JPEE 실천공학교육논문지
JOURNAL OF PRACTICAL ENGINEERING EDUCATION

실천공학교육 특별호
J. Pract. Eng. Educ. 16(1), 51-58, 2024

심층학습을 이용한 한국종합주가지수의 특성분석

Characteristic Analysis of Kospi Index Using Deep Learning

한상일*

한국기술교육대학교 산업경영학과

Snag-Il Han*

Dep. of Industrial Management, Korea University of Technology and Education, Cheonan 31253, Korea

[요 약]

본고는 Kospi와 S&P500 지수를 이용해 한미 주식시장 간 차이를 보고 이를 통해 정책적 시사점을 논하고자 한다. 이를 위해 기존 시계열 분석 방법에 더해 심층학습 방법으로 시장간 비교를 하되 주가 예측력, 자료 생성 능력 측면에서 비교를 했다. 월별 자료에서 시계열간 차이는 크지 않고 일별 자료에서 안정성 측면에서 차이가 약하며, 예측력이나 모의자료 생성에서도 차이가 크지 않았다. 본 연구결과와 같이 시장가격 움직임의 패턴이 한미간에 차이가 크지 않다면, 공매도의 부작용에 대한 대책으로 담보비율, 보고주기와 같은 직접적 규제보다 미국과 유사하게 투자자들의 자산운용 전략에 영향을 미치는 장기 주식보유에 대한 세제혜택과 같은 제도개편이 효과적이라 본다.

[Abstract]

This paper examines the differences between the Korean and American stock markets using the Kospi and S&P 500 indices and discusses policy implications through them. To this end, in addition to the existing time series analysis method, a deep learning method was used to compare markets, and the comparison was made in terms of stock price forecasting ability and data generation ability. In monthly data, the difference between time series was not large, and in daily data, the difference in terms of stability was weak, and there was no significant difference in predictive power or simulation data generation. As shown in the results of this study, if there is not much difference in market price movement patterns between Korea and the United States, tax benefits for long-term stocks investment will be effective against the side effects of short selling.

Key Words: Kospi, LSTM, Predictability, S&P500, Stability, TimeGan

http://dx.doi.org/10.14702/JPEE.2024.051

Received 16 January 2024; **Revised** 23 January 2024
Accepted 31 January 2024

***Corresponding Author**
E-mail: sihan@koreatech.ac.kr

I. 서론 및 기존문헌 연구

A. 연구의 필요성

최근 외국인의 공매도(short sale)에 대한 제한을 요구하는 소액 투자자들의 요구가 늘어나고 있으며 정부는 이를 반영하여 시장제도 개선을 추진하고 있다. 정부당국자는 공매도 금지가 우리나라 주식시장의 예측 가능성을 떨어뜨리지 않도록 조심스럽게 제도개선을 추진하고 있다는 발언을 하였다(연합인포믹스, 2023. 12. 05). 우리나라 시장에서 많은 투자자들은 미국과 유사한 투자전략이 작동한다는 믿음 하에 미국 등에서 개발된 다양한 기술적 패턴분석 등을 이용하여 매매를 한다. 문제는 미국 등과 다르게 국내 시장에서 외국인 비중이 매우 크고 소수의 외국인 기관에 의해 매매가 주로 이루어지므로 미국 시장과 구조에서 차이가 있을 수 있다. 그럼에도 미국시장에서와 마찬가지로 국내 주식시장에서 시장 가격의 예측 가능성은 많은 투자자들의 주된 관심사이다. 한편, 우리나라 세제는 장단기 자본이득을 구분하지 않는다. 반면 미국의 경우 26 U.S. Code 1222를 보면 1년을 기준으로 보유기간 등에 따라 세율이 명확히 구분된다. 이에 따라 미국시장에서 투자자들의 장기보유 유인이 강하고, 단기적인 가격변화에 상대적으로 민감하게 반응하지 않는다. 실제 시장가격 움직임의 패턴이 한미간에 차이가 크지 않다면 공매도의 부작용에 대해 가격에 영향을 미치는 직접적인 규제(예 : 담보비율 조정, 보고의무강화)보다 시장친화적인 참가자들의 자산운용 전략을 개선하는 장기보유에 대해 세제혜택을 부여하는 방안이 효과적이라 본다. 이에 본고는 타 연구와 다르게 인공지능 기법 등을 활용하여 일별자료에 대해 우리나라 주식시장과 미국 시장간의 차이점을 분석하여 정책적 시사점을 보고자 한다.

본고는 최근 발달되어 있는 인공지능 기법인 LSTM(long short-term memory) 모형 등을 통해 양 시장에서 예측 모형의 성과를 먼저 보고자 한다. 그리고 일별 시장 자료에 내재된 잠재변수에 의해 모의 자료를 생성할 수 있는지를 보았다. 지금까지 연구와 다르게 본고는 한미간 주식시장의 특징을 일별 지수자료를 이용하되 전통적 기법인 공적분(co-integration), 자기상관 분석보다는 시계열에 대한 인공지능 분석기법을 적용하고자 한다. 본고의 실증분석에 따르면 미국 주식 시장과 국내 주식시장 가격의 임의 보행적 성격이 유사하며, 예측 가능성도 유사한 것으로 나왔다. 이러한 영향 등으로 모의 가격 생성도 미국 주식 시장과 유사한 것으로 판단된다. 이러한 연구결과는 시장에서 일부 참가자가 주장하는 공매도 제도 개선이 예측력을 떨어뜨리는 경우 부작용이 발생할 수 있음을 시사한다.

