The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 723–734

DOI: https://doi.org/10.5351/KJAS.2021.34.5.723
Print ISSN1225-066X / Online ISSN 2383-5818

Wind power forecasting based on time series and machine learning models

Sujin Park[a], Jin-Young Lee[a], Sahm Kim[1,a]

[a]Department of Applied Statistics, University of Chung-Ang

Abstract

Wind energy is one of the rapidly developing renewable energies which is being developed and invested in response to climate change. As renewable energy policies and power plant installations are promoted, the supply of wind power in Korea is gradually expanding and attempts to accurately predict demand are expanding. In this paper, the ARIMA and ARIMAX models which are Time series techniques and the SVR, Random Forest and XGBoost models which are machine learning models were compared and analyzed to predict wind power generation in the Jeonnam and Gyeongbuk regions. Mean absolute error (MAE) and mean absolute percentage error (MAPE) were used as indicators to compare the predicted results of the model. After subtracting the hourly raw data from January 1, 2018 to October 24, 2020, the model was trained to predict wind power generation for 168 hours from October 25, 2020 to October 31, 2020. As a result of comparing the predictive power of the models, the Random Forest and XGBoost models showed the best performance in the order of Jeonnam and Gyeongbuk. In future research, we will try not only machine learning models but also forecasting wind power generation based on data mining techniques that have been actively researched recently.

Keywords: time series, machine learning, wind power forecasting, random forest, XGBoost

1. 서론

최근 미세먼지와 제철 기온 상승, 또는 지진이 발생할 때마다 원자력발전소의 가동 여부에 대한 소식 등 환경과 안전에 관련된 문제가 제기되고 있다. 2021년 미국은 파리기후변화협약 복귀 행정명령에 서명을 해 글로벌 기후변화에 매우 빠르고 민감하게 반응하며 '청정에너지혁명'이라는 기후변화 대응을 시도하고있다. 또 다른 사례 중 유럽은 '그린딜' 정책을 수립해 2050년까지 이산화탄소 배출과 흡수가 완전히 상쇄되는 탄소 중립을 달성하고자 한다. 이로 인해, 신재생에너지에 대한 관심이 전 세계 각국에서 늘어나고 있는 추세이다. 국내에서는 국제사회의 기후변화 대응 노력에 동참하기 위해 2016년 11월 3일 파리 협정 국내 비준 절차를 완료하고, 유엔(UN)에 비준서를 기탁하여, 12월 3일부터 국내에서 발효되었다. 이렇게 세계적으로 태양광, 풍력에너지 등을 포함한 신재생 에너지의 발전량이 증가하고 있다. 한국 풍력산업협회에 따르면, 2010년 이후로 현재까지 신규설비용량은 꾸준히 늘어나고 있고, 신규 단지수 또한 28개에서 103개로 증가했다. 이에 발맞춰 프로젝트 개발 및 투자가 이루어지고 있다. 이 중 풍력에너지는 빠르게 발전하고있는 재생에너지 중 하나이다.

풍력이란, 바람이 가진 운동에너지를 이용하여 전기에너지를 생산하는 시스템이다. 즉, 에너지 변환과정을 통해 전력을 생산하는데, 특히 우리나라는 해안선이 길어 풍력발전을 하기에 유리한 조건을 가지고 있다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2016R1D1A1B01014954).
[1] Corresponding author: Department of Statistics, Chungang University, 84 Heuksukro, Dongjak-Gu, Seoul 06974, Korea. E-mail: sahm@cau.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

풍력발전은 설치 장소에 따라 육상풍력과 해상풍력으로 구분할 수 있는데 육지에 풍력발전단지를 건설하여 발전한 것을 육상풍력, 바다에 풍력발전단지를 건설하여 발전한 것을 해상풍력이라고 한다. 육상풍력은 바람을 이용하여 환경오염 및 고갈염려가 없으며 해상풍력은 육상풍력 대비 높은 입지제약에서 자유롭고, 대형화로 높은 이용률 확보가 가능하다는 장점이 있다. 풍력발전기의 구조를 살펴보면 바람으로부터 회전력을 생산하는 회전날개(blade)와 회전축을 포함한 회전자(rotor), 이를 적정 속도로 변환하는 증속기(gearbox)로 이루어져 있다. 또한, 풍력발전기는 회전축의 방향에 따라 수직축 풍력과 수평축 풍력 두가지로 분류되는데, 발전기의 회전축에 따라 풍향의 영향을 받기도 한다. 풍력은 이렇게 방향에 따라, 또는 변동성이 강한 대기의 움직임에 따라 역동적인 특성을 가지며 쉽게 예측할 수 없다는 특징을 가지고 있다. 우리나라의 경우, 바람 세기가 일정하지 않다는 점과 발전기에서 발생하는 소음에 대한 주민들의 우려 등으로 인해 풍력에너지의 성장이 더딘 편이었지만, 최근 정부의 적극적인 신재생에너지 정책과 기술력이 발전하고, 정부의 소음 관련 가이드라인 설정 등으로 인해 인근 환경을 고려한 발전소 설치가 추진되고 있어 국내 풍력 보급이 점차 확대되고 있는 추세이다. 이에 따라, 풍력 에너지의 수요를 정확히 예측하기 위한 시도들이 계속되고 있다.

