The Korean Journal of Applied Statistics
2023, Vol. 36, No. 1, 49–62

DOI: https://doi.org/10.5351/KJAS.2023.36.1.049
Print ISSN1225-066X / Online ISSN 2383-5818

A study on solar radiation prediction using medium-range weather forecasts

Sujin Park[a], Hyojeoung Kim[a], Sahm Kim[1,a]

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

Abstract

Solar energy, which is rapidly increasing in proportion, is being continuously developed and invested. As the installation of new and renewable energy policy green new deal and home solar panels increases, the supply of solar energy in Korea is gradually expanding, and research on accurate demand prediction of power generation is actively underway. In addition, the importance of solar radiation prediction was identified in that solar radiation prediction is acting as a factor that most influences power generation demand prediction. In addition, this study can confirm the biggest difference in that it attempted to predict solar radiation using medium-term forecast weather data not used in previous studies. In this paper, we combined the multi-linear regression model, KNN, random fores, and SVR model and the clustering technique, K-means, to predict solar radiation by hour, by calculating the probability density function for each cluster. Before using medium-term forecast data, mean absolute error (MAE) and root mean squared error (RMSE) were used as indicators to compare model prediction results. The data were converted into daily data according to the medium-term forecast data format from March 1, 2017 to February 28, 2022. As a result of comparing the predictive performance of the model, the method showed the best performance by predicting daily solar radiation with random forest, classifying dates with similar climate factors, and calculating the probability density function of solar radiation by cluster. In addition, when the prediction results were checked after fitting the model to the medium-term forecast data using this methodology, it was confirmed that the prediction error increased by date. This seems to be due to a prediction error in the mid-term forecast weather data. In future studies, among the weather factors that can be used in the mid-term forecast data, studies that add exogenous variables such as precipitation or apply time series clustering techniques should be conducted.

Keywords: time series, random forest, K-means clustering, solar radiation forecasting, probability density function

1. 서론

화석연료의 지속적인 소비와 함께 에너지 고갈과 환경오염 문제는 점점 더 심각해지고 있다. 이러한 환경 문제들을 해결하고 지속 가능한 발전을 이룰 수 있는 해결책을 찾는 것이 필요한 시점이다. 이에 발맞추어 재생 가능 에너지는 전 세계적으로 많은 관심을 끌었고 최근 몇 년 동안 빠르게 개발되고 있다 (Ozturk, 2013). 최근 그린뉴딜 정책이 수립됨에 따라, 정부는 2025년까지 신재생에너지인 태양광, 풍력설비의 목표를 48.7GW로 제시하였다. 특히 코로나-19에도 불구하고 재생에너지 설비 연간 목표치는 3년 연속으로 달성하고 있다. 특히

This research was supported by the Chung-Ang University Graduate Research Scholarship Grants in 2022.
[1] Corresponding author: Department of Statistics, Chungang University, 84 Heuksukro, Dongjak-Gu, Seoul 06974, Korea. E-mail: sahm@cau.ac.kr

Published 28 February 2023 / journal homepage: http://kjas.or.kr

태양광 발전량은 높은 설비 비중을 차지하고 있으며 이에 따라 전력수급의 안정성을 위한 태양광 발전량의 정확한 예측 연구가 필요한 시점이다. 또한 일사량이 태양광 발전량을 예측하기 위해 가장 중요한 요소이기 때문에 높은 예측 정확도의 일사량 수요 예측 연구가 증가하고 있는 추세이다. 다시 말하면, 일사량 예측은 태양광 발전량 예측의 핵심 내용이라고 해도 무방하다. 이에 따라, 일사량을 정확하게 예측하기 위한 노력이 세계적으로 진행 중이며 머신러닝과 딥러닝 모형의 사용한 연구가 활발하게 이루어지고 있다 (Kim 등, 2022). 하지만 초단기예보와 단기예보 같은 기상 예보 자료를 이용한 연구는 지속되고 있는 반면, 중기예보 기상 자료를 이용한 태양광 일사량 예측 연구는 부족한 실정이다. 중기 예보와 같은 장기적인 예보 데이터를 적용한 일사량 예측 시스템이 개발된다면 신재생에너지 전력수급 안정화에 큰 도움이 될 것으로 보인다.

Yadav와 Behera (2014)은 일사량 값을 예측하기 위하여 기온과 습도, 풍속, 풍향, 이슬점 온도와 기압과 같은 변수들을 추가하여 recurrent neural network (RNN)과 Wavelet 변형을 적용하였다. 그 결과, Wavelet 변형을 이용한 기법이 MAE, RMSE 9.62%, 14.96%로 우수한 것을 확인하였고, Alzahrani 등 (2017)은 태양광 발전량 예측에 일사량의 정확한 예측이 매우 큰 영향을 끼친다고 판단하여 feedforward neural networks (FNN), RNN, deep recurrent neural networks (DRNN) 그리고 support vector regression (SVR)과 long short-term memory (LSTM)을 일사량 예측 모형으로 이용하였고, 검증 결과 LSTM의 경우가 root mean squared error (RMSE) 값이 0.086으로 가장 우수한 예측 성능을 보유한 모형이라는 것을 확인할 수 있었다. Khosravi 등 (2018)은 일사량을 정확히 예측하기 위해 태양 복사 조도를 예측하는 것이 중요한 요소라고 판단하여, 기압, 온도, 풍속 및 상대 습도를 사용하여 선제적으로 태양 복사 조도를 예측하고, 일사량의 시계열 예측을 진행하였다. 모형으로는 multilayer feed-forward neural network (MLFFNN), radial basis function neural network (RBFNN), SVR, fuzzy inference system (FIS) 그리고 adaptive neuro-fuzzy inference system (ANFIS)을 제안하였다. 결과적으로 조도 예측에는 SVR과 MLFFNN 모형이, 일사량 예측에는 SVR, MLFFNN, ANFIS 모형이 검증 데이터 세트에 대해 0.95 이상의 상관계수를 보여주었다. Al-Hajj 등 (2019)은 태양광 일사량의 예측을 위하여 앙상블 방법을 제안하였다. SVR 모형에 다양한 커널을 적용해보고 그 결과를 decision tree (DT), K-nearest neighbors (K-NN), MLP 방법론을 결합하여 어떠한 모형이 가장 예측성능이 우수한지 확인하였다. 그 결과, SVR 단일 모형으로 예측한 것보다 SVR과 MLP를 결합한 앙상블 모델이 mean absolute error (MAE) 0.03926의 높은 성능을 보여주었다. Karasu와 Altan (2019)은 random forest (RF) 모형의 적합한 변수를 선정하기 위해 기온, 기압, 풍속 그리고 습도와 같은 기상 변수를 수집 및 조합함으로써 다음날의 일사량 값을 추정하는 연구를 진행하였다. 선택된 변수들로 예측한 일사량 값은 10-fold cross validation 방법을 통해 검증되었으며, 0.9963의 높은 설명력을 나타낸 것을 확인하였다. Fan 등 (2020)은 support vector machine (SVM), M5 model tree (M5Tree), RF, extreme gradient boosting (XGBoost) and gradient boosting with categorical features support (CatBoost) 모형을 적용하였다. 외생 변수로는 최저 및 최고 기온, 일사 시간, 습도 등이 사용되었다. 모형들과 기상 변수들을 다양하게 조합해 본 결과, RF 모형이 가장 작은 통계적인 오류를 산출해내었고 그 뒤로는 XGBoost, CatBoost, M5Tree and SVM 모형이 뒤를 이었다. SVM 모델은 훈련 시간이 길다는 단점이 존재하였고, 종합적으로 CatBoost 모형이 여러 지역에서 지역적, 기후적 특성을 가장 잘 보완한 모형이라고 할 수 있었다. Bamisile 등 (2022)은 polynomial regression, RF, SVR 그리고 딥러닝 모형으로는 artificial neural network (ANN), convolutional neural network (CNN), RNN을 이용하여 네 지역의 태양광 일사량을 정확히 예측하고자 하였고, 그 결과 지역마다 편차가 있지만 polynomial regression 모형이 대부분 가장 우수한 성능을 나타내는 것을 확인하였다. Demir와 Citakoglu (2022)은 5가지의 머신러닝 기반 방법론을 사용하여 월별 일사량 추정 연구를 진행하였다. 모형은 SVR, LSTM, Gaussian process regression (GPR), extreme learning machines (ELM) 그리고 K-NN을 사용하였다. 또한 일사량에 영향을 끼치는 기상데이터 등의 입력 변수를 탐색한 후, 영향을 끼치는 변수들을 이용하여 모형에 적합하였다. 그 결과, LSTM 및 GPR 모형이 통계적으로 유효하며 일사량의 정확한 예측에 도움이 될 것이라 주장하였다.