B. 기존 문헌 연구

현실세계에서 시계열로 주어지는 주식시장 자료에 대한 예측은 관심사 중 하나이다. 미국 주식시장에 대한 실증 분석을 보면 일별 주가자료보다 월별 자료에서 자기상관성이 강해지고, 관측 주기를 증가시키면 예측 가능성 수준 자체는 낮으나 증가되는 것으로 보고된다(Campbell 등(1996), Cochrane(2005, pp. 391-395))[1,2]. 본고는 기존의 안정 시계열에 기초한 단변량 ARIMA(autoregressive integrated moving average) 예측 모형에 더해 심층학습 모형중 LSTM 모형을 이용하여 일별 자료의 예측 및 특징을 보고자 한다. 구체적으로 ARIMA, LSTM을 이용하여 예측력을 본 후 Facebook에서 개발된 Prophet과 비교를 통해 시사점을 본다. 이후 TimeGan(time generative adversarial network) 모형을 이용하여 모의 자료의 생성 가능성을 보고자 한다.

기존 계량경제학적 방법을 사용하는 모형의 경우 보통 회귀모형을 이용하여 시장 예측력을 보았는데 종속변수인 y_{t+1}을 미래시점으로 설정한 후 현재시점의 설명변수 X_t를 이용하여 설명계수의 통계적 유의성을 검증한다. Chun(2020)의 연구에 따르면 설명변수로 거시 경제 환경 변수를 사용하면 월별 자료에 대해 의미있는 설명이 가능한 것으로 보고되었다[3]. 한미 주식시장간 차이점에 대한 연구는 Yoon(2007)처럼 주로 시계열 분석기법인 공적분이나 인과관계 분석에 기초해 이루어졌다[4]. 최근 인공지능을 이용한 단일시장 투자분석은 많이 제시되고 있다. 반면 인공지능 기법을 이용한 시장간 비교연구는 아직 초보 단계이다. 인공지능을 이용한 시계열 분석은 미래 자료를 출력으로 과거 자료를 입력으로 모형화할 수 있는 LSTM 기법이 주로 적용되는데, 최근에는 attention[1]만으로 신경망을 구성하는 transformer 기법도 연구되고 있다. 인공지능 기법을 이용한 금융 시계열 예측에 대한 국내연구로 Choi(2021)은 인공지능을 이용한 주가예측 모형 전반을 검토했으며, Hong(2021)은 Facebook의 Prophet 시스템을 이용하여 암호화폐 시장에서 예측력을 보았다[5][6]. Jung-Kim(2020)은 Nasdaq 자료에 대해 일별 자료를 이용하여 LSTM 모형에 따른 예측력을 보았다[7]. 기존 LSTM 모형의 경우 상태변수에 따라 출력 변수 값이 영향을 받는 구조인데 과거시점이 멀어질수록 영향도가 낮아진다. 하지만 과거시점과 유사한 일정 패턴이 나타나면 시점과 무관하

[1]attention은 주어진 입력에 대해 key, value를 이용하여 중요도를 계산하는 기법으로 기존의 encoder, decoder모형과 다르게 상태변수를 이용하지 않는다.

게 동 패턴에 의해 미래 시점의 가격이 예측되는 게 필요하다. 이처럼 과거 시점의 중요도를 반영하여 예측을 하는 것이 transformer에 기반한 모형이다. Yoo 등(2021)은 transformer 기법을 적용하여 복수 주식간 상관성을 반영하면서 예측이 가능한 DTML 모형을 다양한 국가의 일별 자료에 대해 적용했다[7]. Lee 등(2022)은 TimeGan 모형을 이용하여 모의 자료를 생성한 후 트레이딩 시스템의 유용성을 검증했는데 수익률의 변동성이 낮고 수준 자체는 높은 알고리즘 개발에 유용함을 주장하고 있다[9]. 이외에도 ARIMA와 LSTM을 모두 적용하여 Nasdaq 시장에 적용한 연구로 Kobiela(2022)가 있는데, ARIMA 모형이 LSTM에 비해 2배이상 우수한 예측력을 보이는 것으로 보고되고 있다[10]. 한편, Lin 등(2014)은 Finbert 소프트웨어를 이용하여 다양한 특색(features)을 만든 후 가격자료에 추가한 후 Gan 모형을 미국 시장에 적용했다[11].

