Journal of Environmental Science International
26(4); 521~527; April 2017

pISSN: 1225-4517 eISSN: 2287-3503
https://doi.org/10.5322/JESI.2017.26.4.521

ORIGINAL ARTICLE

제주 실시간 일사량의 기계학습 예측 기법 연구

이영미* · **배주현** · **박정근**

㈜에코브레인

A Study on Prediction Techniques through Machine Learning of Real-time Solar Radiation in Jeju

Young-Mi Lee*, Joo-Hyun Bae, Jeong-keun Park

Eco Brain Co., Ltd., Jeju 63309, Korea

Abstract

Solar radiation forecasts are important for predicting the amount of ice on road and the potential solar energy. In an attempt to improve solar radiation predictability in Jeju, we conducted machine learning with various data mining techniques such as tree models, conditional inference tree, random forest, support vector machines and logistic regression. To validate machine learning models, the results from the simulation was compared with the solar radiation data observed over Jeju observation site. According to the model assesment, it can be seen that the solar radiation prediction using random forest is the most effective method. The error rate proposed by random forest data mining is 17%.

Key words : Solar radiation prediction, Machine learning, Data mining, Tree models, Conditional inference tree, Random forest, Support vector machine, Logistic regression

1. 서 론

일사량은 도로에서 발생하는 위험 요소인 결빙에 의한 피해를 방지하는 환경 요인으로 안개, 운량 등 기상요소들과의 상관성이 높아서 그 중요성이 크다. 이에 예측되어지는 기상요소들을 기반으로 한 정확한 일사량 예측은 도로 위험 예보시스템 구축에 있어서 필수요인이라 볼 수 있다. 기존 일사량 예측 연구는 태양에너지 자원량의 정확한 예측을 위해 필수적이기 때문에 다양한 방향으로 진행되어왔다(Lee et al., 2011; Jee et al., 2012; Kim and Kim, 2016).

일반적으로 일사량을 예측하는 통계모형의 경우는 시간별 일사량과 운량, 기온, 상대 습도, 기압 등의 기상 매개 변수 또는 과거의 일사량 관측치 간의 상관관계를 기반으로 개발된다. 일 수평면 일사량 예측에 가장 많이 사용되는 모델은 인공 신경망(Artificial Neural Networks, ANN)이며, 국지적 기상 관측값들과 통계 매개 변수들을 사용해서 가장 단순한 알고리즘에 의해 예보될 수 있다(Martin et al., 2010; Mellit and Massi, 2010; Voyant et al., 2013). 하지만 신경망의

Received 16 March, 2017; **Revised** 4 April, 2017;
Accepted 12 April, 2017
***Corresponding author**: Young-Mi Lee, Eco Brain Co., Ltd., Jeju 63309, Korea
Phone : +82-70-7018-3082
E-mail : leeym@ecobrain.net

경우는 그 해석이 복잡해서 일사량 예측에 있어서 중요한 변수를 찾아내기 힘든 단점이 있다. 그 외 Z&H 모델식을 활용하여 국내에 적합한 수평면전일사량 예측식을 개발하는 연구가 많이 이루어지고 있다(Kim and Kim, 2016).

본 연구에서는 기상청에서 예측되고 있는 기상요소들과 일사량 자료들을 이용하여 더 개선된 방식으로 일사량을 예측하고자 다양한 기계학습 기법들을 이용하여 우리나라 제주 지점의 일사량을 실시간으로 예측하고자 한다. 동시에 통계 모형들의 입력변수 중요도를 함께 분석함으로 인해 일사량 예측에 있어서의 효과적인 방안을 모색하고자 하였으며, 이처럼 정확한 일사량의 실시간 예측을 통해 도로 위에서의 위험 요소인 결빙량을 실시간으로 예측 가능할 수 있음으로 인해 안전한 도로 교통 환경을 조성할 수 있을 것이다.

2. 재료 및 방법

2.1. 기계학습 이론

기계학습(machine learning)이란 컴퓨터에게 배울 수 있는 능력, 즉 코드로 정의하지 않은 동작을 실행하는 능력에 대한 연구분야이며(Arthur, 1959), 인공 지능의 한 분야로써 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다.

일사량 예측 모형을 개발하기 위해 사용한 기계학습 기법들로는 의사결정나무(tree models)와 조건부 추론 나무(conditional inference tree), 앙상블 학습기법인 랜덤포레스트(random forest), 서포트 벡터 머신(Support Vector Machine, SVM), 로지스틱 회귀 분석(logistic regression) 이다.

