시계열 데이터의 성격과 예측 모델의 예측력에 관한 연구

이원하*·최종욱**

Relationships Between the Characteristics of the Business Data Set and Forecasting Accuracy of Prediction models

Won Ha Lee*·Jong Uk Choi**

Abstract

Recently, many researchers have been involved in finding deterministic equations which can accurately predict future event, based on chaotic theory, or fractal theory. The theory says that some events which seem very random but internally deterministic can be accurately predicted by fractal equations. In contrast to the conventional methods, such as AR model, MA, model, or ARIMA model, the fractal equation attempts to discover a deterministic order inherent in time series data set.

In discovering deterministic order, researchers have found that neural networks are much more effective than the conventional statistical models. Even though prediction accuracy of the network can be different depending on the topological structure and modification of the algorithms, many researchers asserted that the neural network systems outperforms other systems, because of non-linear behaviour of the network models, mechanisms of massive parallel processing, generalization capability based on adaptive learning.

However, recent survey shows that prediction accuracy of the forecasting models can be determined by the model structure and data structures. In the experiments based on actual economic data sets, it was found that the prediction accuracy of the neural network model is similar to the performance level of the conventional forecasting model. Especially, for the data set which is deterministically chaotic, the AR model, a conventional statistical model, was not significantly different

* 크낙정보통신
** 상명대

from the MLP model, a neural network model. This result shows that the forecasting model appropriate to a prediction task should be selected based on characteristics of the time series data set. Analysis of the characteristics of the data set was performed by fractal analysis, measurement of Hurst index, and measurement of Lyapunov exponents.

As a conclusion, a significant difference was not found in forecasting future events for the time series data which is deterministically chaotic, between a conventional forecasting model and a typical neural network model.

1. 시계열 데이터의 성격과 예측력

신경망의 연구가 활성화됨에 따라서 시계열 데이터의 예측에서는 비선형 모델에 기초하고 있는 신경망의 예측력이 기존의 선형 모델들보다 우수하다는 논문들이 많이 나타나고 있다 [지원철, 1995; Hsu, Hus, and Tenorio, 1993; Tang and Fishwick, 1991; Jhee and Lee, 1993; Connor and Atlas, 1991; Castillo and Melin, 1995; Tenti, 1995; Jang and Lai, 1993; Tyree and Long, 1995; Bowen, 1991]. 신경망 모델의 우월성 주장은 지금까지 개발된 시계열 예측 및 진단모델들이 전통적인 통계기법, 즉, 단순 회귀 또는 다중 회귀 모델, 연립방정식 및 다변량 분석 모델, 상태 공간 모델 등에 근거하기 때문에 모델의 비적응성, 단순화를 위한 과도한 가정조건, 처리의 복잡성 및 예측의 비정확성 등이 생길 수밖에 없다는 점에 근거한다. 신경망은 적응적 학습기능, 대규모 병렬처리, 그리고 함수의 근사화 및 일반화를 통한 비선형 문제의 해결을 장점으로 들 수 있다.

그러나 이러한 기존의 연구들은 다음과 같은 점에서 일반화에 문제가 있다. 우선, 기존의 연구들이 소수의 시계열 데이터를 대상으로 기존의 선형 모델과 신경망 모델의 예측력을 비교한 것이기 때문에 이들 데이터를 대상으로 행한 실험을 바탕으로 신경망 모델이 기존의 모델들보다 예측력이 높다는 주장은 통계적으로 설득력이 약하다. 각 실험의 일반화를 위해서는 통계적으로 의미있는 개수의 Dataset을 대상으로 테스트가 이루어져야한다. 본 연구에서는 미국 St.Louis 연방은행에서 취합한 65개의 데이터를 대상으로 실험을 하였다.

다음으로는 신경망의 경우, 예측의 정확도는 신경망 구조(Backpropagation, Recurrent, FIR, TDNN 등), 내부구조(Hidden Layer의 수, Hidden Node의 수), 활성화 함수(sigmoid, tanh) 등에 따라 예측력이 달라진다는 문제점을 안고 있다. 신경망의 구조뿐만 아니라 시계열 데이터의 성격에 따라서도 예측력이 달라지므로 이들 예측 모델의 예측력을 일반화할 수 있는 연구, 특히 데이터의 성격과 예측력과의 관계를 규명하는 연구가 필요하다. 본연구에서는 이러한 한계점을 테스트 Data Set의 수를 늘림으로서 극복하려고 하였다. 테스트 Dataset이 많아짐으로서 각 모델 적용에서 발생할 수 있는 오차들을 평균화할 수 있을 것으로 생각한다.

이전의 연구[Yun, Nam-Kung, Shin, Roh, and Choi, 1997; 심기창, 최종욱, 정윤, 1995]에서는 데이터의 성격에 따라 예측 모델의 정확성이 달라진다는 점이 발견되었다. 본연구에서는 데이

터의 성격에 따라서 예측력이 달라진다는 가설하에 데이터의 성격에 따라 데이터를 분류하고 이를 기존의 선형 모델과 신경망 모델을 사용하여 예측한 후, 예측력의 차이를 검정하였다. 본 연구의 궁극적인 목표는 일반적인 주기성과 계절성, 순환성를 가지고 있지 않기 때문에 비교적 예측이 어려운 임의적인(Random) 변화를 갖는 시계열 데이터의 성격을 여러 가지 요소로 측정하고, 이러한 성격에 따라서 각 모델의 예측정도가 달라질 것이라는 가정하에 데이터의 특성에 따른 예측력이 우수한 모델을 제시하려고 하는 것이다.

시계열 데이터의 성격을 규명하기 위해서는 전통적으로 연검정(run-test), 전환점 검정(turning points test)과 같이 단순한 방법이나 자기 상관계수(Autocorrelation coefficient), 회귀분석등 여러 가지 방법이 사용되어 왔으나 비선형 동태성을 가지는 시계열의 경우에는 상관차원 추정후 원시계열에 대한 BDS검정과 [백웅기, 1996] Shuffling BDS검정을 반복적으로 수행하는 검정[남재우, 1994]등을 사용하고 있다. 본연구에서는 실험 대상 시계열 데이터들이 비선형 동태성을 가지고 있다고 가정하고, 최근 이러한 데이터의 성격 규명에 많이 사용되고 있는 R/S 분석과 Hurst계수의 측정, data의 손실정도를 의미하는 Lyapunov계수, 신경망 모델의 입력노드 수를 정하는 Correlation Dimension 등을 데이터의 특징 분석에 사용하였다.

다음으로는 데이터의 성격 규명에서 나타난 특징을 사용하여 기존의 선형 모델과 최근 개발된 신경망 모델을 사용, 각 모델들의 예측 정확도를 측정하였다. 이는 최근 신경망 예측 모델이 기존의 선형 모델보다 높은 예측 정확도를 가진다는 주장을 검정하기 위한 것으로 65개의 경제 시계열 데이터에 대해 테스트하였다. 이러한 연구의 궁극적인 목표는 여러 혼돈현상을 규명하는 측정요소에 따라 각 모델의 예측정도가 달라질 것으로 가정하고, 학습되지 않은 자료에 대해서도 신뢰성을 보장할 수 있는 시스템 개발을 위하여 데이터의 특성과 각 모델들의 예측력을 비교하여 예측력이 우수한 통합모델을 제시하고자 한다. 본 연구에서는 통계모델중 이동평균모델(Moving Average: MA)과 자기회귀(AutoRegressive: AR)모델을 이용하였으며, 신경망 모델로는 다층 퍼셉트론즈(Multi-Layered Perceptrons: MLP)와 회귀 다층 퍼셉트론즈(Recurrent Multi-Layered Perceptrons: RMLP)를 이용하였다.