II. 분석 모형

A. LSTM을 이용한 시계열 자료의 예측

인공지능을 이용한 시계열 분석에 대해 기간 T 동안 수집된 시계열 자료 X를 시간 스탬프 t_i를 이용해 $X = \{(x_1, t_1), ..., (x_T, t_T), x_i \in R^D\}$로 표현하자. 시계열에 기초한 모형은 평균이나 분산이 기간상에서 안정된다는 가정하에 전개되는데 특정 시계열을 추세, 계절 변동 등을 반영하여 변환한 후 안정적 시계열로 만드는 변환 함수를 이용하여 예측을 한다. 따라서 특정 시계열이 먼저 안정적인지 검증이 필요하며 이는 보통 단위근 검증인 ADF 등으로 수행될 수 있다. 또한 복수의 불안정한 시계열간 선형관계를 공분산 모형 등으로 추정하기도 한다. 본고는 안정성 분석후 ARIMA 모형으로 예측을 하여보고자 한다. 이때 단변량 ARIMA를 적용하거나 외생변수를 추가로 반영한 ARIMAX 등이 사용될 수 있는데 본고는 예측 모형에서는 D=1인 일변수만을 고려하겠다.

최근 인공지능 기술이 급속히 발전함에 따라 동 방법을 이용하여 시계열을 분석하는 기법도 늘어나고 있다. 인공지능은 크게 데이터에 정해진 라벨값을 이용해 학습하는 지도학습(supervised learning) 모형과 이를 활용하지 않는 비지도학습 모형 그리고 강화학습(reinforcement learning) 모형으로 구분된다. Foumani 등(2023)은 결과치에 대해 분류형 값을 주는 모형과 수치형 값을 주는 모형으로 구분하고 이에 추가하여 자료확장 및 이전학습 모형으로 구분하고 있다[12]. 본고는 자료확장 모형에 생성형 모형에 의한 자료생성을 추가 하는게 유용하다 본다.

인공지능 기법중 RNN(순환신경망, recurrent neural networks)은 순차적 자료를 처리하는 모형으로 RNN에서 출력은 현재 입력뿐만 아니라 과거 입력 자료 등에 의존하는 구조를 갖는다. 이렇게 자료가 재귀적으로 사용되면 최적화를 위해 사용되는 기울기의 불안정성이 발생하는데 이를 해결하는 방안으로 LSTM 알고리즘이 Hochreiter-Schmidhuber(1997) 등에 의해 개발되었다[13]. RNN은 상태변수 과거치와 현재의 자료인 h_{t-1}, x_t를 입력 변수로 하여 상태변수와 출력값으로 h_t, o_t를 생성한다. 반면 LSTM은 셀상태를 나타내는 C를 추가하여 출력값을 먼저 생성한 후 셀상태 변수와 결합해 새로운 상태변수 h를 만든다. 최근에는 자료의 전체 구조에서 부분의 중요도를 반영해 RNN의 입력 자료를 생성하는 attention 및 attention 만으로 자료를 생성하는 tansformer 모형이 제시되고 있고 이를 시계열 자료에 적용하는 방법이 Zhou 등(2020)에 의해 Informer 소프트웨어로 제시되었다[14].

B. 생성형 TimeGan 모형에 의한 시장 가격의 생성

분포함수를 명시적으로 구하는 경우 자료를 저차원의 몇개의 잠재(latent) 변수 공간에 투사한 후(encoder) 모수를 이용하여 자료를 생성(decode)하여 원자료와 비교를 통해 학습하는 autoencoder 기법이 있다. 한편, Goodfellow 등(2014)은 게임모형에서 mini-max 균형점 개념을 학습에 도입하여 Gan(적대적 생성 모형) 모형을 제시했다[15]. Gan은 묵시적 생성 모형으로 해석되며, encoder와 decoder로 구성된 autoencoder 모형과 유사하되 encoder는 생성자(generator, G) 그리고 decoder는 판별자(discriminator, D)로 구성된다. 생성자는 자료와 최대로 유사한 모의 자료를 생성하는 것을 목적으로 하며 판별자는 자료와 모의 자료를 구분하도록 학습을 하는 것이다. 따라서 두 기능은 서로 적대적인 목적 함수를 갖는다.

2인 게임 모형에서 균형점은 각 참여자들이 상대의 전략에 따른 효용을 극소화하면서 자기의 이익을 극대화하는 전략을 선정하는 경우 mini-max 형태를 띠는 목적함수 $V(G, D)$하에서 균형상태로 정의된다. 이를 수식으로 보면 게임모형에서 목적함수는 minmax $V(a_i, a_{-i})$으로 주어지며 이때 a_{-i}는 참여자 i를 제외한 나머지 참여자의 전략을 나타낸다. 즉 상기 식은 나머지 참여자의 전략은 참여자의 효용을 최소화하면서 참여자 i의 효용을 극대화하는 자신의 전략을 선택한다. 게임모형의 목적함수를 도입하여 Gan 모형은

$$\min_G \max_D V(D, G) =$$
$$E_{x \sim P_{Data}(x)}[\log D(x)] + E_{x \sim p_z(x)}[\log(1 - D(G(z))]$$

을 손실함수로 설정한다. 여기서 p_{Data}는 자료에 기초한 밀도함수, p_z는 잡음 z에 기초한 선험(prior) 밀도함수를 나타낸다. 그리고 $D(G(z))$는 잡음에 기초해 생성한 자료에 대해 판별자 D의 출력 값을 나타낸다. 식의 첫 번째 부분 $\log D(X)$는 입력 x에 대해 판별자를 극대화하는 것을 의미한다. $D(G(z))$는 생성자에 의해 만들어진 정보인 $G(z)$를 판별자에 입력하는 경우 값을 이용해 계산된 값, 즉 판별식을 속이는 경우를 극대화하는 것으로 해석된다.

