulation and Computation*, Vol.22(1993a), pp.205-216.
- [9] Kim, Y., D.S. Park, K.I. Shin and Willemain, "Simulation Output Analysis Using the Threshold Bootstrap," *European Journal of Operational Research*, Vol.134(2001), pp.17-28.
- [10] Kim, Y., T.R. Willemain, J. Haddock and G. Runger, "The Threshold Bootstrap : A New Approach to Simulation Output Analysis," *In Proceedings of the 1993 Winter Simulation Conference*, G. Evans, M. Molgahsemi, E. Russell and W. Biles, eds. IEEE Press : Piscataway, NJ. (1993c), pp. 498-502.
- [11] Künsch, H., "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, Vol.17(1989), pp.1217-1241.
- [12] Leemis, L., "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. (1995) pp.16-23.
- [13] Liu, R. and K. Singh, "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," *In Exploring the Limits of Bootstrap*, R. LePage and L. Billard, eds. Wiley : New York, 1992.
- [14] Melamed, B., J.R. Hill and D. Goldsman, "The TES Methodology : Modeling Empirical Stationary Time Series," *In Proceedings of the 1992 Winter Simulation Conference*, J.J. Swain, D. Goldsman, R.C. Crain, and J.R. Wilson, eds. Piscataway, NJ : IEEE Press, (1992), pp.16-23.
- [15] Omer, F.D. and T.R. Willemain, "Generation of simulation input scenario using

- bootstrap methods," *Journal of Operational Research Society*, Vol.53(2002), pp.69-78.
- [16] Omer, F.D. and T.R. Willemain, "Applications of the Bootstrap in Scenario Generation," unpublished Ph.D dissertation, Rensselaer Polytechnic Institute, 2001.
- [17] Park, D., "The Threshold Bootstrap For Time Series Analysis," Unpublished Ph.D. Dissertation, Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, 1997.
- [18] Park, D. and T.R. Willemain, "The Threshold Bootstrap and Threshold Jackknife," *Computational Statistics and Data Analysis*, 1999.
- [19] Politis, D.N. and J.P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, Vol.89(1994), pp.1303-1313.
- [20] Schruben, L.W., "Establishing the Credibility of Simulations," *Simulation*, Vol.34 (1980), pp.101-1.
- [21] Singh, K., "On The Asymptotic Accuracy of Efron's Bootstrap," *Amm. Statist.*, Vol.9 (1981), pp.1187-1195.
- [22] Turing, A.M., "Computing Machinery and Intelligence," *Mind*, Vol.59(1950), pp.433-460.

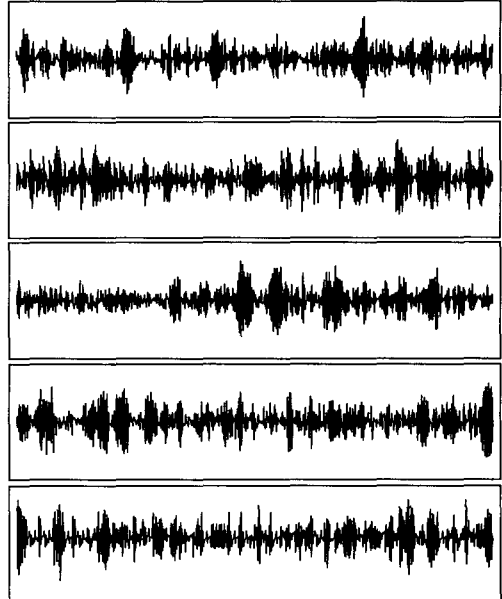
부 록

설문지(1)



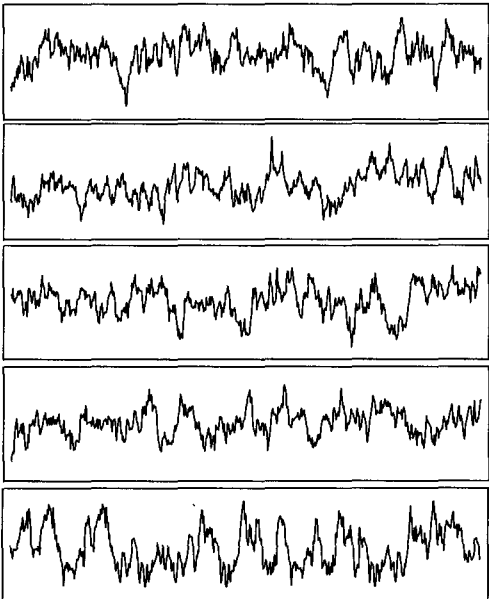
Data series : 10-day log rolling variance(\$/₩), 부트스트랩 데이터는 상위 2번째