본 연구에서 사용한 기계학습 도구 R은 기계 학습을 포함하여 통계, 금융, 생물정보학, 그래픽스에 이르는 다양한 통계 패키지를 갖추고 있으며, 본 연구에서 사용되는 패키지로는 rpart, party, randomForest, kernlab, nnet이다(Table 1).

의사결정나무는 의사결정규칙을 나무 구조로 도표화하여 분류와 예측을 수행하는 분석 방법이며, 조건부 추론 나무는 변수와 반응값 사이의 연관관계를 측정하여 노드 분할에 사용할 변수를 선택하는 다중 가설 검정을 고려한 방법이다. 랜덤포레스트는 입력변수로부터 여러 개의 모델을 학습한 다음, 예측 시 여러 모델의 예측 결과들을 종합 사용하여 정확도를 높이는 기법이고 로지스틱 회귀분석은 단순회귀분석과 다중회귀분석이 선형으로 가정하는데 비해 S자형으로 가정하며, 서포트 벡터 머신은 입력되는 데이터를 두 집단으로 분리하고 분석하는 학습 알고리즘이다(Lee et al., 2016).

Table 1. Statistical packages used in this study

Model	R Package
Tree models	rpart
Conditional inference tree	party
Random forest	random forest
Support vector machine	kernlab
Logistic regression	nnet

2.2. 분석 방법

일사량 예측모형 개발을 위해 제주 지역의 ASOS (Automatic Synoptic Observation System) 일사량 관측자료와 기상청의 통합모델(Unified Model, UM) LDAPS (Local Data Assimilation and Prediction System) 기상정보를 활용하여 통계모형을 개발하였다.

학습기간은 2015.1.1. 01시 ~ 2015.12.31. 23시이고 검증기간은 2016.1.1. 00시 ~ 2016.12.31. 23시이며, 학습과 검증을 위해 사용되는 제주 ASOS 일사량 관측자료와 예측된 LDAPS 기상자료들은 제주 ASOS 관측 지점인 위도 35.5141°, 경도 128.5297° 위치 자료이다.

입력변수는 일사량 LDAPS 예측값, 기온 LDAPS 예측값, 습도 LDAPS 예측값, 풍속 LDAPS 예측값, 운량 LDAPS 예측값, 입력 시각과 지점의 고도각과 방위각, 1시간 전 일사량 관측값, 2시간 전 일사량 관측값, 3시간 전 일사량 관측값, 4시간 전 일사량 관측값, 5시간전 일사량 관측값이며, 출력변수는 예측 일사량으로 통계모형 학습시에는 관측 일사량을 출력값으로 지정하여 훈련하였다.

모형 입·출력변수는 Table 2에 나타내었으며, 입력변수 중 1~5시간 전 일사량은 ASOS 관측값이고 고도각과 방위각을 제외한 나머지 입력변수는 LDAPS

Table 2. Information of input and output variables in machine learning model

Input data								Output data
T	RH	WS	C	A	Z	SR	b1∼b5	S
Air temperatures	Relative humidity	Wind speed	Cloud cover	Azimuth	Zenith angle	Solar radiation	1∼5 hours before solar radiation	Solar radiation
LDAPS predicted weather elements							ASOS observation values	

기상 예측값으로 학습시 출력변수는 관측 일사량 자료로 설정하였다.

검증방법은 2016년 LDAPS 자료를 입력 후, 일사량 예측 결과와 실제 관측값과의 검증이 이루어진다. 본 모형 검증들을 위해서는 아래와 같은 검증방식들을 적용하였다.

먼저, 일사량값과 계산값의 상관계수 R(correlation coefficient)은 상관관계의 크기를 나타내며, R값은 두 변량이 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 -1을 가진다. 오차들의 절대값의 평균을 구하는 방법으로 MAE (Mean Absolute Error)와 시스템의 기상 예측 성능 평가를 표현할 때 사용되는 오차율(%, NMAE: Normalized Mean Absolute Error)을 사용한다. 이는 실제와 예측의 차에 전체기간의 실제값 평균으로 나눈 그 값의 평균을 구하여 100%을 곱한 값이다. 검증식들은 식(1)∼(3)과 같이 표현할 수 있으며, M은 모델값, O는 관측값이며, mean은 평균값을 뜻한다.