풍력발전량의 예측을 위하여 Cadenas와 Wilfrido (2010)은 ARIMA, ANN(인공신경망)을 이용한 Hybrid 모형을 적용해 비선형적으로 발생하는 오류를 줄이기 위한 Hybrid 모형이 풍속데이터의 변동성을 잘 예측해 mean absolute error (MAE)와 mean squared error (MSE) 기준으로 높은 정확도를 가진다는 점을 보였다. Catalão 등 (2011)은 포르투갈에서의 풍력 데이터를 이용하여 wavelet 변형을 적용해 풍력 발전량을 예측하려고 시도했다. Liu 등 (2011)은 미국 콜로라도 주의 풍속을 4가지의 서로 다른 높이에서 측정한 후, ARMA-GARCH, ARMA-GARCH(-M)모형을 다양하게 변형해 10가지의 접근을 시도해 풍속 데이터의 변동성, 이분산성을 효과적으로 개선하였다. Zeng 등 (2011)은 SVM을 응용한 모델을 이용하여 풍력발전량을 예측했고, Pinson (2012)은 덴마크 horn rev의 풍력단지 데이터를 이용하여 generalized logit (GL) transform을 적용해 비선형적, 변동성이 강한 풍력 데이터의 특성을 고려한 후, 개선된 AR, conditional parametric auto-regressive (CPAR) 모형을 이용해 풍력을 예측하였다. Anastasiades와 McSharry (2013)는 quantile regression(분위수 회귀)모형에 여러가지 외생 변수를 추가해 기존 모형의 예측능력을 상승시켰다. Li 등 (2015)는 풍력 발전량 예측을 위한 앙상블 모형으로 NNs(신경망)과 wavelet 변형, 변수 선택법과 partial least squares regression (PLSR)을 이용하였다. 이 모형으로 48시간을 예측했을 때, MAPE 기준 5.0%라는 낮은 수치를 기록했다. Wu 등 (2016)은 중국 북동쪽에 위치한 풍력 발전소에서 수집한 15분 간격의 데이터를 이용하여 DNN(딥러닝 인공신경망) 모형과 LSTM, RNNs(순환 신경망)을 적용해 발전량을 예측하였다. Park과 Kim (2016)은 변동성이 크다는 풍력에너지의 단점을 보완하기 위해 Box-cox변환을 이용하여 데이터의 분산안정화를 한 후에 NNet 신경망 모형을 이용하여 예측하였다. Zhao 등 (2016)은 매우 짧은 기간의 풍력 시계열 예측을 위해 기존 forward forecasting과 backward forecasting을 기반으로 bidirectional forecasting을 이용해 길어지는 예측구간의 정확도를 높이려고 시도했다. Lahouar와 Ben Hadj Slama (2017)는 인공신경망과 Random Forest 모형을 이용하여 풍력 발전량을 예측하였다. 풍속과 같은 외생변수를 추가했을 때, Random Forest 모형이 다른 모형들에 비해 MAE값이 낮은 것을 확인했다. Yu 등 (2017)은 시계열 모형인 ARMA 모형과 Boosting(부스팅 알고리즘)을 사용하여 일별 풍력 발전량을 예측하려고 시도했고, MAE가 57.09로 다른 모형과 비교했을 때 예측력이 가장 높음을 보였다. Hu 등 (2018)은 LSTM을 기반으로 응용한 모형을 사용해 10분 단위와 한 시간 단위의 풍속을 예측했다. Suh 등 (2019)은 시계열의 비선형 패턴을 반영하기 위해 SVM, Hybrid 모형을 적합했고, Halil 등 (2019)은 기계학습 기법인 XGBoost, SVR, Random Forest 중 Random Forest가 가장 예측력이 높다는 결론을 보였다. Zheng와 Wu(2019)는 xgboost 모형을 이용하여 단기간 풍력 발전량을 예측하려 하였다. 또한, Aditya 등 (2020)은 Random Forest와 Decision Tree를 이용하여 풍력 발전량 예측을 시도하였다. Ko 등 (2020)은 “DRNets”이라는 새로운 방법을 bidirectional LSTM(양방향 장기적 기억 신경 네트워크)와 함께 이용하여 풍력 시계열 데이터를 정확히 예측하고자 하였고, Hossain 등 (2020)은 GRU, LSTM, Bi-LSTM, RNN(순환 신경망), NN 신경망 모형을 이용한 Hybrid 모형을 예측 모형으로 사용하였다. Ahmadi 등 (2020)은 10분단위

풍속을 포함한 풍력 데이터를 사용하여 의사결정나무모형을 기반으로 해 Random Forest, AdaBoost, Gradient boosting, XGBoost 등의 모형을 적용하여 MAE 기준 가장 성능이 좋은 모형은 XGBoost 모형이라고 결론지었다. Priya와 Arulanand (2021)은 LSTM 모형을 풍향과 풍속을 포함한 다변량, 풍속만 포함한 단변량 시계열로 나누어 예측하였고 단변량일때의 풍속 예측이 더 정확하다는 것을 보였다.

위에서 언급한 모형들과 같이 최근 연구에서는 시계열 데이터의 예측 정확도를 향상시키기 위해 다양한 측정방법이 적용되고 있다. 본 논문에서는 SVR 모형만을 이용하여 예측 문제에 적용하는 것이 아니라, 최근 추가적으로 진행된 기계학습 기법을 활용한 연구를 적용하여 예측력을 향상시키려 한다. 또한, 풍력 발전량 예측 정확성 개선을 위해 개별 단지의 발전 설비 용량을 고려하여 육상, 해상으로 지역을 나눈 데이터를 활용하였고, 같은 발전기에서 수집되는 풍속과 풍향 데이터를 활용하여 더 나은 예측 성능을 보이고자 한다. 예측오차를 검증하기 위하여 mean absolute error (MAE)와 mean absolute percentage error (MAPE)를 사용한다.