최근에는 클러스터링 기법을 이용하여 데이터를 구분한 뒤, 각 클러스터별 예측을 진행하는 선행 연구들

도 활발하게 진행되고 있다. Benmouiza와 Cheknane (2013)은 K-Means 클러스터링과 nonlinear autoregressive neural network (NAR) 기법을 이용하여 지역을 구분한 후, 클러스터별 예측을 NAR로 진행하였다. K-Means의 적절한 K의 수는 실루엣 기법을 이용하여 산정하였다. Ghofrani 등 (2016)은 K-Means와 Multilayer perceptron neural network (MLPNN) 을 적용하여 각 날짜의 시간별 일사량 평균을 기준으로 날짜들을 구분하였다. 그리고 각 클러스터별 예측을 진행하였고, 그 결과 mean squared error (MSE) 0.0044로 우수한 예측 정확도를 확인할 수 있었다. Jiménez-Pérez와 Mora-López (2016)은 스페인의 시간 단위 일사량 데이터를 사용하여, 클러스터링 기법인 decision tree (DT)와 support vector machine (SVM)을 이용, 시간별 일사량 분포가 비슷한 날짜들을 나누었다. 그리고 각 클러스터의 중심을 비교한 다음, 날짜들의 클러스터를 예측하는 연구를 진행하였다. 그 결과, SVM으로 클러스터를 분류하고, SVR로 예측하는 방법이 MAE 10.5%로 가장 우수한 결과를 보이는 것을 확인하였다. Zhang 등 (2021)은 일사량 예측을 위하여 기온과 습도, 풍속을 이용하였고, fuzzy C-means (FCM) 알고리즘과 클러스터 개수를 정하기 위한 실루엣 계수 공식, ANN과 선형회귀모델을 사용하였다. 정규화된 데이터를 이용하였으며 1월부터 12월까지 월별 데이터를 클러스터 1, 2로 나누어 예측을 진행하였다. 그 결과, ANN과 클러스터링 모형을 사용한 방법이 훨씬 높은 정확도를 보이는 것을 확인하였다. 현재까지 태양광 일사량의 단기예측 또는 단기예보 데이터를 이용한 연구는 다양하게 선행되었다. 하지만, 3일부터 10일까지의 기상예보 데이터인 중기예보 기상데이터를 이용한 일사량 예측 연구는 부족한 상황이다. 그에 따라, 본 논문은 중기예보 기상데이터를 이용하여 일사량을 예측 연구를 진행한다. 본 논문의 2장에서는 연구에서 사용된 예측 및 클러스터링 모형에 대해 소개한다. 클러스터링 기법으로는 K-means와 적절한 K 개수 선정을 위한 Silhouette coefficient 방법 그리고 예측 모형으로는 다중선형회귀모형과 random forest, K-NN 마지막으로 SVR 모형이 일사량 예측에 사용되었다. 3장에서는 중기예보 기상 데이터와 관측 데이터의 설명과 그리고 일사량 예측 결과에 대해 말한다. 마지막으로, 4장에서는 결론과 추후 연구에 대한 방향성을 제시한다.

2. 예측 모형

본 연구에서는 클러스터링 기법과 머신러닝 회귀 예측 모형을 결합하여 분석을 진행한다. 먼저 머신러닝 회귀 모형으로 하루 총 일사량 값을 예측한 후, 기후 인자들을 이용하여 날짜들을 클러스터링하여 일사량의 시간별 확률 밀도 함수를 계산하는 방법을 제안한다.

2.1. 클러스터링

2.2. K-means 모형

K-means 모형은 클러스터링을 수행하는 가장 빠르고 간단한 비지도 학습 알고리즘 중 하나이다. 이 모형은 주어진 데이터를 K 개의 클러스터로 분류하는 것으로 구성된다. 처음으로는 아래 방정식을 이용하여 데이터의 각 점 $x_i^{(j)}$과 클러스터 중심 c_j 사이의 유클리드 거리를 계산하여 가장 가까운 클러스터의 중심과 연결한다.

$$\left\| x_i^{(j)} - c_j \right\|. \tag{2.1}$$

다음으로는 새로운 K 중심의 위치를 다시 계산한다. 중심이 더이상 움직이지 않을 때까지 첫 번째와 두 번째 단계를 반복하면 대상 데이터들이 최적의 중심을 탐색하고, 각 데이터 포인트들이 어떠한 그룹에 속하는지 탐색한다 (Benmouiza와 Cheknane, 2013).