금융 시계열은 자기 상관성을 띠는게 일반적이며 동시에 외생변수의 영향을 받는 시스템으로 볼 수 있다. 그런데 두 영향이 혼재되어 실현이 되고 외생변수를 충분히 인식하기도 어렵다. 또한 설명변수를 늘리는 경우 잡음에 과적합(over-fit)되는 문제가 발생한다. 더욱이 일반적인 금융 시계열은 안정성(stationality)과 비안정 경계선상 특징을 보인다. 이로 인해 실제 금융시장에서 거래 시스템을 구축할 때 과거 자료를 사용하는 경우 자료내(in sample) 편기로 인해 개발된 거래 알고리즘의 적정성 평가에 오류가 발생하기 쉽다. 따라서 적절한 거래 시스템을 구축하기 위해서는 시장과 동일한 구조를 띠면서 과거 자료와는 다른 수치를 가지고 거래 시스템의 성과를 평가하는 것이 중요하다. 이에 적합한 인공 지능 생성모형이 Gan이 될 수 있다.

시계열 자료의 특징을 반영하여 낮은 차원의 잠재변수를 명시적으로 고려하면서 지도학습을 결합한 Yoon 등(2019)의 TimeGan 모형은 autoencoder와 구조가 유사하며 입력 X를 받아들이는 임베딩 e와 복구 r을 통해 입력 추정치 $\tilde{X}$을 출력하는 부분으로 입력 자료를 가공한다[16]. 그리고 Gan 모형에서 입력 자료 X 대신 잠재 상태값 h를 입력하여 판별을 한다. 동 저자는 RC-Gan, C-RNN-Gan 등과 같은 다양한 자료 생성 방법과 TimeGan을 비교하였으며 판별스코어, 예측 스코어에서 TimeGan의 우수성을 제시한다.

III. Kospi 지수에 대한 실증분석

A. 자료의 선정과 학습환경

본고는 우리나라 주식시장과 미국 시장에 대해 ARIMA, LSTM 모형 등으로 예측력을 보고, 생성형 모형을 이용하여 시장간 차이를 분석하기 위하여 KOSPI 종합지수와 S&P500 지수를 선정했다. 지수에는 일반적으로 배당 수익률 정보가 반영되지 않아서 총 수익률 정보측면에서는 한계가 있다. 그럼에도 두 시장간 특징 분석에 사용되는 것에는 문제가 없다고 본다. 일별 지수와 월별 지수를 yfinance를 이용하여 Yahoo에서 추출했으며, 기간은 2006년 1월초부터 2023년 12월말까지의 일별 자료를 사용했다. 일별자료는 영업일을 맞추지 않고 분석을 하되 공적분 추정에서는 자료를 합병하여 처리하였다. 예측모형은 복수의 가격자료를 이용하여 가격간 상관계수를 반영하여 예측하는 게 유용한데 본고는 시장간 관계를 보려는 것이므로 예측에서는 종가만이 이용되었고 생성하는 경우 시고저종 가격이 사용되었다. 일별 가격 자료의 경우 잡음이 상대적으로 높을 것으로 추정되므로 월별 자료를 통해 시장의 특성을 보완 분석하면 유용할 것이다. 월별 S&P500 지수의 경우 Robert Shiller가 제공하는 자료를 사용했는데 1900년 1월부터 2023년 6월까지의 자료이다. 동 자료는 장기간에 걸친 주가지수와 배당 수익률을 제공하므로, 주식시장의 특징인 주식 수익률이 무위험 자산의 수익률보다 높은 주식 프리미엄 현상, 가치주식에서 초과수익이 발생하는 가치 프리미엄 현상 등을 거시적으로 분석하는데 유용하다.

예측 자료의 경우 종가만을 사용하므로 주어진 자료는 1차원 구조이다. 전체 가격 자료를 학습용 80%, 테스트용 20%로 구분하여 사용했으며 ARIMA, LSTM, Prophet 모두에게 적용했다. 또한 통계치는 파이썬 statsmodels와 sklearn을 이용하여 계산하였다. 인공지능 모형의 경우 주어진 가격자료에 대해 maxminScale을 적용하여 가격 자료를 정규화한 후 입력자료로 사용했다. LSTM 모형은 tensorflow 1.5 버전의 keras를 사용했다. 신경망은 각 레이어의 units = 100으로 하면서 LSTM을 2층으로 구성한 후 Dense 함수를 적용하여 1차원으로 줄여서 실측치와 예측치간 차이에 대해 mse(mean squared error)를 손실함수로 사용하여 100번을 반복학습하는 구조로 했다. 관측기간이 T라면 일별 종가에 대해 자료구조는 $T\times1$과 같은 벡터구조를 띠는데 여기에 과거 일정기간 S 동안의 관측치를 반영하면 $T\times S$ 구조가 된다. 물론 n개 자산 가격간 상관계수를 고려하는 경우 $T\times S\times n$과 같은 자료구조를 사용해도 된다. 본고는 S=100을 사용했고 Prophet의 경우 파이썬 1.1.5 버전을 사용했다

TimeGan 모형의 경우 시고저종 모든 자료를 이용하여 학습을 했다. 학습반복(epoch)을 50,000로, 과거 시계열(seq_len) 75, 은닉 상태의 차원(hidden_dim) 75, 각 게이트의 레이어 3개, 배치 사이즈 128로 학습을 했다. Yoon 등(2019)은 구글 가격자료에 대해 TimeGan을 적용했는데 해당 수치는 Yahoo 등에서 제공하는 수치와 차이가 큰데 이는 총수익을 반영하여 가격을 조정한 것으로 추정된다. 일단 모형의 정확성을 검증하기 위해 동 저자들이 제공하는 자료를 학습하여 비슷한 결과가 나오는걸 확인한 후 S&P500과 Kospi에 적용하였다.