다음 2장에서는 풍력 발전량 예측을 위해 사용한 ARIMA, ARIMAX, SVR, Random Forest, XGBoost 모형에 대하여 소개한다. 3장에서는 본 연구에 활용된 풍력발전 데이터, 풍향과 풍속 데이터에 대하여 설명하고, 4장에서는 위에서 언급한 모형을 적용한 예측 결과를 비교, 분석한다. 5장에서는 결론 및 향후 연구 방향에 대하여 제안할 예정이다.

2. 예측 모형

2.1. Auto-regressive integrated moving average (ARIMA) 모형

ARIMA 모형은 대표적인 시계열 분석 모형으로, 자기회귀모형(AR)과 이동평균모형(MA) 결합된 모형을 결합한 것에 시계열의 비정상성(non-stationary)를 설명하기 위해 차분(difference) 절차를 포함한 모형이다. ARIMA(p,d,q) 모형의 일반적인 형태는 다음과 같다.

$$\begin{aligned} \phi_p(B)(1-B)^d Y_t &= \theta_q(B)\epsilon_t, \\ \phi_p(B) &= 1-\phi_1 B-\cdots-\phi_p B^p, \\ \theta_q(B) &= 1-\theta_1 B-\cdots-\theta_q B^q. \end{aligned} \tag{2.1}$$

여기서 $\phi_p(B)$는 자기회귀모형에 관한 식으로, p는 이 모형의 차수, $\theta_q(B)$는 이동평균모형에 관한식으로, q는 이 모형의 차수, d는 1차 차분이 포함된 정도를 의미한다. ϵ_t는 오차항 또는 백색잡음(white noise)를 의미하며, 평균은 0, 분산은 일정한 값을 가진다. B는 후진연산자(backward shift operator)이다.

2.2. Auto-regressive integrated moving average with eXogeneous variable (ARIMAX) 모형

ARIMAX 모형은 앞의 모형인 ARIMA 모형에 외생변수를 추가한 모형이다. ARIMA의 차수가 p,d,q일 때, k개인 외생변수를 라고 할 때 ARIMAX(p,d,q) 모형은 다음과 같다.

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\epsilon_t + \sum_{i=1}^{k} r_i x_{it}. \tag{2.2}$$

여기서 $\phi_p(B)$와 $\theta_q(B)$는 이동평균모형에 관한 식으로, ARIMA 모형식에서 설명한 식 (2.1)과 같다. ϵ_t는 오차항 또는 백색잡음(white noise)를 의미하며, r_i는 외생변수인 x_{it}의 계수이다. 본 논문에서는 외생변수로 풍향과 풍속, 기온, 습도 등을 고려하였다.

2.3. Support vector regression (SVR) 모형

Support vector machine (SVR)은 기계학습 분야 중 하나로 감독학습에 의한 패턴 인식, 자료 분석을 위한 지도 학습 모형이며, 주로 분류와 회귀 분석, 특이점 판별을 위해 사용되는 모형이다 (Halil 등, 2019). SVM은 과적합 되는 경우가 적고, 신경망보다 사용하기 쉽다는 장점이 있다. 이러한 SVM은 크게 support vector classification (SVC)과 support vector regression (SVR) 두 가지로 나누어진다. 본 논문의 목적은 목표로 하는 값을 최대한 정확히 예측하기 위함이므로 SVM을 일반화한 SVR을 이용한다. SVR은 광범위한 변수 세트에 대한 최적화 전략이 향상되어 선형 회귀, KNN 및 Elastic Net과 같은 다른 알고리즘에 비해 더 나은 성능 예측을 보인다. SVR 함수를 찾기 위한 최적화 문제는 비용함수와 라그랑지 함수, 제약조건을 통해 해결할 수 있다.

SVR 모형은 커널함수를 사용하는데, 본 논문에서는 radial basis function (RBF) 또는 가우시안 커널이라고 불리는 커널을 사용한다. RBF 커널의 매개변수인 를 통해 가중치를 적용하게 된다.

2.4. Random Forest 모형

Random Forest (RF) 모형은 분류, 또는 회귀분석 등에 이용되는 앙상블 학습(ensemble learning)방법의 일종으로, 훈련 과정에서 부트스트랩(bootstrap)방식을 이용해 구성한 의사 결정 트리(decision tree)로부터 분류 또는 가중 평균을 이용해 동작한다. RF 모형은 의사 결정 트리 기반의 알고리즘으로 변수 간의 상호작용 및 비선형성을 다루기 용이하고 이상치에 강해 회귀 문제에서 광범위하게 사용되고 있다 (Dorronsoro 등, 2015). RF의 주요 파라미터는 결정 트리의 갯수와 트리의 최대 깊이 등으로 연구자가 지정해야 하는 값이다. 결정 트리의 갯수를 늘리면 연산량이 증가해 속도가 느려지지만 데이터에 대한 과적합을 피할 수 있다. 반면, 트리의 최대 깊이를 늘리면 과적합이 발생한다. 과적합을 해결하기 위해 적정수준의 파라미터를 결정하게 된다. 본 논문에서는 파라미터를 조정하는 과정을 거쳐 예측력이 우수하도록 모형을 적합하였다.