본연구에서는 65개의 시계열 데이터를 대상으로 2개의 전통적 모델(AR, MA)과 2개의 신경망 모델(MLP, RMLP)을 사용하여 예측력을 테스트하였던 바, 데이터의 성격에 따라 각 모델의 예측력에는 차이가 나타났으나 신경망 모델이 기존 모델들 보다 예측력이 우수한다는 가설은 기각되었다.

2. 데이터 성격 분석 방법

시계열 분석 방법은 시계열 구성 요소의 특성에 따라 선택하게 된다. 분석방법의 선택은 일반적으로 자료의 형태, 분석의 용이성, 분석자료 해석의 이해 정도에 따라서 선택되어진다. 전통적인 분석방법으로는 회귀분석 방법, 박스-젠킨스(Box-Jenkins)방법, 지수평활법(exponential smoothing), 시계열 분해방법 등 네 가지로 나눌 수 있다[허명회, 박유성, 1994]. 수학적 이론을 바탕으로 한 회귀분석 방법과 박스-젠킨스 방법은 체계적인 반면, 분해 방법과 지수 평활법은

경험적이고 직관적인 방법이라 할 수 있다. 일반적으로 말해, 박스-젠킨스 방법은 어떠한 형태의 시계열 자료에도 이용할 수 있으나, 특히 시계열 구성요소가 시간의 흐름에 따라 매우 빠르게 변동되는 경우에 효과적이다. 회귀분석 방법은 시계열의 구성요소가 시간에 의존하지 않는 상수효과를 가지고 있을 때 적합한 방법론이다. 한편, 시계열 자료의 분해방법과 지수 평활법은 시계열의 구성요소가 시간의 흐름에 따라 느리게 변동할 때 가장 효과적인 예측방법이고, 특히 시계열을 각 구성요소로 분해할 수 있다는 장점을 가지고 있다.

본 연구에서 대상으로 삼고있는 데이터의 경우, 기존의 선형성 예측 모델에서 변동의 속도와 변동의 주기성으로는 데이터 특성을 표현하기 힘들기 때문에 일반적으로 프랙탈 분석에서 사용하는 지수들을 사용하였다. 본연구에서는 Hurst 계수와 Lyapunov 계수, Correlation Dimension을 데이터의 특성을 나타내는 측정지수로 사용하였다. Hurst 계수는 시계열 데이터의 임의성(Randomness)를 측정하기 위해서 주로 사용되는 지표이며 Lyapunov계수는 데이터의 영향력 손실 정도를 측정하는데 쓰여진다. Correlation Dimension은 시계열 데이터의 움직임을 결정하는 독립변수의 수를 측정하는데 쓰여진다.

2-1. R/S 분석과 Hurst Exponent

경제 시계열 데이터의 경우에는 일반적으로 임의보행과정(random walk)을 따른다고 이야기되고 있다. 이런 임의보행과정의 시작은 아인쉬타인(Einstein)의 1908년 브라운 운동에대한 기술로부터 시작되었다.

$$R = T^{0.50}$$

where, R = distanceecovered, T = a timeindex - (식 1)

만약 n개의 연속적인 값들을 가지는 벡터라고 가정하고, 정규화 과정(Z-score)을 수행하면

$Z_r = (X_r - X_m)$ *where,* $r=1, \cdots, n$ - (식 2)

으로 변환된다. 누적된 시계열(cumulative time series) Y 를 정의하면,

$Y_l = (Z_l + Z_r)$ *where,* $r=2, \cdots, n$ - (식 3)

이다. 여기서 가장 마지막 누적 시계열은 Z의 값이 항상 0이다. 여기서 시계열의 부합범위(adjusted range) R_n을 정의할 수 있다[Peters, 1994]. 즉, R_n은 시계열 지수(time index) n인 시스템의 이동범위를 의미한다. 만약 n=T라면, 초기 브라운 운동을 고려할 때, n의 증가에 따라 시계열 벡터가 독립적이라고 말할 수 있다.

$R_n = Max(Y_1, \cdots, Y_n) - \min(Y_1, \cdots, Y_n)$

where, $R_n > 0$ - (식 4)

이러한 개념을 브라운 운동이 아닌 시계열에 적용하는 경우 (식 6)은 아래와 같이 일반화 된다.

$(R/S)_n = C \cdot n^H$

where, H is Hurst exponent - (식 5)

R/S is Rescaled range

재구성된 범위는 평균값이 0이며, 국부적 표준편차를 의미하고 양변에 log를 취하면 다음과 같이 된다.

$\log(R/S_n) = H\log(n) + \log(c)$ - (식 6)

위에서 얻어진 허스트 계수(H)는 상관 정도(correlation measure)에 영향을 미친다[Peters, 1991]. 상관정도는 다음과 같이 정의되어 진다.

$C = 2^{(2H-1)} - 1$

where, C = correlation measure - (식 7)

H = Hurst exponent

그리고 허스트 계수는 다음과 같이 3가지로 분류되어 질 수 있다.

1) $H = 0.5$

무작위 보행(random walk)을 한다고 말한다. 즉, 발생되는 사건은 (식 7)에서 보면 C = 0으로 아무런 상관관계도 가지지 않는다. 다시 말하면 현재의 추세는 미래에 아무런 영향을 미치지 못하는 것을 의미한다. 맨델로프에 의하면 H=0.5인 경우에는 이차원 프랙탈을 가진다고 말한다. 이는 프랙탈 차원이 허스트 지수의 역수와 같음을 의미한다.

2) $0 \leq H < 0.5$

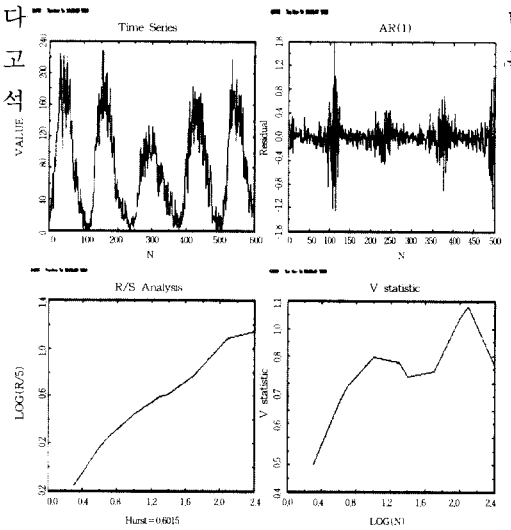
비지속성(antipersistent)을 가진다고 말한다. 만약 시스템이 현재의 주기에서 증가하는 경향을 가진다면 다음 주기에서는 반대로 감소하는 영향을 가진다. 이러한 비지속성은 H가 0에 가까워질수록 강하게 나타나며 C는 -0.5인 음의 상관관계(negative correlation)를 가지게 된다.