2.3. Silhouette coefficient (SC)

Silhouette coefficient는 클러스터링을 평가하는 척도로써, Peter Rousseeuw (1987)에 의해 제안되었다. 이 계수는 군집 내 비유사성('within' dissimilarities)은 작고, 군집 간 비유사성('between' dissimilarities)은 큰, 즉 서로 잘 구분이 되는지 확인하기 위한 척도이다. 이 값은 각 데이터 포인트와 주위 데이터 포인트들과의 거리 계산을 통하여 값을 구하며, 본 연구에서는 K-means 클러스터링과 KNN의 적절한 K 개수를 선정하기 위해 실루엣 계수 방법을 사용하였다. i 번째 데이터 에 대한 실루엣 계수 값은 아래의 식으로 정의된다.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{2.2}$$

$a(i)$는 클러스터 내 데이터 응집도(cohesion)를 나타내는 값이고, $b(i)$는 클러스터간 분리도(seperation)을 나타내는 값이다. 실루엣 값인 $s(i)$은 −1과 1 사이에서 변할 수는 있지만, 약 0일 때에는 $a(i)$와 $b(i)$가 거의 동일한 값임을 의미한다. 즉, 클러스터의 개수가 최적화되어있다면 $s(i)$ 값은 1에 가까운 값이 된다 (Tambunan 등, 2020).

2.4. 예측

2.5. Multiple linear regression (MLR) 모형

다중 선형 회귀 분석 모형은 단순 선형 회귀 분석 모형을 확장한 모형으로써, 둘 이상의 설명 변수를 포함한다. 다중 선형 회귀 방정식은 단순 선형 회귀 방정식과 같은 형태를 갖지만 더 많은 항을 갖는다는 차이점이 있다 (Tranmer와 Elliot, 2008). MLR 모형은 다음과 같은 식으로 나타내어진다 (Abuella와 Chowdhury, 2015).

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon. \tag{2.3}$$

Y는 반응 변수이고, X_k는 k번째 예측 변수이다. β_k는 회귀 계수이며, ϵ는 Y의 변동성을 나타낸다. 반응 변수는 태양광 일사량이며 예측 변수들은 중기예보 기상 데이터에서 사용할 수 있는 기상 변수들이다 (Hong, 2010).

2.6. K-nearest neighbors (KNN) 모형

KNN 회귀 알고리즘은 예측된 데이터 포인트와 알려진 데이터 포인트 사이의 거리를 계산하여 가장 가까운 K 개의 데이터 값 $\{y_1, y_2, \ldots, y_K\}$을 평균내어 값을 예측하는 방법이다. 여기서 y_1은 예측된 데이터에서 가장 가까운 데이터 포인트를 나타내며, y_2는 두 번째로 가깝게 알려진 데이터 포인트를 나타낸다 (Fan 등, 2019). 이 모형은 게으른 학습 접근법이다. 즉, 다시 말하면 훈련 데이터에 대해 모형을 명시적으로 구축하지 않고, 검증 데이터 샘플에서 보이지 않는 기록이 주어지면 훈련 데이터에서 가장 가까운 K 개의 데이터 일치 항목을 탐색한다. 예측 변수는 태양광 일사량으로 연속적인 값이기 때문에, 이러한 K 개의 가장 가까운 데이터들의 평균은 보이지 않는 검증 데이터 샘플에 대한 예측이 이루어진다 (Jawaid와 NazirJunejo, 2016). 이 알고리즘의 순서는 다음과 같이 진행된다. 먼저 K 개수를 선정한 다음, 훈련 데이터 포인트와 출력 값으로 구성된 데이터 테이블을 구성한다. 그리고 검증 데이터 포인트로 훈련 데이터들과의 유클리드 거리를 계산한 후, 계산된 거리를 오름차순으로 정렬한다. 유클리드 거리 계산법은 다음과 같다.

$$d(x, y) = \sqrt{\sum_{i=1}^{K} (x_i - y_i)^2}. \tag{2.4}$$

$x = (x_1, x_2, \ldots, x_K)$와 $y = (y_1, y_2, \ldots, y_K)$는 데이터의 좌표를 나타낸다. 본 연구에서 K 개수는 계절별로 다르게 설정하였으며, 클러스터링 적용에 앞서 실루엣 계수 방법을 이용하여 최적의 K 개수를 선정하였다.

Table 1: Latitude, longitude and number of data by each region

Region	Lat	Log	Count
Seoul	37.57	126.97	1826
Incheon	37.48	126.62	1826
Suwon	37.95	127.75	1824
Chuncheon	35.88	128.65	1822
Daegu	36.37	127.37	1820
Daejeon	37.26	126.98	1786

2.7. Support vector regression (SVR) 모형

Support vector machine 모형은 Vapnik (1999)에 의해 제안되었다. 이 모형은 입력 데이터를 기반으로 알고리즘 구조를 최적화하는 분류 및 회귀 기법이다. 본 연구에서는 연속적인 일사량 값의 예측을 위해 회귀 기법을 적용한 support vector regression 모형을 적용하였다. 이 모형은 선형 모형인 $W \cdot X + b$에 $S = \{(X_p, y_p) : p = 1, \ldots, N\}$을 적합한 후, 다음과 같은 기준 함수인 라그랑지 함수를 최적화하는 과정을 진행한다 (Gala 등, 2013).

$$\begin{aligned} \min_{W,b,\xi} \frac{1}{2} \|W\|^2 + C \sum_i (\xi_i + \xi_i^*), \\ s.t W \cdot X_i + b - y_i \geq -\xi_i - \epsilon, \\ W \cdot X_i + b - y_i < \xi_i^* + \epsilon, \\ \xi_i, \xi_i^* \geq 0, \\ K(x_i, x_{i'}) = \left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2 \right). \end{aligned} \tag{2.5}$$

2.8. Random forest (RF) 모형

Random forest 모형은 Breiman (1984)에 의해 제안되었다 (Breiman, 2001). 의사 결정 트리 학습에서 회귀 및 분류에 가장 널리 사용되는 모형이다. 매우 효율적이며 동시에 다른 회귀 모형보다 정확도가 우수하다는 장점을 지닌다. 이 모형은 훈련 단계에서 많은 수의 의사 결정 트리를 구성한다. 다수의 의사 결정 트리를 개발한 후, 모형의 출력은 모든 개별 트리의 출력 값을 평균화하여 얻는다. 단일 트리를 훈련하기 위해 학습자 bagging 알고리듬은 랜덤 포레스트 모형에서 사용된다. 여기서, bagging은 반복적으로 훈련 세트의 bootstrap 샘플을 선택하고 이러한 샘플의 지니 계수(Gini index)을 사용하여 $t_b(x)$ 트리를 적합한다. 훈련 과정 이후, 예측 값은 다음 식과 같이 모든 회귀 트리의 예측 결과를 평균화하여 계산한다 (Srivastava 등, 2019).

$$y = \frac{1}{B} \sum_{b=1}^{B} t_b(x). \tag{2.6}$$

즉, 단일 트리보다 다양한 트리를 모델링함으로써 이 모형은 더 정확한 예측 성능을 제시한다.