2.5. XGBoost 모형

XGBoost 모형은 의사결정나무 기반의 알고리즘으로써 그래디언트 부스팅(Gradient Boosting)을 개량한 알고리즘으로서 다양한 연구에서 사용되고 있다 (Huan Zheng 등, 2019). 이 모형의 기반인 부스팅 기법(boosting method)는 비교적 약한 모형들을 여러개 만든 후 결합하여 강한 모형을 만들어내는 방법으로서 초반의 간단한 모형에서 학습 후 발생한 오차를 또 다른 모델로 학습시켜 오차를 점차 줄여나간다. 이러한 과정을 거쳐 모형을 생성한 후 가중치를 부여해 통합함으로써 정확도가 높은 모형을 최종적으로 만들어낸다. 이러한 XGBoost 모형은 병렬 처리를 사용하기에 학습과 분류가 빠르고, 과적합이 잘 일어나지 않는 장점이 있다. 하지만 파라미터의 개수가 많아 복잡하다는 단점이 있다. 본 논문에서는 그 중 5개의 파라미터를 조정하여 예측력이 우수하도록 모형을 적합하였다.

3. 데이터 및 자료 분석

3.1. 데이터 및 분석 방법

본 연구에서 사용된 데이터는 전력거래소에서 제공받은 전남지역과 경북지역의 총 풍력 발전량 데이터이다. 2018년 01월 01일 1시부터 2020년 11월 01일 24시까지의 한시간 단위 데이터를 사용하여 분석을 실시하였다. 2018년 01월 01일부터 2020년 10월 24일까지의 자료는 훈련용 데이터(training data)로 모형 적합에 이용하였고, 2020년 10월 25일부터 2020년 11월 1일까지의 자료는 테스트 데이터(test data)로 모형의 성능을 평가하였다. 전남지역의 풍력발전량은 전라남도 지역 영광풍력발전과 약수풍력, 영광백수풍력발전에서 수

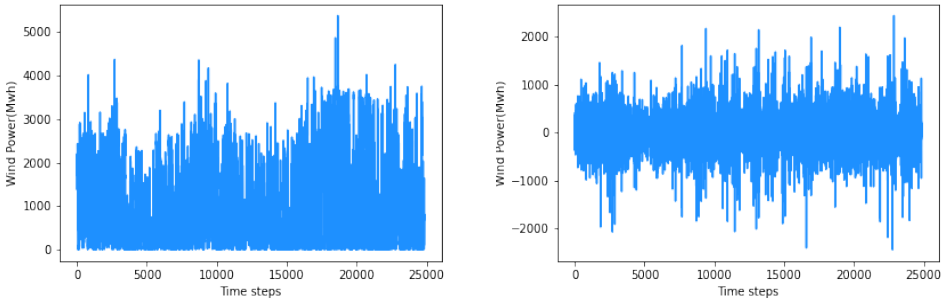


Figure 1: *Original graph and differencing graph of wind power in Gyeongbuk.*

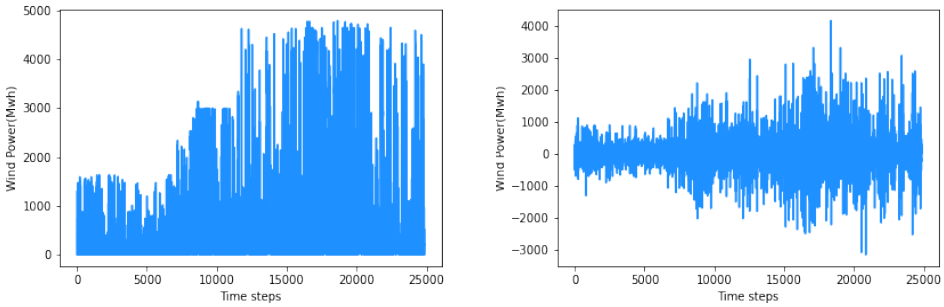
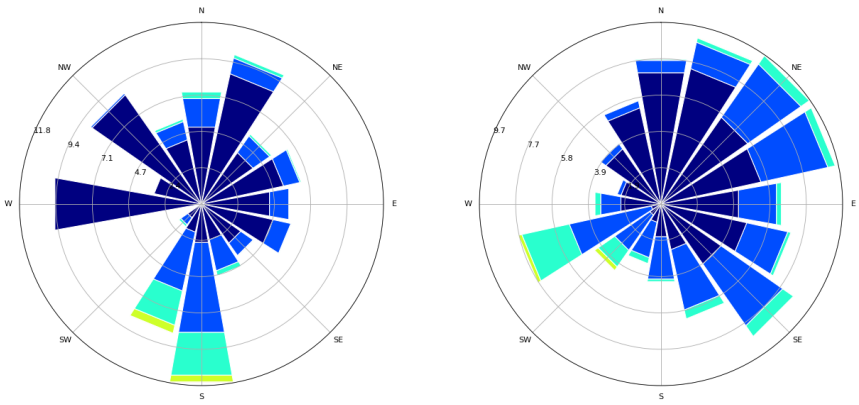


Figure 2: *Original graph and differencing graph of wind power in Jeonnam.*



(a) Gyeongbuk (b) Jeonnam

Figure 3: *Wind Rose of Gyeongbuk and Jeonnam.*

집하여 총량을 계산한 후 평균으로 전처리하였다. 경북지역의 풍력발전량 또한 경상북도 지역의 양구리풍력, 영양풍력, GS 영양풍력에서 수집하여 총량을 계산한 후 평균으로 전처리하였다.