3) $0.5 < H \leq 1$

지속성(persistent) 또는 강제적 추이(trend-reinforcing)를 가진다고 말한다. 만약 현재주기가 증가(감소)하고 있다면 다음 주기에서도 증가(감소)하는 경향을 가진다. H가 1에 수렴하면 C는 2에 수렴하게 된다. 만약 마지막 움직임이 양의방향이고 H=0.6이라면, 다음 움직임이 증가할 확률이 60%임을 의미한다.

카오스계로부터 생성된 시계열의 허스트지수 H는 N값이 작은 경우 데이터 자체가 추세성을 가질 수밖에 없으므로 H>0.5이다. 그러나 N이 커지면 데이터간의 무질서도가 증가하여 H가 0.5로 수렴되어 가는 과정을 볼 수 있다. 따라서 N이 증가함에 따라 H값이 변하는 시점의 N은 기억효과의 지속성을 의미한다고 볼 수 있다.

다 다
고 분
석



[그림 1] 혹점 생성 개수에 따른 R/S 분석 및 Hurst 지수

2-2. 프랙탈 차원(fractal dimension)

아날로그 신호를 $f(t)$라고 한다면[1] 우리는 t를 적정한 간격 Δt로 샘플링하여 이산신호를 얻을 수 있고, Δt를 충분히 작게 하면 원하는 신호를 복원할 수 있다. 시간구간 [a, b]를 등분해서 샘플링한 신호 f를 N점의 샘플값 계열에서 순서로 된 N개의 수치가 짝으로 표시되는 양을 N차원 벡터라고 한다.

$$f(f_1, f_2, \cdots, f_n) \quad \text{- (식 8)}$$

이라고 쓰면 f는 N차원 벡터(vector)로 표현되게 된다.[2]

(x_1, y_1)부터 (x_2, y_2)까지의 직선을 생각해 보면, 두 점간의 거리는 피타고라스의 정리에 의해 (식 14)과 같이 나타낸다. 그러나 곡선에 대해서는 곡선의 부분을 조각(segment)으로 나뉘어 각

1) 일반적으로 시계열 데이터는 일정한 시간간격 t를 가지고 샘플링한 함수라고 볼 수 있다.
2) 순서로 된 N개의 수치가 짝으로 표시되는 양을 N차원 백터라고 한다.

점들간의 직선의 거리를 구하는 방법(polygonal approximation)을 적용하여 구하는 방법밖에 없다.

$$\sqrt{(x_1-x_2)^2 + (y_1-y_2)^2} \quad \text{-(식 9)}$$

다음의 그림은 polygonal approximation의 방법을 나타낸 것이다.

[그림 2] Polygonal approximation of a curve

그러나 [그림 2]과 같은 방법으로는 프랙탈 구조를 가지는 Koch 눈송이의 길이를 측정하는 것은 거의 불가능하다.

이는 프랙탈 구조와 같이 아주 미세하게 변하는 시계열의 경우에는 독립변수로써는 측정할 수 없는 차원이 존재할 수 있음을 의미하는 것과 마찬가지이다. 위상 차원(topological dimension)과 마찬가지로 프랙탈 차원도 주어진 체적 $B(x_0, r)$의 주어진 집합 S를 둘러쌀 수 있는 C로부터 시작한다. 직경이 r인 체적요소를 둘러싸기 위한 최소단위를 ϵ을 만족하기 위해서는 $r < \epsilon$이어야 한다. 만약 S가 고정되어 있고 체적요소를 둘러싸기 위한 최소의 값을 C라 하면 최소의 체적요소의 개수는 $N(\epsilon)$이어야 한다. 프랙탈 차원 D_f는

$$D_f = \lim_{\epsilon \to \infty} \frac{\ln N(\epsilon)}{\ln 1/\epsilon}$$

과 같이 정의되어 진다[Ott et. al., 1994]. 프랙탈 차원이 존재한다는 것은 끌개를 구성할 수 있다는 것을 의미하며, 시계열 예측의 가능성과 예측범위를 판단할 수 있다[백웅기, 1996].

2-3. 리아프노프(Lyapunov) 지수

혼돈 시스템의 특징중 하나는 "초기조건의 민감성"이다. 비선형 동태 방정식을 가지는 시스템처럼 랜덤한 요소를 가지는 경우에는 어떤 시점에서 발생한 충격이 유한한 시간이 지난 후에 영향력을 모두 잃어버리게 된다. 즉, 시스템의 운동 방향은 충격에 따라 어떠한 방향으로 진행되고, 소멸되어 진다. 이런 시스템의 확장 또는 수축을 의미한는 것이 리아프노프 지수(Lyapunov exponent)이다. 그러나 일반적인 소산(dissipative)시스템에서는 추세성 및 변동성을 보는 방법이 단순히 Fourier Spectrum을 보거나 시간에 따른 출력값을 본다고 해서 나타나지 않는다. 그러므로 위상공간(Phase space)에서의 고찰이 일반적이다.[3] 또한, 그 속에서 Fractal 구조를 발견할 수 있다면 리아프노프값을 측정할 수 있다[이주장, 1993].

k-dimensional벡터 x_n을 가정하여 보면 (식 9)과 같이 나타낼 수 있다.

$$x_{n+1} = G(x_n)$$

여기서 원래의 오비트(orbit)와 멀어진 오비트를 고려하면, δx_n을 무한벡터라 가정할 때 $x_{n+1} \rightarrow x_n + \delta x_n$로 나타내어지며

$$\delta_{i+1} = DG(x_n)\delta x_n$$

Where, $DG(x) = k * k$ *Jacobian Matrix of partial derivative of* $G(x)$ -(식 10)

라 할 수 있다. 여기서

$$y_n = \frac{\delta x_n}{|\delta x_0|}$$

$$y_{n+1} = DG(x_n)y_n \quad \text{-(식 11)}$$

3) 시계열 자료를 위상공간에 재구성하는 이유는 시계열 자체가 하나의 독립변수를 나타내는 관계로 프랙탈차원을 측정하는 경우 1에서 2사이의 값을 가지게 된다. 그러나 위상공간에서 재구성하는 경우에는 모든 영향을 미치는 요소(독립변수)를 고려할 수 있다는 것이 Rulle에 의해 증명되었다.

이라고 정의할 수 있으며, y_n은 탄젠트 벡터 또는 탄젠트 스페이스라고 부른다. 위 수식에서 보듯이 y_n은 초기조건 $\{y_n\}$과 초기의 탄젠트벡터 y_0에 의존하여 변화된다. 여기서 주목할 것은 y 값이 매회 반복됨에 따라 탄젠트 공간이 확장 또는 수축되는지 알아보기 위하여 다음과 같이 정의하고

$$h(x_0, y_0) = \lim_{\epsilon \to \infty} h(x_0, y_0, n) \quad \text{-(식 12)}$$

$$h(x_0, y_0, n) = \frac{1}{n} \ln |y_n| \quad \text{-(식 13)}$$

$h(x_0, y_0)$를 리아프노프 지수라 하며, $h(x_0, y_0, n)$을 유한시간(Finite time) 리아프노프 계수라고 한다[Ott et. al., 1994].