$$R=\frac{\sum(O_i-O_{mean})(M_i-M_{mean})}{\sqrt{\sum(O_i-O_{mean})^2\times\sum(M_i-M_{mean})^2}} \quad (1)$$

$$MAE=\frac{1}{N}\sum_{i=1}^{n}|M_i-O_i| \quad (2)$$

$$NMAE(\%)=\frac{1}{N}\sum_{i=1}^{n}\frac{|M_i-O_i|}{O_{mean}}\times 100\% \quad (3)$$

3. 결과 및 고찰

3.1. 분석 결과

R을 통한 5가지 통계 패키지를 활용하여 2015년 제주 지점의 일사량과 LDAPS 기상요소들, 고도각과 방위각 등으로 학습을 해서 모형을 개발하고 이 모형들을 이용하여 2016년 일사량을 예측하고 검증하였다.

의사결정나무를 통해 제주지역에서의 변수별 일사량 예측 중요도를 확인하였다. 의사결정나무의 경우 조건부 추론 나무와 마찬가지로 입력변수의 분류를

Table 3. Verification results in each model

Model	R	MAE(W/m^2)	NMAE(%)
Tree models	0.939	47.891	32.237
Tree models (correction)	0.939	43.769	29.463
Conditional inference tree	0.967	28.748	19.352
Random forest	0.973	25.197	16.961
Random forest (correction)	0.973	25.251	16.998
Support vector machine	0.973	29.744	20.014
Support vector machine (correction)	0.974	25.761	17.334
Logistic regression	0.968	35.698	24.021
Logistic regression (correction)	0.969	29.103	19.583
LDAPS	0.930	49.307	33.190

통해 예측을 시도했으며, 본 연구에서 진행한 의사결정나무의 분류도를 Fig. 1에 나타내었다. 13개의 입력변수 중 SR 즉 LDAPS에 의해 예측된 일사량이 가장 큰 분류 기준이 되며, 두 번째로 1시간 전 관측 일사량의 영향이 큰 것을 살펴볼 수 있었다.

앙상블 학습 모형인 랜덤포레스트 모형에서의 변수 중요성을 분석하였으며 Fig. 2에서 살펴볼 수 있다. 2가지 방식으로 변수들을 테스트하며, 왼쪽 그래프에서는 각 변수 취환 후 테스트 과정을 통해 평균제곱오류(Mean Squared Error, MSE)를 기록한 그림으로 차이가 클 수로 변수가 더 중요한 것을 나타낸다. 오른쪽 그래프는 각 변수들의 노드 순도를 나타내며 그래프들을 종합해 보건데, 변수 정확도와 정확도 개선에 중요한 변수로는 LDAPS 일사량과 운량, 1시간 전 관측 일사량임을 알 수 있었다.

일사량 예측모형 개발을 위한 중요 입력변수로는 무엇보다도 기상수치모델을 통해 예측된 일사량, 즉 본 연구에서는 LDAPS 일사량과 예측 이전 시간대의 관측 일사량, 그리고 운량, 고도각과 방위각이므로, 일사량 예측 모형 구축에 있어서는 무엇보다도 본 변수들을 중심으로 한 학습 데이터셋을 구성할 필요가 있을 것으로 판단된다.

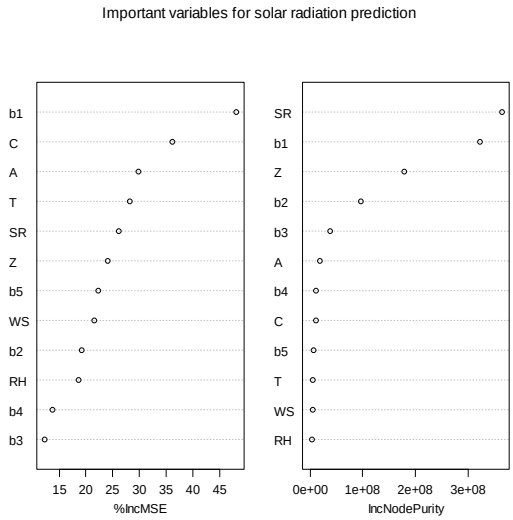


Fig. 2. Importance of solar radiation predicting variables in random forest model.