하드웨어는 Intel I-7, Titan X 3-GPU 환경으로 설정했고 2개의 GPU에 작업을 할당하여 Tensorflow 1.5를 이용하여

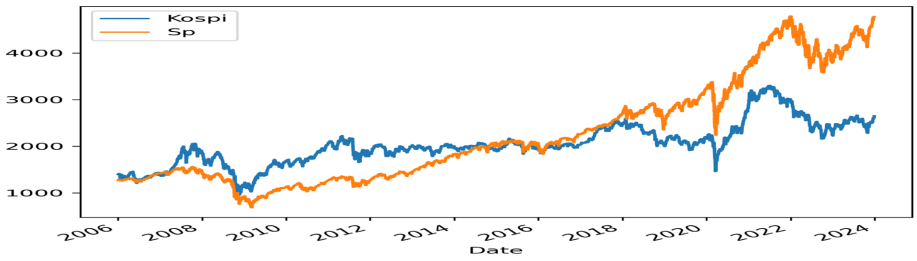

그림 1. 2006~2023년간 Kospi와 S&P500
Fig. 1. Kospi and S&P500 from 2006 to 2023.

표 1. Kospi와 S&P 500 지수 통계량
Table 1. Kospi and S&P 500 Index Statistics

	Kospi	S&P500	비고
기간	'06.1.2.~23.12.28	'06.1.3.~23.12.29	
관측갯수	4441	4529	
평균	2039.11	2246.83	
표준편차	441.51	1092.26	
최소값	938.75	676.53	
최대값	3305.21	4796.56	
ADF(p값)	−65.72(0.0)	−15.28(0.0)	차분
표준편차	22.84	28.10	차분

결과를 얻었다. 대부분 작업에 대한 소요시간은 30분 이내이나 TimeGan의 경우 계산에 22시간 정도 소요되었다.

B. 예측력

두 시장간 차이를 간단히 시계열로 보면 그림 1과 같다. 비교를 위해 초기 값을 표준화해도 되나 초기 시점인 2006년 1월 두 주가지수가 유사한 수준에 있어 이를 수행하지 않았다. 그림 상으로 보면 최근에 올수록 미국 지수 자체는 상승곡선을 그리고 있고 우리나라 지수는 상대적으로 횡보를 보여준다. 표 1은 두 시장간 통계량의 차이를 보여주는데 일단 영업일 차이가 연 5일 정도 미국이 많고 평균 차이는 크지 않다. 오히려 표준편차는 미국시장이 2배 정도 커서 시장의 변동폭이 미국 시장에서 향후 클 것으로 판단된다. 금융시계열의 안정성(stationality)은 시간에 걸쳐 평균, 표준편차와 같은 통계량이 변화하는 지를 나타내며 보통 단위근 검정 통계량인 ADF로 측정이 가능하다. 이에 두 시장의 1차 차분 자료에 대해 ADF 통계량으로 안정성을 검증해 보았으며 S&P500 지수 및 Kospi 모두 월간 자료에서 안정성을 보였다(표 1).

Cochrane(2005) 등에 따르면 S&P500 지수의 경우 주식 프리미엄이 장기적으로 관측된다고 보고되고 있고 Kospi 시장도 유사한 현상을 보일 것이다[2]. 지난 2016년 이후 우리나라 성장률이 미국보다 높았음에도 불구 미국의 첨단 산업 중심의 시장구조 개편으로 Kospi의 일별 수익률은 상대적으로 낮은 것으로 보인다. 일별 자료를 보면 미국 시장과 한국 시장간에는 미세한 차이가 있는 것으로 보인다. 환율변환을 하지 않고 단순 일별 가격만을 사용하여 S&P500 지수와 Kospi간에 공적분(co-integration)을 보았다. 가격 수준에서는 p-value가 0.20, 차분에서는 0.0을 보여 1차 차분 자료에 공적분이 존재하며 차분은 수익률로 해석되므로 수익률 측면에서 S&P500 시장이 원인변수로 작동할 것으로 보인다.

그림 2는 시장가격의 일차차분에 대해 자기상관 함수를 보여주고 있는데 두 시장 모두 자기 상관성이 있는 것을 알 수 있다. 다만 미국 자료의 경우 우리나라보다 자기 상관성이 높은 것으로 나왔다. 표 1과 그림 2에서 보듯 Kospi 및 S&P500 지수 모두 1차 차분가격 자료에서 자기상관성이 존재하며, 1차 차분에 대한 ADF 통계량에 따르면 안정성을 띠는 것으로 보인다.