리아프노프 지수는 계의 자유도(Degree of Freedom)만큼 존재한다. 자유도 2를 갖는 계에 대해서 주 점의 시간경로는 서로 분리되거나 합해진다. 만약 분리점들이 항상 서로 접근하는 방향으로만 움직인다면 Lyapunov지수는 음의 값을 가지게 되며 양의 값은 서로 멀어지는 발산의 형태를 가진다. 이런 Lyapunov 지수의 특성에 의해 계는 축소 또는 확장되어지고 흡인집합에서 주름진 형태나 이상한 형태의 끌개를 구성하게 되는 것이다.

2-4. 상관 차원(correlation dimension)

시계열 자료에 있어서 가장 좋은 예측모델은 운동방정식을 알고 있는 경우일 것이다. 그러나 비선형 동태 방정식의 형태를 가지는 시스템의 경우에는 운동방정식으로 나타내는 것에 한계가 있을 뿐만 아니라 원하는 방정식을 찾아내기도 어렵다. 일반적으로 예측모델의 구성에 있어서 가장 어려운 부분중 하나가 독립변수의 개수를 지정하는 일이다. ARMA나 ARIMA와 같은 고전 통계모델의 경우에는 자기 상관 계수(Auto Correlation Factor : ACF) 등을 이용하여 모델의 p값과 q값을 찾아내고 있으나, 비선형 동태 방정식의 경우에는 이러한 해를 찾기가 힘들다. 이런 경우에는 카오스 검정에서 가장 중요한 단계로 취급하는 매립과정과 프랙탈 차원, 상관차원을 이용하여 해결한다.

시계열 $\{a_t\}$로 m-차원에서 매립화 시키는 과정은 $(a_t, a_{t+1}, \cdots, a_{t+m-1})$과 같이 단순히 관측치 m개를 중첩시킴으로써 완료된다. 매립과정을 통해 우리는 시계열 자료가 결정론적 설명력을 가진다고 말할 수 있으며[Takens, 1996], 원래의 상태벡터 $\{x_t\}$ 대신 $\{a_t^m\}$을 살펴봄으로써 저차원 카오스적 끌개가 있는지를 검정할 수 있는 근거를 제시하였다.

Grassberger와 Proccacia에 의해 제시되어진 상관차원은 위상공간에서 시계열을 재구성하는 방법을 이용하여, 프랙탈 차원의 단점을 극복하였다. 상관 차원은 매립차원(embedding dimension)을 2부터 증가 시키며, 프랙탈 구조의 직경(diameter)을 증가시켜 측정한다.

$$C_m(R) = \left(\frac{1}{N^2}\right) * \sum_{i,j=1}^{N} Z(R - |x_i - x_j|) \quad \text{-(식 14)}$$

Where, $Z(x) = 1$ *if* $R - |x_i - x_j| > 0$ *:0 otherwise*

N = *Number of observations*

R = *distance*

C_m = *correlation integral for dimension m*

여기서 상관차원은 t와 t+1점간에서 프랙탈 직경 안에 두 점이 있을 확률을 의미한다.

3. 예측 모델과 데이터 분석

본연구에서는 시계열 데이터의 예측을 위해서는 4가지 모델이 사용되었다. 고전통계모델로

는 AR모델과 MA모델이 사용하였으며 신경망 모델로는 MLP와 RMLP를 사용하였다. AR모델의 p와 신경망 모델의 입력노드를 결정하기 위하여 카오스 검정법의 하나인 상관차원(Coorelation dimension)을 사용하였으며 MLP 신경망 모델의 은닉노드(Hidden Node)의 수는 입력노드의 두 배로 결정하였으며, RMLP 모델의 경우 Context node의 수는 두 개로 고정하여 측정하였다. 일반적으로 신경망모델이 구간예측에 탁월한 성능을 보인다고 하고 있으나, 본 연구에서는 점예측만을 하였다. 이는 시계열 데이터에 내재되어 있을 수 있는 카오스성질 자체가 오랜 시간 지속되지 못한다는 점에 기인하고 있다.

3-1. 자료의 설명

본 연구에 사용된 자료는 미 St. Louis 연방은행에서 취합되어진 자료로써, 미 정부의 경제지표 조사 데이터로써 사용되며, 아무런 제한없이 복사 또는 무료로 사용할 수 있다. 이 자료들은 FRED BBS라는 공공 DB를 통하여 다운로드(download)받은 것이다.

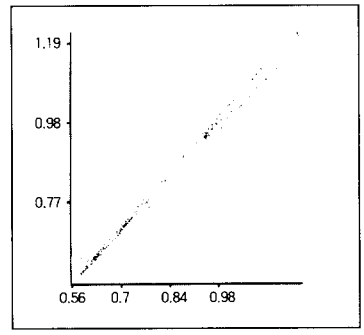
연구에 사용된 자료는 크게 계절성(Seasonality)이 있는 것과 없는 것으로 구분되어 있으며, 주식, 예금, 고용 및 기타자료로 구성되어 있다. 예측에 사용된 데이터 패턴에 관한 특징은 다음과 같다.

〈표 1〉 예측에 사용된 패턴의 특성

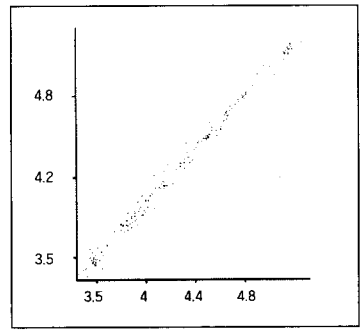
종 류	계절성(Seasonality)			총 Pattern수
	있음	없음	모름	
주 식	3	5	0	8
예 금	8	18	0	26
고 용	0	0	20	20
기 타	1	4	6	11
총 Pattern수	12	27	26	65

3-2. 자료의 성격 분석

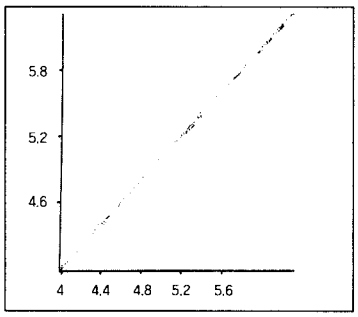
본 연구에서는 경제 시계열 패턴의 형태가 카오스의 성질을 가지고 있는지를 알아보는 과정으로 먼저 흡인집합을 구성하였다. 흡인집합을 구성한 결과 65개 시계열 자료에 대해 일정한 영역을 벗어나지 않는 범위내에서 끌개(attractor)를 구성하고 있음을 볼 수 있었다. 다음은 구성한 끌개중 각 예측모델이 예측한 값의 오차가 최소인 패턴들의 흡인집합(absorption set)이다.