3. 데이터 및 자료 분석

3.1. 데이터 및 분석 방법

본 연구에서 사용된 데이터는 한국전력거래소에서 제공받은 서울, 인천, 수원, 춘천, 대구, 대전의 설비용량 대비 일사량 데이터이다. 각 지역의 위도와 경도는 Table 1과 같다. 2017년 3월 1일 0시부터 2022년 2월 28

Table 2: Mean and variance of each month

Month	Lowest temperature		Highest temperature		Morning cloud		Afternoon cloud		ei		Solar radiation	
Statistic	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
January	−5.66	26.87	3.51	22.50	1.42	0.16	1.77	0.44	16.94	1.26	8.93	9.91
Feburary	−3.78	22.85	6.00	28.64	1.45	0.17	1.82	0.44	21.72	2.76	11.95	16.01
March	2.32	12.83	13.13	18.99	1.62	0.25	1.95	0.58	28.64	4.43	15.24	30.01
April	7.34	13.48	18.39	20.24	1.62	0.26	2.00	0.54	34.90	2.15	18.48	51.02
May	13.03	9.60	23.73	16.25	1.73	0.25	2.11	0.56	38.70	0.59	19.83	58.73
June	18.28	5.20	27.72	10.15	1.99	0.27	2.38	0.57	41.38	0.03	20.51	50.17
July	22.82	6.51	29.94	13.21	2.14	0.21	2.66	0.42	40.48	0.45	16.70	58.37
August	23.35	6.67	30.44	12.12	2.15	0.18	2.69	0.31	36.85	2.05	15.54	45.62
September	17.58	9.41	25.86	5.58	1.81	0.21	2.27	0.47	30.58	3.60	14.47	33.48
October	10.09	23.46	19.97	13.35	1.64	0.24	2.01	0.51	23.77	4.20	12.33	18.98
November	2.82	19.72	12.71	22.99	1.59	0.25	1.88	0.54	18.00	1.82	9.09	12.94
December	−4.15	23.39	4.66	23.65	1.41	0.17	1.72	0.42	15.46	0.23	7.99	8.21

Table 3: Mean and variance of each region

Month	Lowest temperature		Highest temperature		Morning cloud		Afternoon cloud		ei		Solar radiation	
Statistic	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
Daegu	9.87	97.61	19.58	93.01	1.70	0.28	2.12	0.58	29.36	80.76	14.90	44.22
Daejeon	9.31	110.7	19.11	97.85	1.71	0.27	2.12	0.55	29.31	83.58	15.36	57.99
Chuncheon	5.89	134.13	17.55	117.0	1.77	0.26	2.14	0.57	28.67	89.70	14.21	50.67
Seoul	9.26	116.8	17.94	114.0	1.71	0.30	2.11	0.61	28.83	88.40	13.72	49.43
Suwon	8.47	116.4	17.97	108.1	1.69	0.28	2.06	0.58	28.94	87.09	13.62	44.53
Incheon	9.74	107.6	16.39	103.1	1.72	0.29	2.09	0.60	28.88	88.15	13.79	48.54

Table 4: Sunrise and sunset time by season

Season	Month	Sunrise time	Sunset time
Spring	3,4,5	7AM	7PM
Summer	6,7,8	6AM	8PM
Autumn	9,10,11	7AM	7PM
Winter	12,1,2	8AM	6PM

일 23시까지의 한 시간 단위 데이터를 이용, 중기예보 데이터는 일별로 산출되기에 데이터를 일별 데이터로 변환하여 분석을 실시하였다. 추가적으로 일사량을 예측하기 위해서는 대기권 밖 일사량이 큰 영향을 미치기 때문에 지점의 위도와 경도를 이용하여 대기권 밖 일사량을 계산하여 외생 변수로 사용하였다. 또한 대기권 밖 일사량, 일사량 값이 새벽에도 0이 아닌 경우가 존재하는데, 이러한 부분을 보완하기 위하여 계절별 일출, 일몰 시간을 반영하여 새벽시간대에는 값이 산출되지 않도록 전처리하였다. 데이터에 적용한 계절별 일출몰 시간과 사용한 지역별 데이터 수는 Table 4와 같고, Tables 2와 3은 지역별, 월별 변수 통계량을 나타내었다.

본 연구의 순서는 크게 관측 데이터를 이용한 모형 선정과 예보 데이터를 이용한 예측 부분으로 나누어지고, 아래설명 및 Figure 1과 같다.

(1) 일별 데이터에서 기상 변수들과 대기권 밖 일사량을 이용하여 K-means 클러스터링을 수행한다. 클러스

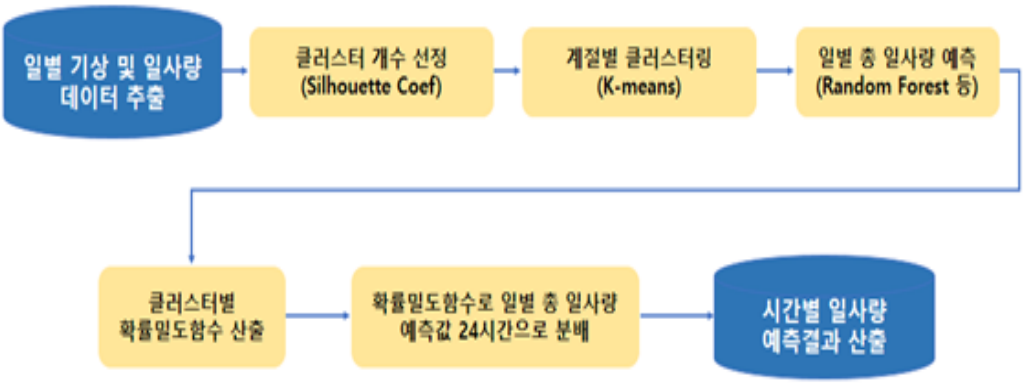

Figure 1: *Data analysis progress.*

터링에 앞서, 최적의 클러스터 개수는 SC 방법을 통해 선정한다.