표 2는 상기 시계열 자료에 대해 예측 모형의 성과를 보여주는데 MAPE(mean absolute percentage error)[2]를 이용해 시장간 비교를 보면 ARIMA 모형의 예측력이 우수하며 ARIMA, LSTM 모두 MAPE가 9% 전후 값을 보여준다. 반면 Prophet 모형의 성과는 매우 낮은데 이는 동 모형이 직전일 자료에 기반한 예측이 아닌 추세선을 예측하는 것에 기인한다. MAPE가 비교적 높은 수준인 9% 전후에서 나타났다. 학습 모수를 고정한 상태에서 검증기간을 약 3년 넘게 설정해 장기간 예측한 것에서 오차 값이 커진 것으로 보인다. 이를 보완하기 위해 윈도우 등을 설정한 후 각 윈도우별 학습과 검증을 하면서 모형의 정확성이 실시간 학습과 유사한 구조하에서 평가될 수 있을 것이다. 본고는 시장간 비교를 하는 것을 목적으로 하므로 심층학습 방법 등을 개선하지는 않았다.

[2] $MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$ 로 정의된다. 이때 $y, \hat{y}$는 실측치, 예측치를 나타낸다.

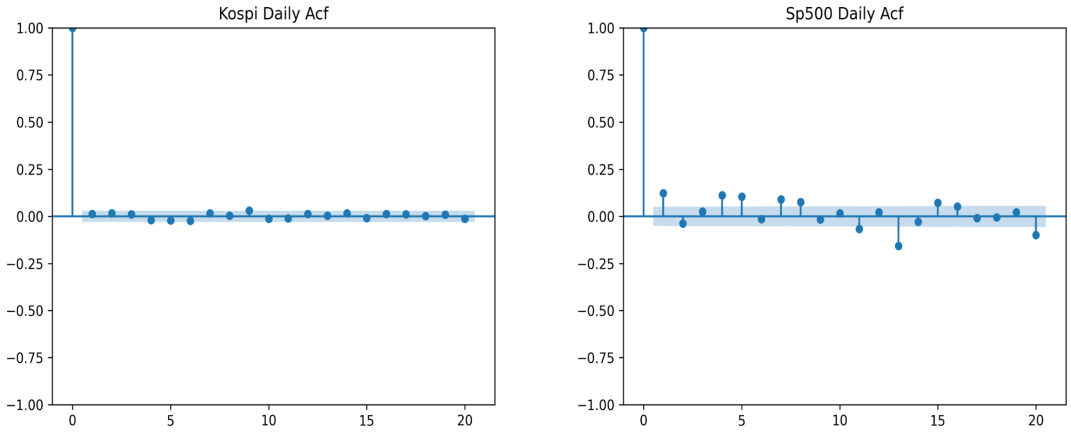

그림 2. 일별 차분 가격자료의 자기상관함수

Fig. 2. Autocorrelation function of daily differential price data.

표 2. 예측 모형별 오차 통계량

Table 2. Error statistics for each prediction model

	ARIMA		LSTM		Prophet	
	Kospi	S&P500	Kospi	S&P500	Kospi	S&P500
MSE	821.91	2025.10	982.53	5598.69	1094322	543472
MAE	22.02	33.94	24.55	63.14	993.21	651.89
RMSE	22.69	45.00	31.35	74.82	1046.10	737.21
MAPE	0.092	0.092	0.097	0.124	0.606	0.392

C. 자료의 생성

PCA(principle component analysis) 및 t-Sne(t-distributed stochastic neighbor embedding)를 이용하여 차원 축소된 자료에 대한 성과를 볼 수 있다. PCA의 경우 각 시점 자료에 대해 특성치를 구하여 2차원 공간에 뿌릴 수 있다. t-Sne은 두 자료간 거리를 구해 확률치로 변환한 값을 이용해 Kullback-Leibler divergence를 최소화하는 낮은 차원의 새로운 변수를 학습하여 2차원 공간에서 비교하는 기법이다. 따라서 2가지 2차원 이미지를 이용하여 직관적으로 두 자료간 비교가 가능

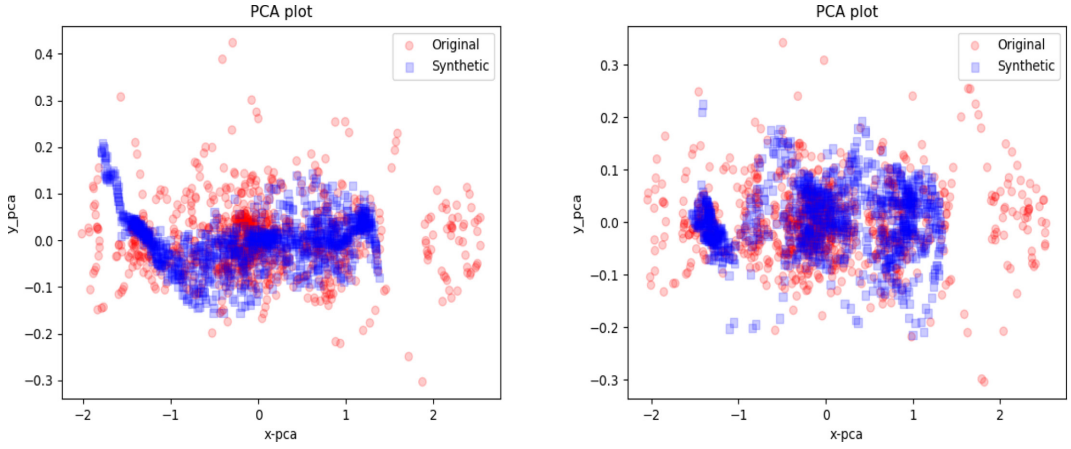

그림 3. Kospi와 S&P500의 PCA

Fig. 3. PCA of Kospi and S&P500.