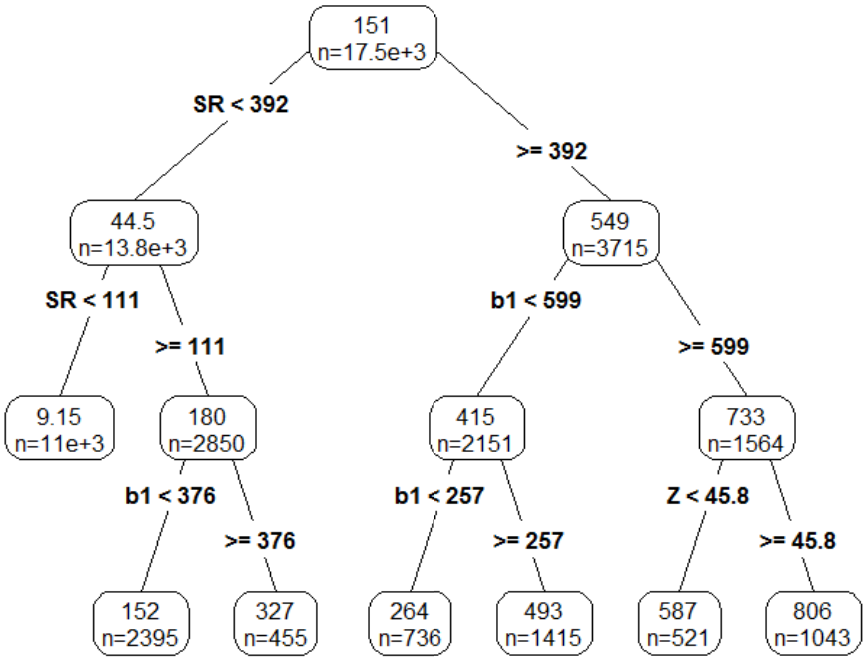
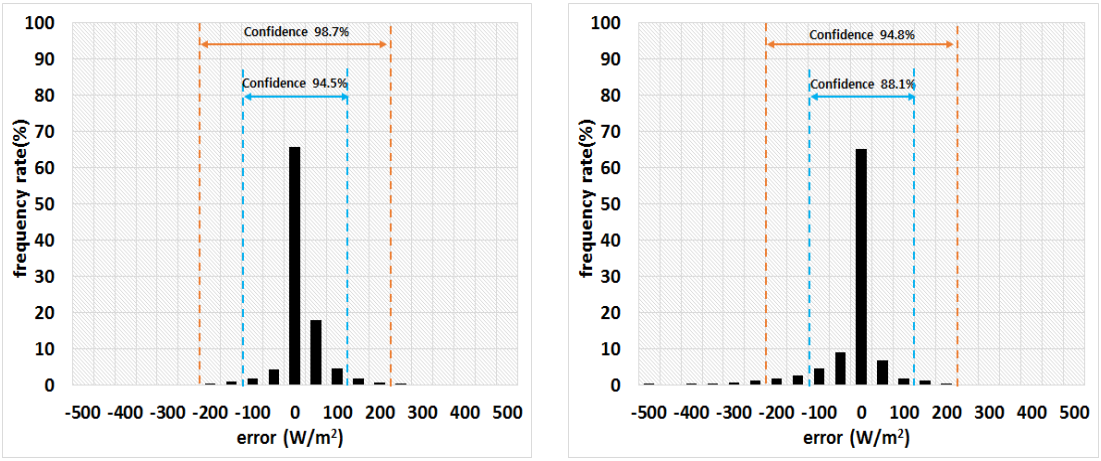


Fig. 1. Classification map of tree models at Jeju (184).



(a) (b)

Fig. 3. The frequency rate of measured and predicted solar radiation differences by (a) random forest (b) LDAPS.

3.2. 검증 결과

일사량 예측 모형들의 성능 비교를 위해 R, MAE, NMAE(%) 검증 결과를 분석해 보았다. Table 3을 살펴보면, 각 기계학습 결과들 중 밤 동안에 일사량이 0이 아닌 경우는 전부 0의 수치로 보정하는 단계를 거쳤고 함께 제시하였다. 조건부추론나무의 경우는 밤 동안에 일사량이 없는 것을 제대로 학습하여 예측 결과에서도 정확히 0의 값을 출력하였기에 보정 절차를 거치지 않아도 되는 편리성을 가진다.