(2) 일별 데이터를 이용, 기상 변수인 최저/최고 기온, 오전/오후 하늘 상태를 운량값으로 변환, 대기권 밖 일사량으로 하루 총 일사량 값을 예측한다. 모형으로는 예측 모형 부분에서 설명한 MLR, KNN, SVR, RF 모형을 적용한다.

(3) 클러스터별 시간별 일사량 값의 확률 밀도 함수를 계산한다.

(4) 계절별, 클러스터별 확률 밀도 함수를 이용하여 예측한 하루 총 일사량 값을 24시간으로 분배하여 산출한다.

따라서, 본 연구에서는 새롭게 일별 일사량 예측 모형과 K-means 그리고 확률밀도함수를 결합한 방법론을 제시한다. 중기예보 데이터를 이용한 예측은 일자별로 예측이 이루어진다. 예를 들면, 현재 시점의 3일 이후 예보 데이터로 3일 이후 일사량 값을 예측한다. 따라서, 3일 10일 이후 일사량 예측은 관측 데이터로 훈련한 모형에 예보 데이터를 적합하여 산출한다. 중기예보에서 8일부터 10일 이후의 예보 데이터 중 하늘 상태는 오전, 오후가 아닌 일별로 통합되어 나오므로, 본 연구에서는 해당 시점은 오전, 오후의 하늘 상태가 같다는 가정하에 연구를 진행한다. 마지막으로, 예측 모형 훈련과 예측 성능을 확인하기 위하여 2017년 3월 1일부터 2021년 2월 28일까지를 훈련 데이터로 2021년 3월 1일부터 2022년 2월 28일까지를 검증 데이터로 분할하였다. 본 논문에서 태양광 일사량 시계열 데이터 예측을 위해서 통계 프로그래밍 언어인 Python 3.6을 이용하였으며 머신러닝 모형 적합 및 예측에는 KMeans, silhouette_score, LinearRegression, KNeighborsRegressor, SVR 패키지와 RandomForestRegrssor 패키지를 사용하였다. 추가적으로 이 모형들은 초기값에 따라 결과가 크게 달라질 수 있으므로 본 연구에서는 random_state 설정과 같은 방법을 사용하여 초기값이 고정되도록 하였다.

3.2. 모형 성능 평가

본 논문에서의 모형 성능 평가를 위해 MAE (mean absolute error)와 RMSE (root mean square error)를 사용하였다. 일반적으로 MAPE (mean absolute percentage error)는 모델을 평가하는 데 널리 사용되지만, 일사량이 0인 경우가 많아 MAPE 계산을 적용하기 어렵다는 단점이 있다. 따라서, 본 연구에서는 다음과 같이 정의되는 평균 절대 오차(MAE)와 루트 평균 제곱 오차(RMSE)의 척도로 정확도를 평가하였다. MAE와 RMSE는

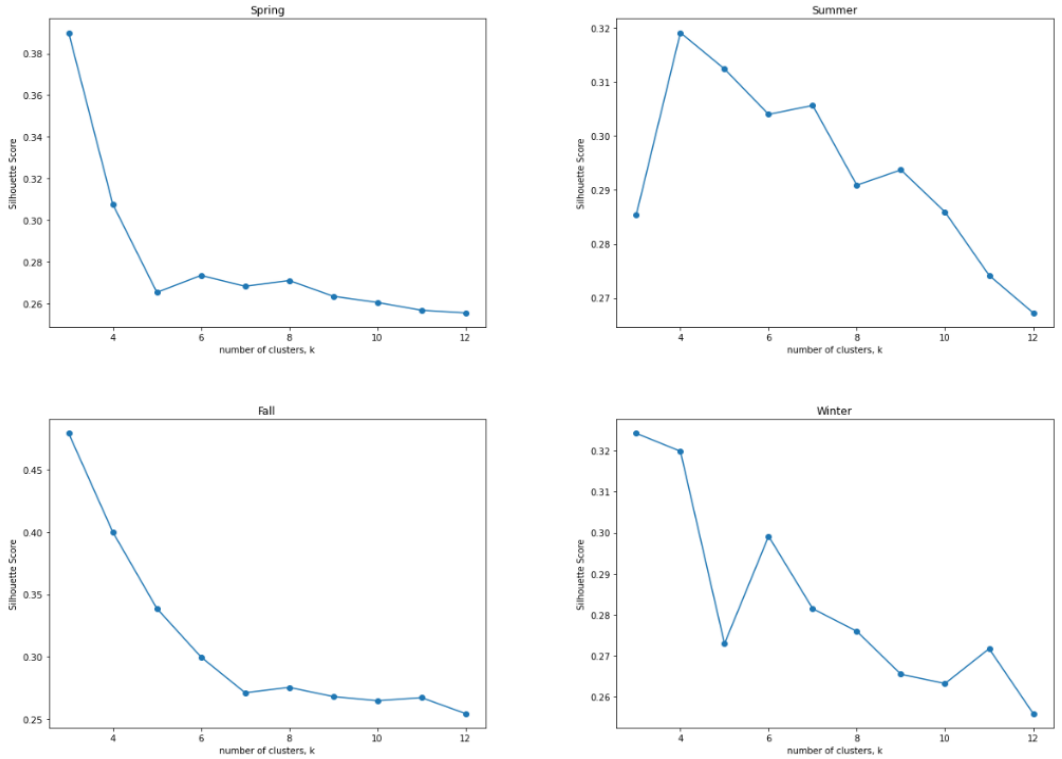

Figure 2: *Silhouette coefficient by season.*

다음과 같이 정의된다.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |Y_t - F_t|}{n},$$
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - F_i)^2}{n}}. \tag{3.1}$$

이 때, n은 예측한 데이터의 개수이며 Y_i는 i 시점에서의 실제 값, F_i는 i 시점에서의 예측값을 의미한다. MAE 와 RMSE 값이 작을수록 모형의 예측 성능이 우수하다는 것을 의미한다 (Park 등, 2021).