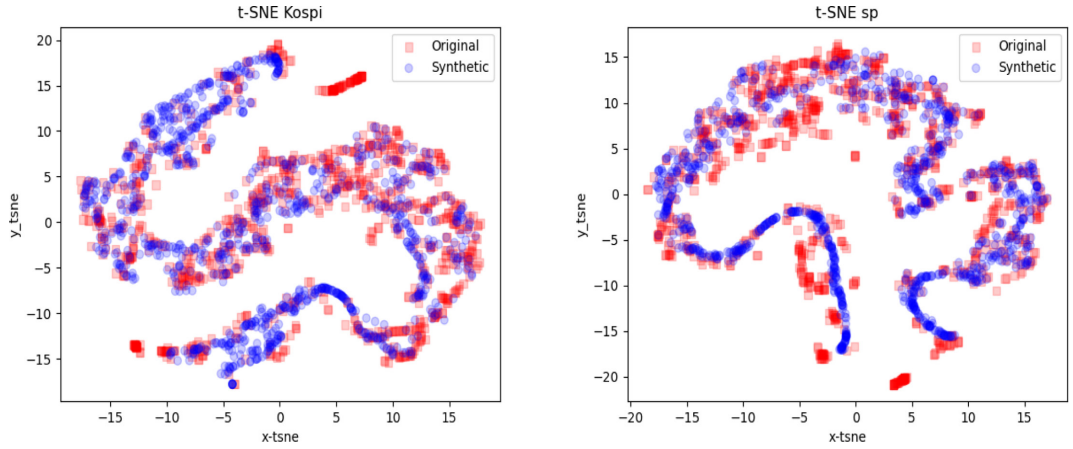

그림 4. Kospi와 S&P500의 t-Sne

Fig. 4. t-Sne of Kospi and S&P500

표 3. Kospi와 S&P500의 생성 점수

Table 3. Generative Score of Kospi and S&P500

	Kospi	S&P500
판별 점수	0.2192	0.1494
예측 점수	0.0814	0.0166

하다. 이외에도 추정치에 대해 구별능력 및 예측 능력 값을 볼 수 있다.

미국 S&P500 지수 및 Kospi 일별 자료에 대해 TimeGan으로 생성 모형을 적용한 결과를 보기 위해 PCA 및 t-Sne를 보면 다음과 같다. 먼저 PCA를 보면 한미 양국 모두 네모로 표기된 원자료와 원으로 표현된 생성 자료간 특성치 값들에서 유사도가 높음을 알 수 있다(그림 3). 차원을 축소하여 2차원 공간에 원 자료와 생성 자료를 보여주는 t-Sne를 보면 미국 자료의 경우 원 자료와 모의 자료의 유사성이 매우 높으며 우리나라 자료 역시 유사성이 높다(그림 4).

따라서 2차원 이미지 자료를 활용한 한미간 주식시장 자료로 생성형 자료가 만들어진다는 것이다. 이를 수치적 측면에서 보기 위해 파이썬 sklearn을 이용하여 판별 점수와 예측 점수[3]를 계산했다. 표 3을 보면 판별 점수는 Kospi의 경우 0.2192로서 S&P500의 0.1494보다 약간 크고 예측 점수도 Kospi의 경우 0.0814로서 S&P500의 0.0166 보다 높음 수준이다. 우리나라 시장에서 자료의 생성이 미국보다 어려우나 생성 자체는 가능한 것으로 판단된다. 2000년 이후 인터넷을 통한 매매가 전세계적으로 이루어지고 결제 시스템의 유사성이 높아지면서 시장간 자료의 유사성이 높아진 것으로 판단된다.

[3]동 점수는 낮을수록 정확성이 높다.

IV. 결 론

본고는 Kospi와 S&P500 지수를 최근 자료를 이용해 한미 주식시장 간 차이를 살펴보았다. 이를 위해 기존 시계열 분석 방법에 심층학습 방법을 추가하여 예측력과 자료 생성 능력을 보았다. 월별 자료에서 시계열간 안정성 차이는 크지 않으며 일별 자료에서도 마찬가지였다. 또한 예측력이나 모의자료 생성에서도 차이가 크지 않았다. 따라서 Lee 등(2022)과 같은 거래 시스템 구축이 가능함을 시사한다[9]. 또한 attention으로 시계열상 공분산을 반영하여 예측하는 Yoo 등(2021) 등에 의한 transformer 방법이 우리나라 시장에도 적용 가능함을 시사한다[8].