Table 1: Parameter estimation of ARIMA(2, 1, 0) in Gyeongbuk

Parameter	Estimation	S.E
ϕ_1	0.2053	0.0063
ϕ_2	–0.1029	0.0063

Table 2: Parameter estimation of ARIMA(2, 1, 0) in Jeonnam

Parameter	Estimation	S.E
ϕ_1	0.1690	0.0063
ϕ_2	–0.1002	0.0063

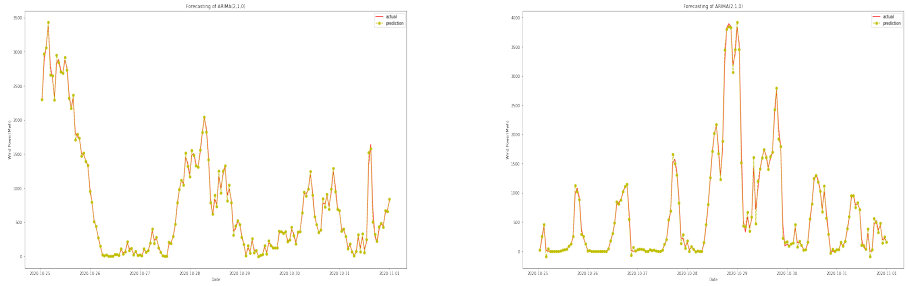


Figure 4: *Forecasting of ARIMA model in Gyeongbuk and Jeonnam.*

외생변수로 사용한 풍향, 풍속, 습도, 기압과 같은 기상 데이터는 전남지역의 경우 무안, 영광, 완도 기상대에서 관측한 시간별 데이터로 풍력발전량 데이터와 같은 기간으로 기상청에서 수집하였고, 경북지역의 경우 영양군과 가까운 청송군, 안동시, 의성군에서 수집하여 총량을 구해 평균으로 전처리하여 데이터로 사용하였다.

Figure 1과 Figure 2는 해상지역과 육상지역의 풍력발전량의 원데이터와 차분을 적용한 데이터이다. 풍력 발전량은 불규칙적인 특성이 있다는 점이 두드러져 분석에 용이하게 하기 위해 차분을 적용, 예측값을 원자료로 변형하여 성능을 평가하였다.

Figure 3은 수집한 풍향데이터를 풍배도(WINDROSE)로 나타낸 것이다. 왼쪽 그림은 경북지역, 오른쪽 그림은 전남지역을 나타낸다. 본 논문에서는 16방위를 이용하여 풍향에 대한 16방위를 분류하여 영향력이 큰 풍속 변수만을 더미변수(dummy variable)로 생성하여 하나의 풍향변수로 결합해 사용하였다. 경북지방에서는 S(남), SSW(남남서), NNE(북북동)의 풍향을 선택하였고, 전남지방에서는 NE(북동), ENE(동북동), WSW(서남서)의 풍향을 선택하였다.

본 논문에서 풍력 시계열 데이터 예측을 위해서 통계 프로그래밍 언어인 R과 Python 3.6을 이용하였으며 시계열 모형 적합 및 예측에는 forecast 패키지를 사용하였고, 머신러닝 모형 적합 및 예측에는 SVR 패키지와 RandomForestRegrssor, xgboost 패키지를 사용하였다.

3.2. 모형 적합 결과

ARIMA, ARIMAX 모형은 Akaike's information criterion (AIC)를 기준으로 그 값이 최소인 모형을 선택하였다. ARIMA 모형의 적합 결과, 경북지역과 전남지역 모두 ARIMA(2, 1, 0) 모형이 최적 모형으로 나타났다. 경북지역과 전남지역에서 해당 모형의 모수추정치는 Table 1, Table 2와 같다.

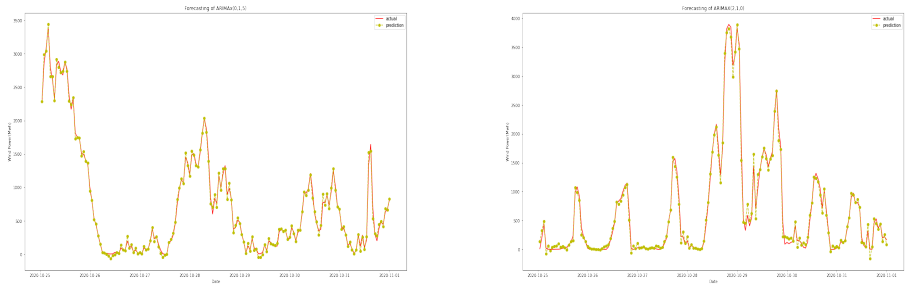


Figure 5: *Forecasting of ARIMAX model in Gyeongbuk and Jeonnam.*

Table 3: Parameter estimation of ARIMAX$(5,1,0)$ in Gyeongbuk

Parameter	Estimation	S.E
θ_1	0.1796	0.0066
θ_2	–0.0869	0.0066
θ_3	–0.0461	0.0069
θ_4	–0.0681	0.0069
θ_5	–0.0474	0.0067
γ_{dir}	0.5378	0.3587
γ_{speed}	–0.7475	0.3801
γ_{humid}	0.0713	0.0577
γ_{hpa}	3.9167	0.7898

Table 4: Parameter estimation of ARIMAX(2,1,0) in Jeonnam

Parameter	Estimation	S.E
ϕ_1	0.1552	0.0064
ϕ_2	–0.1027	0.0063
γ_{dir}	–0.5718	0.4012
γ_{speed}	1.5472	0.4034
γ_{humid}	–2.1020	0.3407
γ_{hpa}	0.2031	0.0522

적합된 모형을 사용하여 풍력 발전량의 실측값과 예측값을 비교한 결과는 Figure 4와 같다. 빨간색 선은 실측값을 나타내며 연두색 선은 예측값을 나타낸다. 대체적으로 예측값이 실측값을 잘 따라가는 경향을 보이지만, 몇몇 값에서 과대 또는 과소 추정이 발생하는 것을 확인할 수 있다.