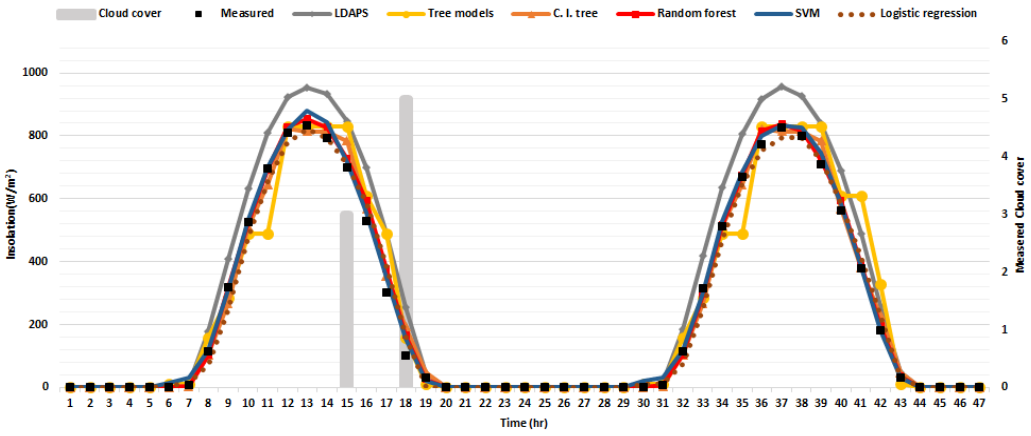
검증 결과를 살펴보면, 랜덤포레스트 보정 결과가 상관계수 R 0.97, MAE 25.3 W/m^2, NMAE 오차율이 17%를 나타내며 가장 좋은 결과를 보여주었다. 두 번째로는 SVM 보정 결과가 NMAE 오차율 17.3%, 다음이 조건부추론나무 모형이 19.4%의 오차율을 나타내었다. 33.2%의 NMAE 오차율을 나타내는 LDAPS에 비해 기계학습으로 예측된 모형 결과들이 상당히 일사량의 예측률을 증가시켰음을 알 수 있었다.

대표적으로 Fig. 3의 랜덤포레스트 보정 결과와 LDAPS 예측 일사량의 관측값과의 차이에 따른 오차 빈도율에서 보듯이 200 W/m^{2}의 절대 오차 범위 안에 랜덤포레스트 결과는 98.7%, LDAPS는 94.8%의 신뢰도를 나타내는 것을 볼 수 있으며, 랜덤포레스트가 LDAPS 보다 좋은 결과를 나타냄을 알 수 있었다.

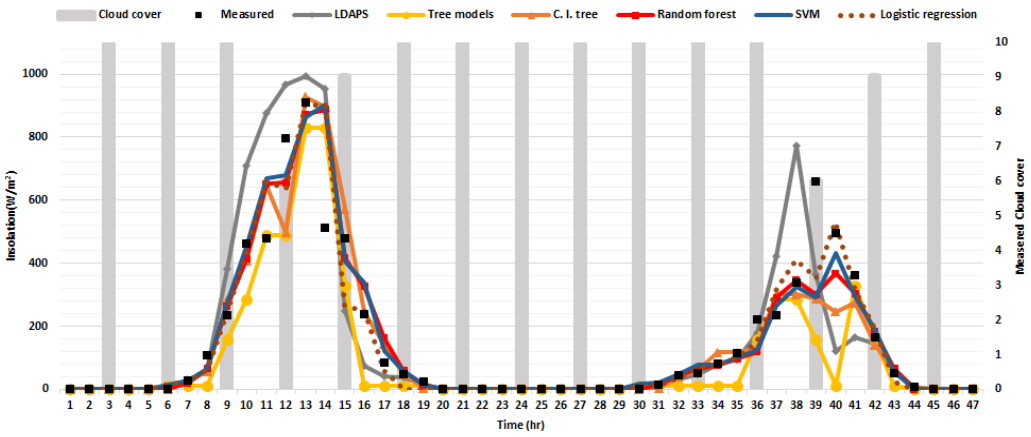
맑은 날과 흐린 날의 사례에 대한 시계열 변화를 살펴보기 위해, 관측운량이 이틀간 대부분 0~1을 나타내는 2016년 3월 27일~28일과 관측 운량이 9~10을 나타내는 2016년 5월 27일~28일의 관측일사량과 LDAPS, 그리고 각 기계학습 모형 결과에 따른 시계열 변화를 함께 살펴보았다.