최근 정부는 공매도에 대한 제도 개선을 추진하고 있는데 외국에는 없는 강한 제도를 도입하는 경우 오히려 부작용이 커질 수 있다. 공매도 제한이 시장의 효율성을 떨어뜨리는지에 대해서는 논란이 많다[4]. 근본적으로 공매도가 타 국가보다 많이 이루어진다면 가격이 내재가치에 대한 판단신호로서 기능 미약에 기인한다고 본다. 이는 시장 참가자들의 심리적 요인에 의해 가격의 변동이 심한 측면에 기인할 수 있다. 우리나라 주식시장 가격의 움직임은 미국과 통계적으로 차이가 높지 않으므로 미국과 유사하게 기관 투자자에 의해 특정 종목을 장기간 보유하는 유인 체계의 검토가 필요하다. 미국의 경우 기관투자자중 주식 회전율이 높지 않는 다양한 전략적 투자자가 많다. 미국의 경우 26 U.S. Code 1222를 보면 1년을 기준으로 보유기간 등에 따라 세율이 명확히 구분하고 분리과세 등이 이루어진다. 이와 같이 장기보유를 우대하

[4]"The impact of short sale bans during crises: A closer look at the 2020 Covid crash response", CEPR, 2023, https://cepr.org/voxeu

므로 투자자들은 단기적인 가격변화에 반응하여 이익을 추구하는 것보다 장기적 이득을 추구하는 것으로 판단된다. 반면 우리나라 소득세제는 특정 종목의 장기보유에 대한 세제 혜택이 없고, 투자이익의 수준을 기준으로 일정금액(3억) 이상인지 여부에 따라 부과되는 제도이다(소득세법 104조 1항 11). 오히려 장기보유에 따른 과세혜택 부여에 대한 사회적 논란만 많은 형국이다. 우리나라에서도 전략적 투자자를 육성하기 위해서는 주식투자 자체에 대한 세제 혜택보다 복수의 특정 종목으로 구성된 포트폴리오의 장기 보유에 대한 세제 혜택을 강화하는 게 좋다고 본다. 이럴 경우 투자자의 장기 보유 관점에서 매매가 유도됨으로써, 공매도에 따른 급등락에 영향을 받는 게 줄어들어 공매도의 시장가격에 대한 부정적 영향이 완화되리라 본다.

참고문헌

[1] J. Campbell, A. Lo, and A. MacKinlay, The Econometrics of Financial Markets, 1997.

[2] J. Cochrane, Asset Pricing: Revised Edition, Princeton, 2005.

[3] S. Chun, "Predicting Korean stock market return with Financial and macro variables," *The Journal of Insurance and Finance*, vol. 31, no. 1, pp. 87-113, 2020.

[4] J. I. Yoon, "The evaluations and the international comparisons of the integration with U.S. stock market," *Journal of Money & Finance*, vol. 21, no. 1, pp. 55-92.

[5] H. Choi, "Stock prediction analysis through artificial intelligence using big data," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 25, no. 10, pp. 1435-1440, 2021.

[6] S. H. Hong, "Cryptocurrency automatic trading research by using facebook deep learning algorithm," *Journal of Digital Convergence*, vol. 19, no. 11, pp. 359-364, 2021.

[7] J. Jung and J. Kim, "A performance analysis by adjusting learning methods in stock price prediction model using LSTM," *Journal of Digital Convergence,* vol. 18, no. 11, pp. 259-266, 2020.

[8] J. Yoo, Y. Soun, Y. Park, and U. Kang, "Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts," ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2021.

[9] J. Y. Lee, J. H. Lee, B. G. Choi, and J. W. Song, "Trading algorithm selection using Time-series generative adversal networks," *Smart Media Journal*, vol. 11, no. 1, pp. 38-45, 2022.

[10] D. Kobiela, D. Krefta, W. Król, and P. Weichbroth, "ARIMA vs LSTM on NASDAQ stock exchange data," *Procedia Computer Science,* vol. 207, pp. 3836-3845, 2022.

[11] H. Lin, C. Chen, A. Jafari, and G. Huang, "Stock price prediction using generative adversarial networks," *Journal of Computer Science,* vol. 17, no. 3, pp. 188-196, 2021.

[12] N. Foumani, L. Miller, C. Tan, G. Webb, G. Forestier, and M. Salehi, "Deep learning for time series classification and extrinsic regression: A current survey," arXiv: 2302.02515, 2023.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation,* vol. 9, no. 8, 1735-1780, 1997.

[14] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," arXiv 2012.07436, 2021.

[15] I. Goodfellow, P. Jean, M. Mehdi, X. Bing, W. David, S. Ozair, C. Aaron, and Y. Bengio, "Generative adversarial nets," *NIPS*, pp. 2672-2680, 2014.

[16] J. Yoon and D. Jarrett, "Time-series generative adversarial networks," Neural Information Processing Systems (NeurIPS), 2019.

한 상 일 (Sang-Il Han)_정회원

1998년 8월 서강대학교 경영학과 박사, 재무관리
1999년 ~ 2004년 : 한국금융연구원 연구위원
2005년 ~ 현재 : 한국기술교육대학교 산업경영학과 교수
〈관심분야〉 인공지능, 금융시장 예측, 단백질 동학 및 설계