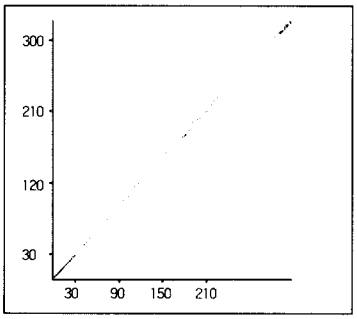
[그림 3] AR모델이 가장 잘 맞춘 흡인집합 (Usmine)



[그림 4] MA모델이 가장 잘 맞춘 흡인집합 (Uscons)



[그림 5] MLP모델이 가장 잘맞춘 흡인집합(Uswtrade)

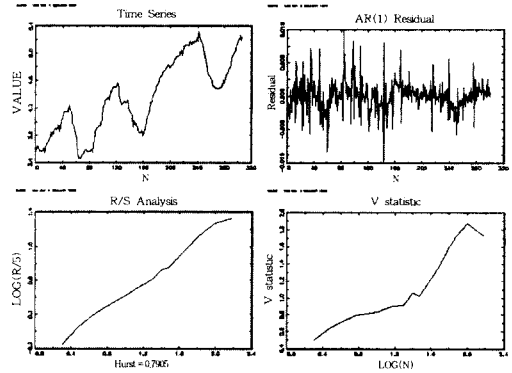


[그림 6] RMLP모델이 가장 잘맞춘 흡인집합(Ira)

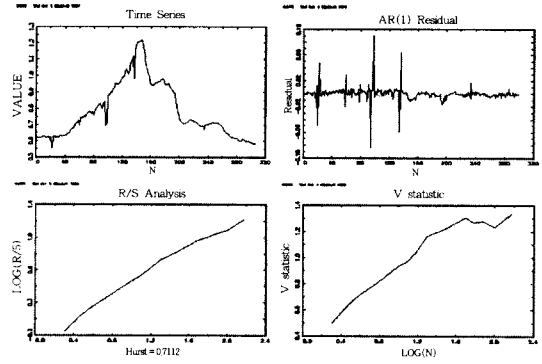
위의 그림에서처럼 변동이 많은 자료라 할지라도 경제 데이터들이 강한 상관관계를 가지며 끌개를 구성하는 것을 볼 수 있다.

카오스검정중 일반적으로 많이 사용하는 것은 리아프노프(Lyapunov) 지수를 측정하는 것이라 할 수 있다. 리아프노프(Lyapunov) 지수는 데이터 패턴 전부분에 관해 측정하는 Large 리아프노프(Lyapunov)계수와 국부적으로 측정하는 Local Lyapunov계수가 있으나, 본 연구에서는 전자의 검정만을 실시하였다.

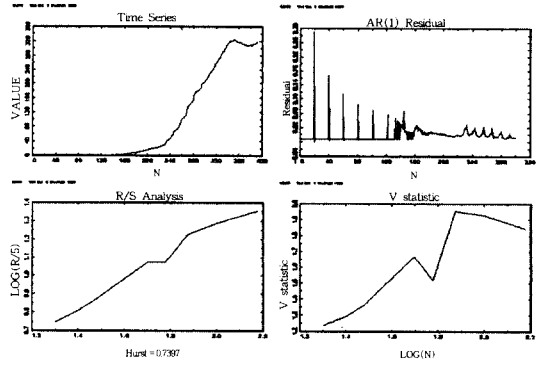
또한 데이터의 추세성과 변동성을 측정할 수 있는 Hurst계수는 잔차분석(Residual Analysis)후 측정하였으며, [그림 7]부터 [그림 10]까지는 Hurst계수를 측정한 결과를 나타내는 그림으로, 각 예측모델이 예측한 값의 오차가 최소인 패턴들의 측정결과이다.



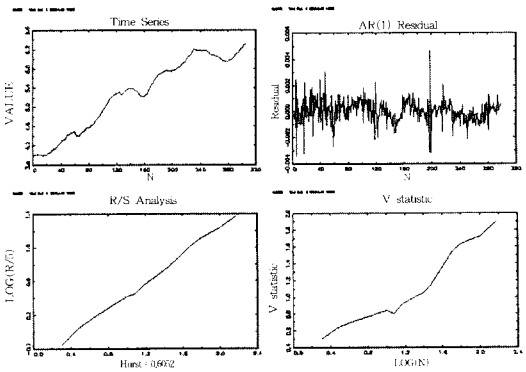
[그림 7] MA모델의 특성을 가장 잘 반영한 패턴의 Hurst계수 측정결과(Uscons)



[그림 8] AR모델의 특성을 가장 잘 반영한 패턴의 Hurst계수 측정결과(Usmine)



[그림 9] RMLP모델의 특성을 가장 잘 반영한 패턴의 Hurst계수 측정결과(Ira)



[그림 10] MLP모델의 특성을 가장 잘 반영한 패턴의 Hurst계수 측정결과(Uswtrade)

위의 그림들에서 볼 수 있듯이 Hurst계수 측정결과 자료들은 비교적 양의 지속성(persistent)을 가지고 있는 것으로 보여지고 있으며, 일부 자료의 경우에는 음의 비지속성(anti-persistent)를 가지고 있는경우도 있었다. 그러나 대부분의 자료가 고전통계에서 말하는 무작위보행(RandomWalk), 즉 Hurst계수 0.5를 가지는 값은 없는 것으로 측정결과 나타났다.

그리고 신경망의 입력노드의 수와 AR모델의 계수 p를 결정하기 위한 방법으로는 매립차원(Embedding Dimension)을 10으로 둔 상태에서 상관차원을 측정하였으며, 대부분의 자료들이 3 또는 4의 차원을 가지는 것으로 나타났다. 즉, 주어진 패턴들은 일반적으로 3개 또는 4개의 외부변수들에 의해 모형화되어진다고 말할 수 있다.

본 연구에서는 데이터특성에 따른 예측력을 비교하기 위하여 AR과 MA 고전통계모델을 SAS를 이용하여 모델링하였으며, 신경망 모델로는 MLP와 RMLP를 일반적으로 많이 사용되고 있는 NNDT라는 프로그램을 이용하여 구현하였다. 신경망의 은닉층(Hidden Layer)의 수는 한 개로 지정하였으며, 은닉층 노드의 수는 입력노드의 수에 두배만큼으로 지정하였다. 또한, 점예측을 위하여 출력노드는 하나로 고정하고 구간예측은 하지 않는 것으로 하였다. RMLP모델에서도 역시 은닉층을 하나 두었으며 MLP구조와 동일한 형태를 취하는 것으로 하였다. 그리고 부가적으로 출력층에 회귀노드를 하나더 포함시켜 입력노드로 시간지연을 주게 하였으며, 총 회귀되어지는 노드의 수는 두 개로 하였다. 이는 입력노드의수에 절반을 고려한 것이다.

4. 예측 결과 분석

65개 미국 경제시계열 데이터를 이용하여 카오스검정과 모델별(AR, MA, MLP, RMLP) 예측력을 조사하여 보았으며 예측력 측정 결과 데이터 패턴에 따른 비선형적 특성과 모델특성이 다음과 같이 밝혀졌다.

1) AR모델과 MLP모델에 있어서 예측력의 차이는 크게 나지 않는다.

본 연구에서는 미 경제 시계열 데이터에 대해 AR모델과 MLP모델이 보인 예측오차를 가지고 MLP모델의 예측력이 AR보다 좋은지를 검증하기위해 두 개의 짝지어진 표본(paired sample)에 관한 paired-t test를 수행하였다. 다음은 이에대한 귀무가설과 대립가설이다.