3.3. 모형 적합 결과

본 논문에서는 훈련 데이터를 이용해 적합한 모형으로 기준 시점의 3일 이후부터 10일 이후까지의 값들을 예측하여 모형의 성능을 비교하고자 한다. 외생 변수로는 각 지역에서 표출되는 중기예보 기상 요소인 최저기온, 최고기온, 오전 운량, 오후 운량 그리고 각 지역의 위도 경도로 산출할 수 있는 대기권 밖 일사량을 변수로 일치시켜 예측하였다. 실험환경은 Python 3.6 버전의 각 방법론에서 초매개변수(hyperparameter)값을 기본값으로 선정하였고, SVR은 kernel을 'rbf'로, RF 모형은 max_depth를 100, KNN은 계절별 실루엣 계수로 선정한 K값을 조정하였다.

Table 5: Linear regression coefficient value by season

Season	Linear regression coefficient				
	Lowest temperature	Highest temperature	Morning cloud	Afternoon cloud	ei
Spring	−0.840	0.762	−3.045	−1.840	0.593
Summer	−1.211	1.570	−1.603	−2.623	0.414
Autumn	−0.700	0.744	−2.584	−0.812	0.530
Winter	−0.377	0.261	−2.083	−0.911	0.643

Table 6: Performance result by model

Model	MLR		KNN		RF		SVR	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Solar radiation	0.186	0.356	0.182	0.364	0.181	0.358	0.182	0.346

Table 7: Performance result by season

Season	MLR		KNN		RF		SVR	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Spring	0.248	0.448	**0.231**	0.452	0.234	0.451	0.241	0.433
Summer	0.220	0.385	0.219	0.394	**0.218**	0.386	0.221	0.386
Autumn	0.159	0.319	0.161	0.331	**0.156**	0.324	0.152	0.302
Winter	0.113	0.226	0.114	0.237	**0.112**	0.226	0.105	0.213

3.4. 모형 선정 결과

우선, 기후가 유사한 날짜들을 분류하기 위한 K-means 클러스터링을 진행하기 앞서, 실루엣 계수 방법을 이용한 적절한 K 개수 선정 결과는 Figure 2와 같다.

Figure 2에서 확인할 수 있듯이, KNN과 K-means 클러스터링에 적용하기 위한 계절별 K값은 3,4,3,3 로 설정하였다. MLR로 확인한 변수별 회귀 계수값은 Table 5과 같다.

모형 선정을 위해 Table 6을 확인하였을 때, 일별 예측 모형으로는 MAE 기준으로 RF와 클러스터링 및 일별 예측값을 24시간으로 분배하기 위해 K-means 그리고 확률 밀도 함수를 적용한 방법이 가장 예측 성능이 우수하다는 것을 확인할 수 있었다.

또한 계절별 데이터로 모형 적합과 클러스터링이 이루어졌었기 때문에 계절별 예측 성능을 Table 7로 확인하였을 때에도 대체적으로 RF와 K-means 그리고 확률 밀도 함수를 결합한 방법이 가장 성능이 우수하였다.

3.5. 예보 데이터를 이용한 예측 결과

결과적으로, RF와 K-means 그리고 확률 밀도 함수를 결합한 방법을 이용하여 중기예보 데이터를 이용한 일사량 예측결과는 다음과 같다. 일별 일사량 값 예측을 위해 예보 데이터를 RF 모형에 적합하였을 때, 변수별 중요도는 Figure 3과 같으며 일사량 예측에 가장 영향을 크게 미치는 변수는 대기권 밖 일사량(ei), 오전 운량(amcloud), 최소기온(mintemp),최고 기온(maxtemp), 오후 운량(pmcloud) 순인 것을 확인할 수 있다.

일자별 일사량 값을 24시간으로 분배하기 위하여 K-means 클러스터링 모형을 이용, 기후가 유사한 날짜들을 계절별로 분류하였을 때, Figure 4과 같이 클러스터별로 일사량 확률 밀도 함수가 차이나는 것을 확인할 수 있다. 이는 실제로 기후가 비슷한 날들을 분류하였을 때, 일사량 산출 패턴을 비슷하지만 기후에 따라 시

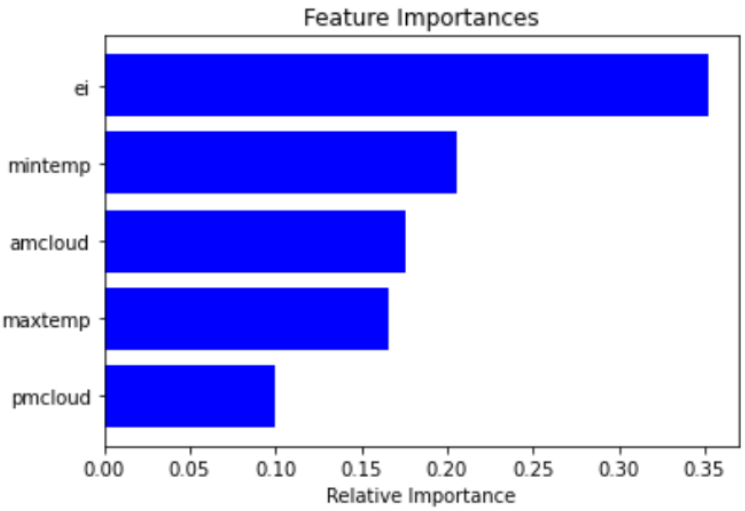

Figure 3: *Feature importances by random forest regression.*

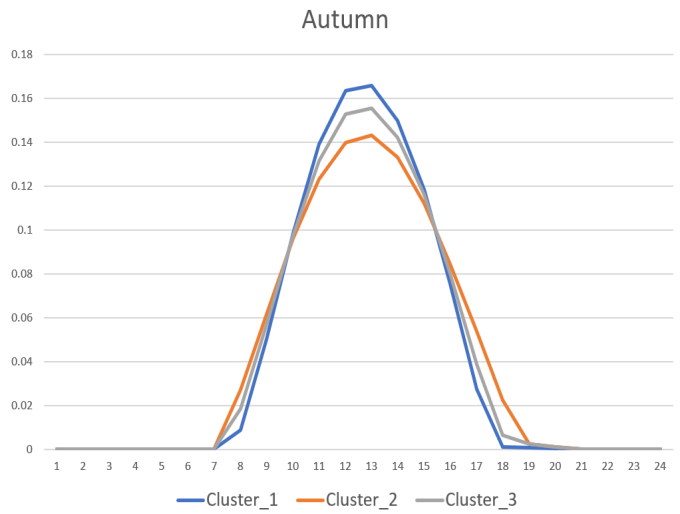

Figure 4: *Probability density function graph by cluster in autumn.*

간별 비율이 다르다는 점을 확인할 수 있다. 확률 밀도 함수는 클러스터별 시간별 일사량 평균값을 하루 총 일사량 평균값으로 나누는 계산법을 사용하였다. 이러한 방법을 통해 각 클러스터별 총합이 1인 시간별 일사량 비율을 나타낼 수 있다. 예를 들면, 가을의 경우 분류되는 Cluster_1의 포함된 일자들의 시간별 일사량 평균값을 계산한 후, 24시간 총 일사량 값에 24시간 각각의 일사량 값에 나누는 방식이다.