해당 모형에 대한 모수 추정값은 Table 1, Table 2와 같다.

ARIMAX 모형의 적합 결과, 경북에서는 ARIMAX$(0,1,5)$ 모형, 전남지역에서는 ARIMAX$(2,1,0)$이 최적 모형으로 나타났다. 외생변수로서 경북지역에서 영향력이 큰 풍향인 S(남), SSW(남남서), NNE(북북동)를, 전남지역에서는 NE(북동), ENE(동북동), WSW(서남서)를 더미변수로 만들어 하나의 변수로 합친 풍향변수(γ_{dir})와 풍속(γ_{speed}), 습도(γ_{humid}), 기압(γ_{hpa})을 사용하였으며, AIC가 최소인 모형을 선택하였다. 적합된 모형을 사용하여 풍력 발전량의 실측값과 예측값을 비교한 결과는 Figure 5와 같다. 빨간색 선은 실측값을 나타내며 연두색 선은 예측값을 나타낸다. ARIMA 모형에 비해 예측값이 실측값을 벗어나는 경향을 보인다.

해당 모형에 대한 모수 추정값은 Table 3, Table 4와 같다.

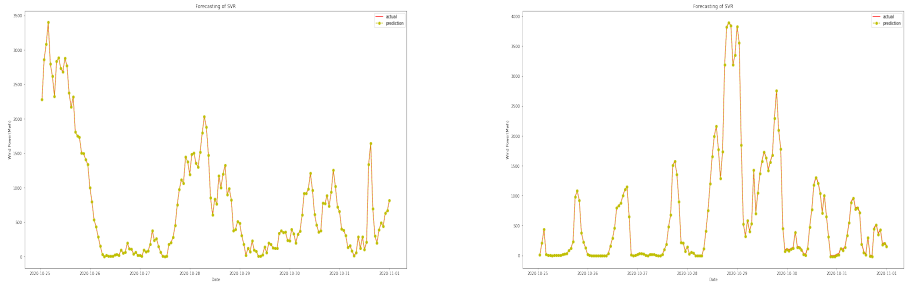


Figure 6: *Forecasting of SVR model in Gyeongbuk and Jeonnam.*

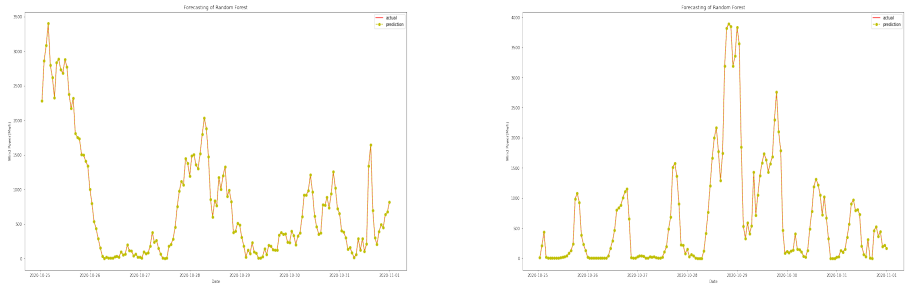


Figure 7: *Forecasting of RF model in Gyeongbuk and Jeonnam.*

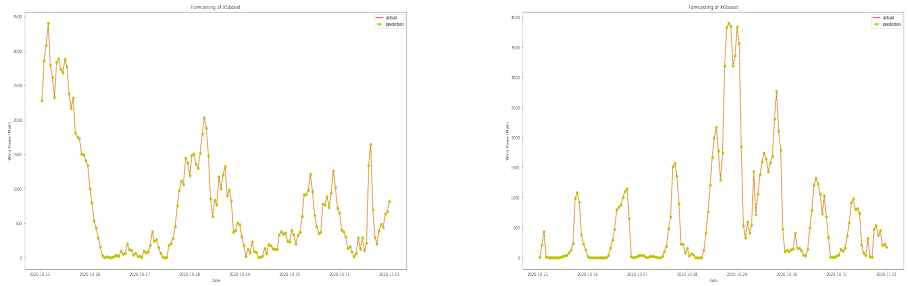


Figure 8: *Forecasting of XGBoost model in Gyeongbuk and Jeonnam.*

SVR 모형의 경우, ARIMAX와 같은 변수를 외생변수로 사용하여 풍력 발전량을 예측하고자 하였다. 본 논문에서는 SVR의 Kernel함수로 radial basis function (RBF) 함수를 사용하고자 하였다. 적합된 모형을 사용하여 풍력 발전량의 실측값과 예측값을 비교한 결과는 Figure 6과 같다. 빨간색 선은 실측값을 나타내며 연두색 점선은 예측값을 나타낸다. ARIMA와 ARIMAX 모형에 비해 훨씬 실측값을 잘 따라가는 경향을 보인다. RF 모형도 동일한 외생변수들을 사용하였으며, 파라미터 값을 조정하여 예측력을 높이고자 하였다. 적합된 모형을 사용하여 풍력 발전량의 실측값과 예측값을 비교한 결과는 Figure 7과 같다. 빨간색 선은 실측값을 나타내며 연두색 점선은 예측값을 나타낸다.