ASOS 지점에서의 운량은 3시간 간격으로 관측이 이루어지고 있으며 먼저 27일 15시에 운량 3, 18시에 운량 5의 값을 나타내며 나머지 시간대에는 모두 0의 운량을 보였던 맑은 날의 경우(Fig. 4(a)), LDAPS를 제외한 5개 모형 결과들은 대체로 관측값을 잘 모사하고 있음을 알 수 있으며, 특히 랜덤포레스트 결과가 상당히 관측값과 일치하는 패턴을 보여주고 있다. Fig. 4(b)에서는 운량이 대체로 9~10이였던 흐린 날의 일사량 시계열 변화이며 맑은 날 사례에 비해 예측값들이 관측값과 정확히 일치하는 패턴을 나타내지 않고 있으며, LDAPS는 과대평가를, 나머지 모형들은 과소평과 하는 추세이지만 역시 랜덤포레스트와 SVM 결과가 우수한 것을 알 수 있다. 특이한 점은 27일 11시에 약 480 W/m^{2}의 관측 일사량을 의사결정나무 결과가 490 W/m^2 정도의 가장 가까운 값을 나타내면서 일치성이 큰 것을 살펴볼 수 있었다.

특정 지점과 사례별로 학습모형의 결과가 조금씩 다르기 때문에 한 가지 기계학습 기법만이 우수하다고는 볼 수 없지만 제주 지점의 일사량 예측에 있어서는



(a)



(b)

Fig. 4. The time series variation of measured and predicted solar radiations at (a) clear (2016.3.27.~28) and (b) cloudy (2016.5.27.~28) day.

랜덤포레스트 기법을 활용하는 방안이 가장 좋을 것으로 밝혀졌다.

4. 결 론

일사량 예측의 정확도 개선을 위해 다양한 데이터 마이닝 기법을 이용하여 기계학습을 실시하였다. 본 연구에서 시행한 학습으로는 의사결정나무, 조건부 추론 나무, 랜덤포레스트, 서포트 벡터 머신, 로지스틱 회귀 분석의 5가지 기법들이다.

기계학습 모형들의 예측 결과가 LDAPS 일사량 예측값을 많이 보정하는 역할을 수행하여 관측값과의 오차를 감소시키고 예측률을 향상시킬 수 있었다. LDAPS의 33.2%의 오차율을 랜덤포레스트 보정 결과를 통해 17%로 감소시킬 수 있었으며, 이에 가장 중요한 입력변수로는 LDAPS 예측 일사량과 예측 운량, 그리고 실시간 제공되는 1시간 전 관측 일사량임을 알 수 있었다.

본 연구를 통해 겨울 동안의 결빙으로 인해 발생하는

도로 교통사고를 방지하는 차원에서 정확한 일사량 예측에 따른 노면 온도를 추정함으로 인해 도로 결빙 여부를 사전에 방지할 수 있기에 제주 지역에서의 도로 기반 안전 정보 제공 시스템 구축에 큰 도움이 될 수 있으리라 기대된다.

감사의 글

이 논문은 2015년도 지역주력산업육성(R&D) 기술개발 사업인「운전자 환경 반응형 CEV (Connected Electricity Vehicle) 서비스 개발」사업(R0003890)으로 산업통상자원부의 지원을 받아 연구한 논문임.

REFERENCES

Arthur, S., 1959, Some studies in machine learning using the game of checkers, IBM Journal, 3(3), 210-229.

Jee, J. B., Lee, S. W., Choi, Y. J., Lee, K. T., 2012, The generation of typical meteorological year for research of the solar energy on the Korean Peninsula, New & Renewable Energy, 8(2), 12-23.

Kim, H. Y., Kim, J., 2016, Prediction correlation of solar insolation using relationships between meteorological data and solar insolation in 2012(I), Journal of KSES, 36(1), 1-9.

Lee, K. T., Zo, I. S., Jee, J. B., Choi, Y. J., 2011, Temporal and spatial distributions of the surface solar radiation by spatial resolutions on Korea Peninsula, New & Renewable Energy, 7(1), 22-28.

Lee, Y. M., Bae, J. H., Park, D. B., 2016, A Study on fog forecasting method through data mining techniques in Jeju, Journal of Environmental Science International, 25(3), 417-424.

Martin, L., Zarzalejo, L. F., Polo, J., Navarro, A., Marchante, R., Cony, M., 2010, Benchmarking of different approaches to forecast solar irradiance, Solar Energy, 84(10), 1772-1781.

Mellit, A., Massi, P. A., 2010, A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy, Solar Energy, 84(5), 807-821.

Voyant, C., Randimivololona, P., Nivet, M. L., Poli, C., Muselli, M., 2013, Twenty four hours ahead global irradiation forecasting using multilayer perceptron, Meteorological Applications, 1387.