지역별 중기예보를 이용한 일사량 예측 결과는 Table 8과 같다. 지역별로 3일자부터 10일자까지 지속적으로 예측 오류가 상승하는 것을 확인할 수 있다. 이러한 패턴은 중기예보 일자별 예측 성능 저하로 인하여 발생할 수밖에 없는 양상으로 보인다.

운량별 중기예보를 이용한 일사량 예측 결과는 Table 9과 같다. 운량 별 예측 결과의 경우, 계절별 일출몰

Table 8: Performance result by region using mid-range weather forecast

Day	+3	+4	+5	+6	+7	+8	+9	+10
Region	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE
Daegu	0.194	0.208	0.210	0.211	0.219	0.226	0.238	0.238
Daejeon	0.197	0.203	0.213	0.211	0.217	0.239	0.236	0.244
Chuncheon	0.195	0.205	0.220	0.218	0.231	0.246	0.252	0.264
Seoul	0.211	0.217	0.226	0.223	0.228	0.244	0.267	0.267
Suwon	0.195	0.198	0.205	0.205	0.211	0.225	0.241	0.246
Incheon	0.222	0.221	0.227	0.233	0.231	0.242	0.268	0.265

Table 9: Performance result by cloud using mid-range weather forecast

Day	+3	+4	+5	+6	+7	+8	+9	+10
Cloud	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE
Sunny	0.418	0.422	0.422	0.402	0.408	0.448	0.461	0.473
Cloudy	0.422	0.429	0.440	0.437	0.427	0.461	0.496	0.490
Overcast	0.427	0.438	0.434	0.422	0.421	0.470	0.503	0.499

Table 10: Performance result using mid-range weather forecast

Day	+3	+4	+5	+6	+7	+8	+9	+10
Solar	MAE	MAE	MAE	MAE	MAE	MAE	MAE	MAE
Radiation	0.203	0.209	0.217	0.217	0.223	0.237	0.251	0.255

시간을 적용하였다. 운량이 증가할수록 예측 성능이 하락하는 것을 확인할 수 있었고, 일자가 지나도 일정한 패턴으로 예측 성능이 하락하기보다는 일자 중간에 예측값들이 좋아지는 상황도 확인 가능하다. 이러한 양상은 예보 기상 데이터의 예측 성능 불확실성으로 인해 발생하는 것으로 보인다. 결과적으로 중기예보를 이용한 일사량 예측 결과는 Table 10과 같다. MAE와 RMSE의 95% 신뢰구간은 (0.254, 0.291), (0.466, 0.504)로 나타내어진다.

4. 결론

본 논문에서는 최근 그린뉴딜과 같은 신재생에너지 발전 정책으로 인해 급증하게 된 태양광 에너지의 수요에 따라 발전량의 정확한 예측에는 일사량이 가장 중요한 역할을 한다고 여겨져 중기예보 데이터를 이용한 일사량의 예측을 하고자 하였다. 특히, 본 연구는 기존의 단기예보 데이터를 이용한 연구가 아닌, 중기예보 데이터를 이용하여 3일부터 10일 이후까지의 일사량을 예측하고자 한 것이 기존 선행 연구들과의 가장 큰 차이점이다. 서울, 인천, 수원, 대구, 대전, 춘천의 중기예보 데이터인 최저기온, 최고기온과 오전, 오후 하늘상태는 운량(1,3,4) 범주로 변경하여 사용하였으며, 각 지역의 위도와 경도를 이용하여 산출 가능한 대기권 밖 일사량(ei)과 일사량 데이터를 이용하여 예측을 실시하였다. 모형적으로는 기본 모형인 다중선형모형회귀, K-NN, RF 그리고 SVR을 이용하였고, 클러스터링 기법으로는 K-means를 이용하였다. 또한 일별 데이터인 중기예보를 이용하여 예측한 일별 일사량 값을 24시간 형태로 나타내기 위하여 시간별 확률 밀도 함수를 계산하는 방법을 제안한다. 관측 데이터를 중기예보 데이터 형식으로 변환한 다음 머신러닝 모형을 선정한 결과, 일별 일사량 예측에 사용된 RF 모형과 유사한 날씨들을 분류한 K-means 그리고 시간별 확률 밀도 함수를 적용한 방법이 MAE 0.181로 가장 예측성능이 우수한 것을 확인할 수 있었다. 또한 이러한 모형으로 중기예보 데이터에 모형을 적합한 후 3일 이후부터 10일 이후까지의 일사량을 예측한 결과, MAE 0.203에서

0.255로 일자가 지날수록 점점 MAE가 증가하는 경향을 파악할 수 있었다. 이는 중기예보의 기상예보 데이터가 일자별로 예측 성능이 하락하기 때문에 발생하는 패턴으로 여겨진다. 본 논문에서는 RF, SVR 그리고 KNN 회귀모형과 K-means 클러스터링과 같은 다양한 분야에서 사용되는 머신러닝 모형을 이용하였고, 확률밀도 함수를 계산하는 방법을 결합하여 일사량 예측을 실시하였다. 추후, 다른 시계열 클러스터링 기법이나 확률 분포 함수 추정 방법 등을 적용하거나 중기예보 데이터에서 활용할 수 있는 또 다른 기상 변수를 활용한 연구가 추가적으로 필요하다고 할 수 있을 것이다.

References

Abuella M and Chowdhury B (2015). Solar power probabilistic forecasting by using multiple linear regression analysis. In *Proceedings of SoutheastCon 2015*, Fort Lauderdale, FL, 1–5.

Alzahrani A, Shamsi P, Dagli C, and Ferdowsi M (2017). Solar irradiance forecasting using deep neural networks, *Procedia Computer Science*, **114**, 304–313.

Al-Hajj R, Assi A, and Fouad MM (2019). Stacking-based ensemble of support vector regressors for one-day ahead solar irradiance prediction. In *Proceedings of 2019 8th IEEE International Conference on Renewable Energy Research and Applications (ICRERA)*, 428–433.

Bamisile O, Oluwasanmi A, Ejiyi C, Yimen N, Obiora S, and Huang Q (2022). Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions, *International Journal of Energy Research*, **46**, 10052–10073.