귀무가설(H0) : 두 모델에 관한 차이(difference[4])가 명확이 있다. 즉,

$H_0 : \delta > 0$

4) 여기서 말하는 차이는 하나의 패턴에 대해 두 개의 예측모델이 가지는 RMS를 가지고 AR과 MLP의 차를 구한것이다.

이고, 대립가설은

대립가설(H1) : 두 모델에 관한 차이가 없다.

즉,

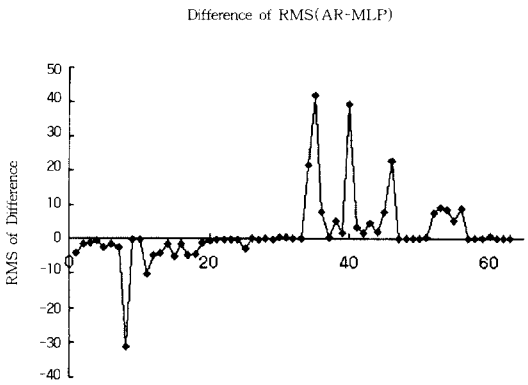
$H_0 : \delta = 0$

으로 주어진다. 주어진 가설에 대한 검정을 수행하면

$E(Di) = E(AR_i - MLP_i) = \delta$

$$\overline{D} = \sum_{i,\ j=1}^{n} \frac{D_i}{n} = 1.857627,$$

$$SD = \sqrt{\sum_{i,\ j=1}^{n} \frac{(D_i - \overline{D})^2}{(n-1)}} = 9.706112$$



[그림 11] Paired t-test를 위한 AR모델과 MLP모델의 차

귀무가설(H0) : $\delta > 0$

유의수준 0.05인 기각역 : $|T| \le t(n-1, \alpha)$

$= t(63, 0.05)$

이므로, 여기서 검정 통계량 관측값은

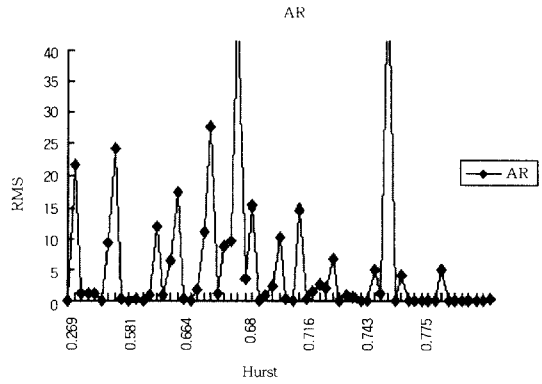
$$t = \frac{\overline{D} - \mu}{\frac{S}{n}} = \frac{1/857627 - 0}{\frac{9.706112}{\sqrt{63}}} = 1.51909$$

과 같이 주어진다. 그런데 $t(63, 0.05) = 1.670$이므로

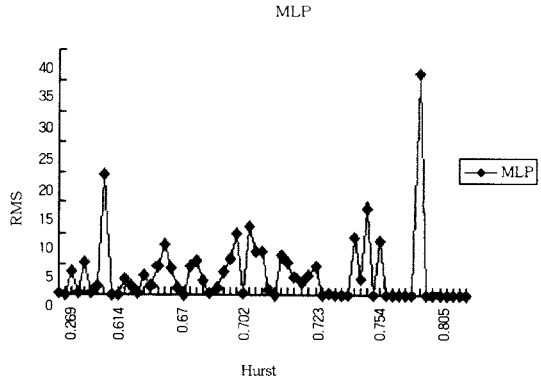
$|T| \le t(62, 0.05)$

이다. 그러므로 귀무가설(H0)를 기각한다. 즉, 데이터 패턴에 따른 AR 모델과 MLP모델의 예측력의 차이는 없다. 즉, 기존의 신경망 모델이 고전통계모델보다 좋다고 이야기한것에 대해 AR모델과 MLP를 비교한 결과 그렇지 않음이 나타나고 있다.

2) Hurst계수가 증가함에 따라 모델들의 예측능력은 향상되었다.



[그림 12] Hurst 계수에 따른 AR모델의 RMS

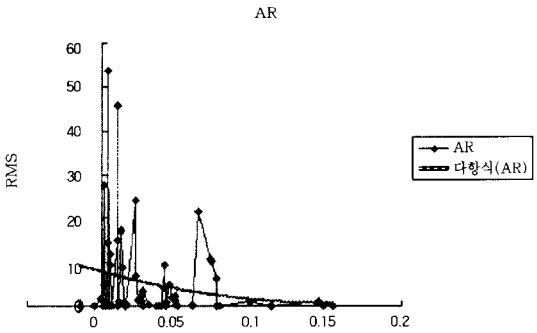


[그림 13] Hurst 계수에 따른 MLP모델의 RMS

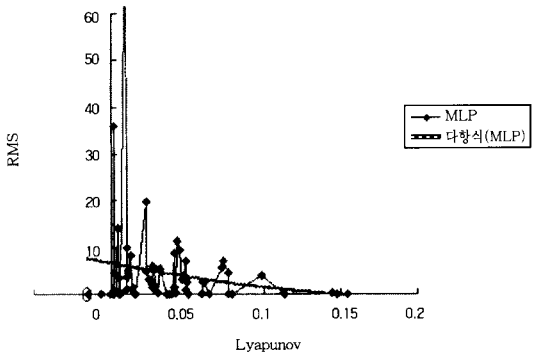
[그림 12]에서는 AR모델의 경우 Hurst 계수 값이 증가하면서 안정된 추세를 보이나, 편차가 심한 것을 볼 수 있다. 반면 [그림 13]에서는 MLP

모델의 경우 Hurst 계수값이 증가하여도 편차는 적은 것을 볼 수 있다. 또한 Hurst계수가 증가하면서 AR모델과 MLP모델이 모두 RMS가 적어지며 예측모델이 안정화되는 것을 볼 수 있다. Hurst계수는 데이터의 추세성과 변동성을 나타낸다. 즉, 시계열 자체의 무작위정도를 의미하며, 계수가 0.5에 가까워질수록 고전통계에서 말하는 무작위보행(Random Walk)을 따르게 된다. 본 실험예측결과 이러한 무작위보행을 갖는 경향이 적어질수록 예측오차가 줄어드는 것이 발견되었다.

3) 리아프노프 지수가 클수록 일반적으로 MLP모델의 예측능력이 좋다.



[그림 14] Lyapunov값에 따른 AR모델의 RMS



[그림 15] Lyapunov값에 따른 MLP모델의 RMS

[그림 14]와 [그림 15]에서 보면 MLP모델과 AR모델의 경우, 리아프노프(Lyapunov)계수의 값이 증가함에 따라 RMS가 수렴하는 것을 볼 수 있다. 그러나 MLP모델의 경우에는 값이 증가하면서 상당히 안정적으로 수렴하는 것이 보이나, AR의 경우에는 그렇지 않은 것을 도식적으로 볼 수 있다. 리아프노프 지수는 현재의 상태에 충격을 주었을 때 데이터 패턴에 영향력을 미치는 것을 의미한다. 즉, 신뢰도가 중요시 되는 예측모델의 경우에는 수렴속도보다 안정성에 신경을 두어야 하고, 이런 경우 리아프노프(Lyapunov)계수는 예측모델을 지정하는 하나의 척도로 사용할 수 있을 것이다. 그러나 그림에서 보듯이 리아프노프(Lyapunov)계수값이 작은 경우에는 두 모델이 모두 불안정한 상태를 나타내고 있어 이에관한 연구가 더 필요할 것이다.