Table 5: Hyperparameter of models

Method	Package	Hyper parmeter
ARIMA	forecast	
ARIMAX	forecast	
SVR	SVR	kernel = 'rbf'
		C = 1000
		gamma = 0.0001
RF	RandomForestRegressor	max_depth = 100
		n_estimators = 150
XGBoost	xgboost	colsample_bytree = 1
		gamma = 0.001
		learning_rate = 0.08
		max_depth = 20
		n_estimators = 1000

Table 6: Result of forecasting accuracy by MAE and MAPE

Method	Gyeongbuk		Jeonnam	
	MAE	MAPE(%)	MAE	MAPE(%)
ARIMA	28.63	13.21	41.23	32.12
ARIMAX	31.24	21.53	51.67	43.45
SVR	0.45	0.38	3.35	2.54
RF	**0.29**	**0.18**	**2.02**	**1.53**
XGBoost	0.96	0.71	2.82	1.04

마지막으로 XGBoost 모형 또한 동일한 외생변수들을 사용하였으며, 파라미터 값을 조정하여 예측력을 높이고자 하였다. 적합된 모형을 사용하여 풍력 발전량의 실측값과 예측값을 비교한 결과는 Figure 8과 같다. 빨간색 선은 실측값을 나타내며 연두색 점선은 예측값을 나타낸다.

3.3. 모형 성능 평가

본 논문에서의 모형 성능 평가를 위해 mean absolute error (MAE)와 mean absolute percentage error (MAPE)를 사용하였다. MAE와 MAPE는 다음과 같이 정의된다.

$$\begin{aligned} \text{MAE} &= \frac{\sum_{i=1}^{n} |Y_t - F_t|}{n}, \\ \text{MAPE} &= \frac{\sum_{i=1}^{n} |\frac{Y_t - F_t}{Y_t}|}{n} \times 100 \end{aligned} \tag{3.1}$$

이 때, n은 예측한 데이터의 개수이며 는 t 시점에서의 실제 값은 t 시점에서의 예측 값을 의미한다. MAE와 MAPE 값이 작을수록 모형의 예측 성능이 우수하다는 것을 의미한다.

본 논문에서는 훈련 데이터를 이용해 적합한 모형으로 168시간 후까지의 값들을 예측하여 모형의 성능을 비교하고자 한다. 모형에 따른 평가지표 비교결과는 다음 Table 6과 같다. 외생변수를 사용하는 ARIMAX, SVR, RF, XGBoost 모형에서는 각 지역에 영향을 끼치는 풍향, 풍속, 습도, 기압을 변수로 일치시켜 예측하였다. 실험환경은 R 3.6.0 버전의 RStudio와 Python 3.6 버전의 각 방법론에서 사용된 초매개변수(hyperparameter)값을 선정하였고, 이는 Table 5과 같다.

예측 결과, 시계열 모형인 ARIMA나 ARIMAX 모형에 비해 기계학습 기법인 SVR, RF, XGBoost 모형의 예측력이 뛰어난 것으로 보인다. 특히 RF모형이 MAE와 MAPE가 각각 경북지역에서는 0.29와 0.18으로, 전남지역에서는 2.02와 1.53으로 가장 작아 우수한 모형으로 나타난다.

4. 결론

본 논문에서는 최근 발전하고 있는 신재생 에너지 중 하나인 풍력에너지의 수요가 중요해짐에 따라 발전량의 정확한 예측을 하고자 하였다.

경북지역과 전남지역의 풍력발전량 데이터, 기상 데이터를 활용하여 시계열 모형, 기계학습 모형을 이용하여 예측을 실시하였다. 본 논문에서는 많은 데이터에서 효율적으로 사용할 수 있는 모형인 Random Forest와 XGBoost 모형의 이용을 제안하였다. 이 때, 기상데이터인 풍속과 풍향은 육상, 해상지역에 따라 크게 영향을 미치는 부분이 각각 다르므로 더미 변수를 분류하여 풍향 외생변수로서 사용하였다. 7일 간의 시간별 예측을 평가한 결과, 예측 모델의 성능은 Random Forest 모형이 MAE 지표가 각각 경북지역에서는 0.29으로, 전남지역에서는 2.02이고 MAPE 지표가 0.18, 전남지역에서는 1.53으로 가장 작아 우수한 모형으로 나타나 최적 모형으로 선정되었다.

본 논문에서는 Random Forest와 XGBoost 모형과 같은 기계학습 단일 모형만을 이용하여 예측 문제에 적용하였지만, 다른 데이터 마이닝 기법이나 적용할 수 있는 외생변수의 갯수, 다양한 데이터 전처리 변형 기법 등을 활용한 연구가 추가적으로 필요하다고 할 수 있을 것이다.

References

Aditya C, Akash S, Ayush K, Karan D, and Neeraj K (2020). Short term wind power forecasting using machine learning techniques, *Journal of Statistics and Management Systems*, **23.1**, 145–156.

Ahmadi A, Nabipour M, Mohammadi-ivatloo B, Rho SM, and Piran J (2020). Long-term wind power forecasting using tree-based learning algorithms, *IEEE Access*, **8**, 151511–151522.