Benmouiza K and Cheknane A (2013). Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models, *Energy Conversion and Management*, **75**, 561–569.

Breiman L (2001). Random forests, *Machine learning*, **45**, 5–32.

Demir V and Citakoglu H (2022). Forecasting of solar radiation using different machine learning approaches, *Neural Computing and Applications*, **35**, 887–906.

Fan GF, Guo YH, Zheng JM, and Hong WC (2019). Application of the weighted k-nearest neighbor algorithm for short-term load forecasting, *Energies*, **12**, 916.

Fan J, Wang X, Zhang F, Ma X, and Wu L (2020). Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data, *Journal of Cleaner Production*, **248**, 119264.

Gala Y, Fernández Á, Díaz J, and Dorronsoro JR (2013). Support vector forecasting of solar radiation values. In *International Conference on Hybrid Artificial Intelligence Systems : Vol. 8073. LNAI* (pp. 51-60), Springer, Berlin.

Ghofrani M, Ghayekhloo M, and Azimi R (2016). A novel soft computing framework for solar radiation forecasting, *Applied Soft Computing*, **48**, 207–216.

Hong T (2010). *Short Term Electric Load Forecasting*, North Carolina State University, Raleigh.

Jawaid F and NazirJunejo K (2016). Predicting daily mean solar power using machine learning regression techniques. In *Proceedings of 2016 6th IEEE International Conference on Innovative Computing Technology (INTECH)*, 355–360.

Jiménez-Pérez PF and Mora-López L (2016). Modeling and forecasting hourly global solar radiation using clustering and classification techniques, *Solar Energy*, **135**, 682–691.

Kim H, Park S, and Kim S (2022). Solar radiation forecasting using boosting decision tree and recurrent neural networks, *Communications for Statistical Applications and Methods*, **29**, 709–719.

Karasu S and Altan A (2019). Recognition model for solar radiation time series based on random forest with feature selection approach. In *Proceedings of 2019 11th IEEE International Conference on Electrical and Electronics Engineering (ELECO)*, 8–11.

Khosravi A, Koury RNN, Machado L, and Pabon JJG (2018). Prediction of hourly solar radiation in Abu Musa Island using machine learning algorithms, *Journal of Cleaner Production*, **176**, 63–75.

Ozturk I (2013). Energy dependency and energy security: The role of energy efficiency and renewable energy sources, *The Pakistan Development Review*, **52**, 309–330.

Park S, Lee JY, and Kim S (2021). Wind power forecasting based on time series and machine learning models, *The Korean Journal of Applied Statistics*, **34**, 723–734.

Rousseeuw PJ (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Srivastava R, Tiwari AN, and Giri VK (2019). Solar radiation forecasting using MARS, CART, M5, and random forest model: A case study for India, *Heliyon*, **5**, e02692.

Tambunan HB, Barus DH, Hartono J, Alam AS, Nugraha DA, and Usman HHH (2020). Electrical peak load clustering analysis using K-means algorithm and silhouette coefficient. In *Proceedings of 2020 International Conference on Technology and Policy in Energy and Electric Power (ICT-PEP), IEEE*, Bandung, 258–262.

Tranmer M and Elliot M (2008). Multiple linear regression, *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, **5**, 1–5.

Vapnik V (1999). *The Nature of Statistical Learning Theory*, Springer science & business media, NewYork.

Yadav AP and Behera L (2014). Solar Radiation forecasting using neural networks and Wavelet Transform, *IFAC Proceedings Volumes*, **47**, 890–896.

Zhang Z, Wang C, Peng X, Qin H, Lv H, Fu J, and Wang H (2021). Solar radiation intensity probabilistic forecasting based on K-means time series clustering and Gaussian process regression, *IEEE Access*, **9**, 89079–89092.

Received October 13, 2022; Revised November 9, 2022; Accepted November 9, 2022

중기예보를 이용한 태양광 일사량 예측 연구

박수진[a], 김효정[a], 김삼용[1,a]

[a]중앙대학교 응용통계학과

요 약

급속적으로 비중이 증가하고 있는 태양광 에너지는 지속적인 개발 및 투자가 이루어지고 있다. 신재생 에너지 정책인 그린뉴딜과 가정용 태양광 패널의 설치가 증가함에 따라 국내 태양광 에너지 보급이 점차 확대되어 그에 맞추어 발전량의 정확한 수요 예측 연구가 활발하게 진행되고 있는 시점이다. 또한, 일사량 예측이 발전량 수요 예측에 가장 영향을 미치는 요소로 작용하고 있다는 점에서 일사량 예측의 중요성을 파악하였다. 덧붙여, 본 연구는 선행 연구들에서 사용되지 않은 중기예보 기상 데이터를 활용하여 일사량 예측을 하고자 하였다는 점에서 가장 큰 차이점을 확인할 수 있다. 본 논문에서는 서울, 인천, 수원, 춘천, 대구, 대전의 총 여섯 지역의 태양광 일사량 예측을 위하여 다중선형회귀모형, KNN, Random Forest 그리고 SVR 모형과 클러스터링 기법인 K-means 기법을 결합한 후, 클러스터별 확률밀도함수를 계산하여 시간별 일사량 예측을 진행하고자 하였다. 중기예보 데이터를 사용하기 전, 모형 예측 결과를 비교하기 위한 지표로서 MAE (mean absolute error)와 RMSE (root mean squared error)를 사용하였다. 데이터는 2017년 3월 1일부터 2022년 2월 28일까지의 시간별 원 관측 데이터를 중기예보 데이터 양식에 맞추어 일별 데이터로 변환하였다. 모형의 예측 성능 비교 결과, Random Forest로 일별 일사량을 예측한 후, K-means 클러스터링으로 기후요인이 유사한 날짜들을 분류한 뒤 클러스터별 일사량의 확률밀도함수를 계산하여 시간별 일사량 예측값을 나타낸 방법이 가장 우수한 성능을 보였다. 또한 이 방법론을 이용하여 중기예보 데이터에 모형 적합 후, 예측 결과를 확인하였을 때, 일자별로 예측 오류가 상승하는 것을 확인할 수 있었다. 이는 중기예보 기상데이터의 예측 오류로 인한 것으로 보인다. 향후 연구에서는 중기예보 데이터에서 활용할 수 있는 기상요인 중, 강수 여부와 같은 외생 변수를 추가하거나 시계열 클러스터링 기법을 적용한 연구가 이루어져야할 것으로 보인다.

주요용어: 시계열, 중기예보, 일사량 예측, Random Forest, K-means clustering, 확률 밀도 함수

이 논문은 2022년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.

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