5. 결 론

본 연구는 적응성있는 예측 모델을 위한 사전적 연구로써 총 65개의 미국 경제 시계열을 이용하여 카오스 검정을 통한 패턴 조사와 4가지(MA, AR, MLP, RMLP) 예측 모델을 이용한 예측능력 비교를 해보았다.

〈표 2〉 모델별 예측 성능 비교

	AR	MA	MLP	RMLP
BEST	30	1	33	1

시계열 데이터에 아무런 전처리도 거치지 않은 상태로 학습한 결과 AR 모델과 MLP모델의 예측능력이 다른 두 모델에 비해 좋은 것으로 나타났으며, 예측력(forecasting accuracy)이 기존의

일반적인 이야기와는 달리 AR모델과 MLP모델의 예측능력이 크게 차이가 나지 않는 것으로 밝혀졌다. 또한, 추세의 정도 및 무작위정도를 의미하는 Hurst계수의 경우, 패턴이 안정될수록 예측모델들의 오차가 줄어드는 것을 볼 수 있었다. 이는 데이터가 무작위 보행을 따르는 경우 예측이 어렵다는 것과 부합되는 결과라 할 수 있다. 그리고 카오스 검정의 대표적인 리아프노프(Lyapunov)계수에 따른 모델들의 예측 결과에서 보듯이 MLP모델의 경우 안정적으로 수렴되어 가는 것을 볼 수 있었다. 이는 MLP모델의 입력노드결정을 위한 상관 차원으로 신경망 모델을 만든 것이 안정적으로 설계되었다고 유추할 수 있을것이다. 즉, 신경망 모델의 입력노드를 지정하기 위하여 수행하는 번거로운 사전작업 보다는 확정적 혼돈을 측정하는 방법중의 하나인 상관차원을 사용함으로써 예측시스템의 단점중 모델링시간을 단축할 수 있을 것이다.

예측결과 특이한 현상은, 예측결과에서 가장 좋은 모델이 AR모델로 지정되는 경우에는 MA모델의 예측력이 RMLP보다 좋은 것으로 나타났으며, MLP예측력이 가장 좋은 경우에는 RMLP의 예측력이 MA보다 좋음이 나타난 것이다. 이는 예측을 위한 모델구성시 주어진 패턴의 상관관계를 따지는 고전통계 구조와 패턴자체를 하나의 정보로 인식하는 신경망 구조가 패턴의 특성에 영향을 받는것으로 볼 수 있다.

마지막으로 관측되어지는 데이터에 대한 특성을 분류해 낼 수 있는 방법으로 카오스 검정이 사용될 수 있음을 밝혔으며, 이는 적응적 통합예측툴을 만들 경우 데이터 성격 규명의 척도로 사용할 수 있을 것이다.

본 연구에서는 신경망과 고전 통계모델의 예측성능을 비교함으로써 통합형 이론의 기초를 마련하였고, 고전 통계모델에서 이야기하는 무작위보행(RandomWalk)에 관한 설명이 무질서도(Chaos) 이론을 통하여 설명되었다.

추후에 연구되어야 할 사항으로는 신경망 예측 모델에 있어서 문제점으로 남고 있는 신경망 구조와 이에 따른 예측정도를 규명하며, 데이터의 특성에 따라 적응성 있는 자기 구조적 신경망 모델을 연구해야 할 것이다. 또한 시계열 데이터의 자기 복제성에 관해 좀더 연구하여 일반적 성질을 규명해야 할 것이다. 그리고 본 연구에서 제외된 동일한 패턴에 대한 추세성과 계절성의 첨가 및 삭제, 장기 구간 예측 등에 관한 연구가 수행되어야할 것이다.

References

남재우, 한국 거시 경제변수의 비선형성에 관한 연구, 석사학위 논문, 한국과학기술원(KAIST) 경영정책 과학과, 1994.

백웅기, "산업별 주가지수의 카오스 검정 및 비모수적 예측", Working Paper, 상명대 경제학과, 1996.

심기창, 최종욱, 정윤, "시계열 데이터의 성격과 Time-Delay 신경망의 예측력", 한국전문가시스템학회 추계학술대회 논문집, 1995.

이주장, "혼돈이론의 비선형 시스템에의 응용," 전자공학회지 제20권 제 3호, 1993.

지원철, "신경망을 이용한 시계열 분석 : M1-Competition Data에 대한 예측성과 분석", 한국전문가시스템학회지, 1995. 1, 135~148.

허명회·박유성, 시계열 자료 분석, 자유아카데미, 1994.

Bowen, J. E. "Using Neural Nets to Predict Several Sequential and Subsequent

Future Values from Time Series Data", *Proceedings of the First International Conference on Artificial Intelligence Applications on Wall Street*, 1991, 30~34.

Castillo, O., and P. Melin, "Intelligent Model Discovery for Financial Time Series Prediction Using Non-Linear Dynamical Systems Theory and Statistical Methods," Software Engineering Press, 1995, 80~89.

Connor, J., R.D. Martin, and L. Atlas, "Reccurent Neural Networks and Time Series Prediction." *IEEE Transactions on Neural Networks*, Vol.5, No.2, 1991, 240-254.

Hsu, W., L.S. Hus, and M.F. Tenorio, "A Clustnet Architecture for Prediction", *Proceedings of IEEE International Conference on Neural Networks*, Vol. I, 1993, pp329-334.

Jang, G-S., and F. Lai, "Intelligent Stock Market Prediction System Using Dual Adaptive Structure," Software Engineering Press, 1995, 88-97.

Jhee, W.C., and J. K, Lee "Performance of Neural Networks in Managerial Forecasting," *Int'l J. of Forecasting*, Vol. 2, 1993, 55~71.

Ott, E., T Sauer, J.W Yorke, Coping with Chaos, Wiley, 1994

Otawara, K., and L.T. Fan, "Synchronization in Chaotic System with Artificial Neural Networks", *Proceeding of IEEE International Conference on Neural Networks*, Vol.V. 1994, 3137-3142.

Peters, E.E., *Fractal Market Analysis*, John Wiley & Sons, Inc., 1994.

Peters, E.E., *Chaos and Order in the Capital Market*, John Wiley and Sons, Inc., 1991.

Tang, Z., and P.A. Fishwick, *Feed-Forward Neural Nets as Models for Timeseries Forecasting*, Technical Report, TR91-008, Computer and Information Sciences, Univ. of Florida, 1991.

Tenti, P., "Forecasting Currency Future Using Recurrent Neural Networks," Software Engineering Press, 1995, 243~252.

Tyree, E.W., and J.K. Long, "Forecasting Currency Exchange Rates : Neural Networks and the Random Walk Model," Software Engineering Press, 1995, 53~62.