Anastasiades G and McSharry P (2013). Quantile forecasting of wind power using variability indices, *Journal of Applied Meteorology and Climatology*, **6.2**, 662–695.

Brown BG, Richard WK, and Allan HM (1984). Time series models to simulate and forecast wind speed and wind power, *Journal of Applied Meteorology and Climatology*, **23.8**, 1184–1195.

Cadenas E and Wilfrido R (2010). Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model, *Renewable Energy*, **35.12**, 2732–2738.

Catalão, João Paulo da Silva, Hugo Miguel Inácio Pousinho, and Vıctor Manuel Fernandes Mendes (2011). Short-term wind power forecasting in Portugal by neural networks and wavelet transform, *Renewable energy*, **36.4**, 1245–1251.

Halil D, Ahmet SD, Alper E, and Murat G (2019). Wind power forecasting based on daily wind speed data using machine learning algorithms, *Energy Conversion and Management*, **198**, 111823.

Hossain A, Chakrabortty RK, Elsawah S, and Ryan MJ (2020). Hybrid deep learning model for ultra-short-term wind power forecasting, *2020 IEEE International Conference on Applied Superconductivity and Electromagnetic Devices (ASEMD)*.

Hu YL and Liang C (2018). A nonlinear hybrid wind speed forecasting model using LSTM network, hysteretic ELM and differential evolution algorithm, *Energy conversion and management*, **173**, 123–142.

Ko MS, Lee KS, Kim JK, Hong CW, Dong ZY, and Hur K (2020). Deep concatenated residual network with

bidirectional LSTM for one-hour-ahead wind power forecasting, *IEEE Transactions on Sustainable Energy*, **12**, 1321–1335.

Lahouar A and Ben Hadj Slama J (2017). Hour-ahead wind power forecast based on random forests, *Renewable energy*, **109**, 529–541.

Li S, Peng W, and Lalit G (2015). Wind power forecasting using neural network ensembles with feature selection, *IEEE Transactions on sustainable energy*, **6.4**, 1447–1456.

Liu H, Erdem E, and Shi J (2011). Comprehensive evaluation of ARMA–GARCH (-M) approaches for modeling the mean and volatility of wind speed, *Applied Energy*, **88.3**, 724–732.

Park SH and Kim S (2016). A study on short-term wind power forecasting using time series models, *The Korean Journal of Applied Statistics* , **29.7**, 1373–1383.

Pinson P (2012). Very-short-term probabilistic forecasting of wind power with generalized logit–normal distributions, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61.4**, 555–576.

Priya B and Arulanand N (2021). Univariate and multivariate models for Short-term wind speed forecasting. *Materials Today: Proceedings*, 2214–7853.

Suh YM, Son HG, and Kim S (2018). Solar radiation forecasting by time series models, *The Korean Journal of Applied Statistics* , **31.6**, 785–799.

Yu J, Chen X, Yu K, and Liao Y (2017). Short-term wind power forecasting using hybrid method based on enhanced boosting algorithm, *Journal of Modern Power Systems and Clean Energy*, **5.1**, 126–133.

Zheng H and Wu Y (2019). A xgboost model with weather similarity analysis and feature engineering for short-term wind power forecasting, *Applied Science*, **9.15**, 3019.

Zeng J and Wei Q (2011). Support vector machine-based short-term wind power forecasting, *2011 IEEE/PES Power Systems Conference and Exposition. IEEE*, 1–8.

Zhao Y, Ye L, Li Z, Song X, Lang Y, and Su J (2016). A novel bidirectional mechanism based on time series model for wind power forecasting, *Applied Energy*, **177**, 793–803.

Received June 17, 2021; Revised August 10, 2021; Accepted August 11, 2021

시계열 모형과 기계학습 모형을 이용한 풍력 발전량 예측 연구

박수진[a], 이진영[a], 김삼용[1,a]

[a]중앙대학교 응용통계학과

요 약

빠르게 발전하고 있는 재생에너지 중 하나인 풍력에너지는 기후변화 대응에 맞추어 개발 및 투자가 이루어지고있다. 신재생에너지 정책과 발전소 설치가 추진됨에 따라 국내 풍력 보급이 점차 확대되어 수요를 정확히 예측하기 위한 시도들이 확대되고 있다. 본 논문에서는 전남지역과 경북지역의 풍력 발전량 예측을 위하여 시계열 기법인 ARIMA, ARIMAX 모형과 기계학습 모형인 SVR, Random Forest, XGBoost 모형들을 비교 분석하였다. 모형의 예측 결과를 비교하기 위한 지표로서 mean absolute error (MAE)와 mean absolute percentage error (MAPE)를 사용하였다. 2018년 1월 1일부터 2020년 10월 24일까지의 시간별 원 데이터를 차분한 후 모형을 훈련시켜 2020년 10월 25일부터 2020년 10월 31일까지의 168시간에 대한 풍력 발전량을 예측하였다. 모형의 예측력 비교 결과, Random Forest와 XGBoost 모형이 전남지역, 경북지역 순으로 가장 우수한 성능을 보였다. 향후 연구에서는 기계학습뿐 아니라 최근 활발한 연구가 이루어지는 데이터 마이닝 기법 기반의 풍력 발전량 예측을 시도할 것이다.

주요용어: 기계학습, 풍력 데이터, Random Forest, XGBoost, 시계열 모형

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A1B01014954).

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 통계학과. E-mail:sahm@cau.ac.kr