설문지(3)



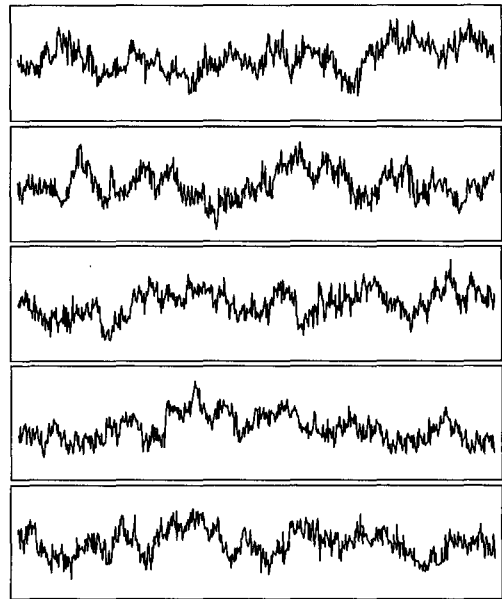
Data series : AR(1),  $\phi = -0.9$  부트스트랩 데이터는 상위 5번째

설문지(2)



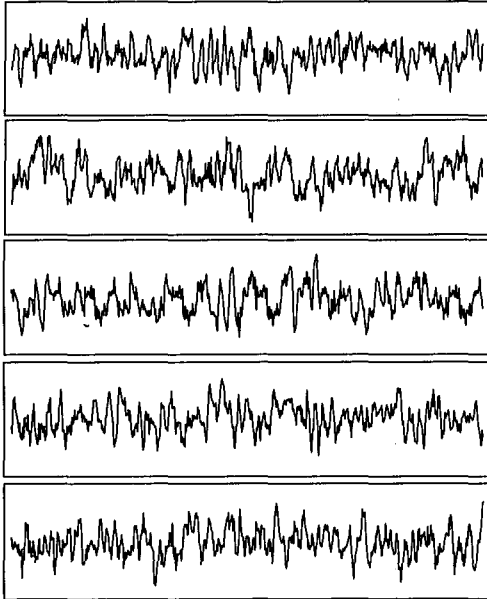
Data series : AR(1),  $\phi = 0.9$ , 부트스트랩 데이터는 상위 5번째

설문지(4)



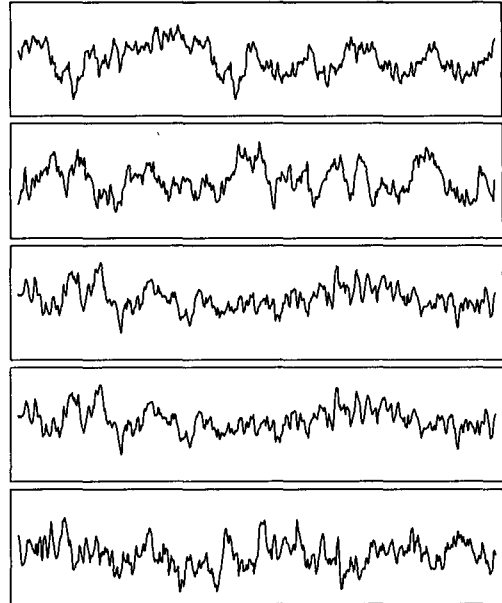
Data series : AR(5) ( $\phi = 0.4, 0.2, 0.1, 0.1, 0.1$ ), 부트스트랩 데이터는 4번째

설문지(5)



Data series : MA(4) ( $\theta = -1.0, -1.0, -1.0, -1.0$ ),  
부트스트랩 데이터는 3번째

설문지(6)



Data series : ARMA(1, 1) ( $\phi = 0.9, \theta = -0.9$ ),  
부트스트랩 데이터는 1번째