Weigend, A.S., and N.A. Gershenfeld, *Time Series Prediction : Forecasting The Future and Understanding The Past*, Addison-Wesley Publishing Company, 1994.

Yun, S.Y., S. Nam-Kung, S.W. Shin, J.H. Roh, and J.U. Choi, "A Performance Evaluation of Neural Network Models in traffic Volume Forecasting," accepted by and to be published in *the Journal of Mathematical and Computer Modelling*, 1997.

Yun, S.Y., S. Nam-Kung, S.W. Shin, J.H. Roh, and J.U. Choi, "Application of a Recurrent Neural Network to Traffic Volume Forecasting," *Proceedings of ITS World Congress*, Orlando, FL, October, Proceedings CD, 1996.

[부록-1] 실험 데이터의 카오스 분석(일부 예시)

File name	Hurst	T-Statistics	Corre Dim.	BDS	Lyapunov	Range
ambns	0.5027	6.8457e-3	8.3562e-1	1.8697e1	0.010862	6.311-467.155
cel6ov	0.7246	1.1968e-2	9.8904e-1	2.1616e0	0.015501	57.172-125.274
clf16ov	0.6266	2.0327e-2	9.8414e-1	3.0635e0	0.003210	59.972-132.737
cnp16ov	0.7548	2.0074e-2	9.9644e-1	7.3302e1	0.002030	102.603-198.453
cpiaucns	0.9024	5.3672e-3	7.7092e-1	2.1736e1	0.005971	9.7-152.5
currdd	0.7163	2.3397e-2	8.9228e-1	1.1708e1	0.024363	138.5-754.2
dddfoins	0.4563	4.1376e-3	8.9500e-1	5.1931e0	0.063756	0.8-3.9
debtns	0.6643	1.8406e-2	7.1518e-1	1.4033e1	0.010284	64-13174.6
debtsl	0.6762	1.8419e-2	7.1634e-1	1.4045e1	0.010065	643.7-13261.3
demdepns	0.5613	2.2551e-2	9.3627e-1	7.6624e0	0.041627	108.7-401.8
demdepsl	0.7105	2.2879e-2	9.3804e-1	8.2462e0	0.049457	110-388.6
em1664	0.7310	2.0140e-2	9.8511e-1	2.7962e0	0.015865	54.299-121.571
euro	0.6142	3.5172e-2	8.2935e-1	7.1312e0	0.059817	0.0-36.4
indpro	0.6179	1.2996e-2	7.3326e-1	2.0477e1	0.027260	6-122.1
ira	0.7397	1.5577e-2	6.0788e-1	1.1741e1	0.030459	0.1-324.3
lgtdcbns	0.6698	2.3657e-2	6.3216e-1	1.2790e1	0.072375	1.1-380.2
lgtdcbsl	0.6551	2.3677e-2	6.3271e-1	1.2799e1	0.075980	1.2-380
lqdassns	0.7470	2.3983e-2	6.6761e-1	1.3420e1	0.004140	375.7-5461.3
ltdns	0.7059	1.8228e-2	6.2566e-1	1.2673e1	0.072721	1.1-549
M1	0.6648	1.3846e-2	8.5445e-1	2.0965e1	-0.001009	273.5-1155.3
M1ns	0.5663	1.8901e-2	7.7963e-1	1.4124e1	0.022560	137.6-1173.5
M1sl	0.7214	1.8920e-2	7.7760e-1	1.4324e1	0.022560	138.9-1152.3
M2	0.6790	1.5243e-2	9.7777e-1	5.0820e0	0.027262	1624.3-3705.9
M2ns	0.7486	2.5345e-2	6.7825e-1	1.3594e1	0.009234	287.7-3687-7
M2sl	0.7120	2.5348e-2	6.7911e-1	1.3610e1	0.003379	1624.3-3705.9
M3ns	0.7794	2.4990e-2	6.5611e-1	1.3260e1	0.001696	289.9-4444
M3sl	0.6698	2.4991e-2	6.5706e-1	1.3277e1	0.001412	289-4450.4
Manemp	0.7348	4.8995e-2	9.9626e-1	8.1601e1	0.075989	9.859-21.162
Nmfemp	0.7717	1.9967e-2	8.4402e-1	1.7062e1	-0.014426	19.882-98.058
Payems	0.7493	2.0805e-2	8.9711e-1	1.4241e1	0.004788	29.783-116.479
Pop	0.5813	1.7054e-2	9.9624e-1	7.5193e1	0.000537	151.13-261.755
Savingns	0.7003	1.8342e-2	7.2626e-1	1.3941e1	0.042992	136-1223.4
Savingsl	0.7516	1.8346e-2	7.2668e-1	1.3955e1	0.041487	136-1222

[부록-1] 카오스 검정 결과

[부록-2] 실험 데이터의 예측력 테스트(일부 예시)

filename	AR	MA	MLP	RMLP	Best
ambns	1.2003	181.9705	5.256	210.2	AR
cel6ov	0.7607	20.1530	0.2884	33.39	MLP
clf16ov	0.9284	21.2502	0.4327	39.94	MLP
cnp160v	0.1321	27.8557	0.03623	0.5587	MLP
cpiaucns	0.2620	56.1216	0.03945	62.06	MLP
currdd	1.5686	218.8479	3.051	243.3	AR
dddfoins	21.6198	0.0914	0.03945	0.4378	MLP
debtns	0.0966	4535.7062	1.105	2508	AR
debtsl	45.7075	4574.5494	3.766	2472	MLP
demdepns	9.3710	15.0335	1.447	177.4	MLP
demdepsl	0.3456	87.3738	0.9548	118.7	AR
em1664	0.6839	19.7041	0.078572	33.10	MLP
euro	0.2475	1.3485	2.648	13.40	AR
indpro	0.0912	37.8802	1.424	45.57	AR
ira	0.7857	115.9155	0.3729	0.1280	RMLP
lgtdcbns	10.9501	88.8620	5.675	58.91	MLP
lgtdcbsl	6.2774	84.3312	4.675	68.76	MLP
lqdassns	53.5490	1720.1267	14.23	1952	MLP
ltdns	10.3365	103.0196	7.006	26.95	MLP
M1	1.7875	262.3376	0.1253	239.2	MLP
M1ns	24.1846	319.1691	19.55	819.6	MLP
M1sl	6.8514	366.1244	4.862	451.1	MLP
M2	3.3827	456.6800	5.933	1001	AR
M2ns	14.6752	1041.2190	82.63	3391	AR
M2sl	14.6207	1104.7741	6.574	1318	MLP
M3ns	4.9877	1281.6119	36.18	496.0	AR
M3sl	27.6045	1332.6980	4.679	707.8	MLP
Manemp	0.0426	0.1266	0.03638	8.672	MLP
Nmfemp	0.0127	23.9467	0.03363	38.38	AR
Payems	0.0643	24.4212	0.09514	40.12	AR
Pop	0.1271	27.9231	0.01034	39.79	MLP
Savingns	0.9490	298.6923	11.26	558700000	AR
Savingsl	4.1677	296.9440	8.886	54070	AR

[부록-2] MA, AR, MLP, RMLP 예측 결과