Comparison between the empirical and machine techniques to predict global solar irradiance for Mutale area in Limpopo Province, South Africa

T.W Murida[1], T.S Mulaudzi[1], N.E Maluta[1,2], N Mphephu[3]

[1]Department of physics, University of Venda, Thohoyandou, South Africa
[2]National Institute for Theoretical and Computational Sciences (NITheCS), Gauteng, South Africa
[3]Standard Bank South Africa

E-mail: thaluwhitney@gmail.com

Abstract. The prediction of solar irradiance is of utmost importance in guiding solar power conversion systems with a specific focus on design, modelling, and operation. Availability of solar irradiance data plays a signicant role for decision-makers responsible for future investment policies about green energy. The lack of weather stations and measured solar parameters in most areas in the developing countries have contributed to the development of solar prediction models. However, reliable prediction of solar irradiance is dependent on the availability of quality data and the prediction methods. Empirical models have been developed and used in the past; however, in recent times intelligent algorithms have proved to have more predictive power due to the availability of high-frequency data. The study use two empirical models namely: the Clemence model and Hargreaves and Samani model to predict the global solar irradiance at Mutale station in the Limpopo province in South Africa. Furthermore, machine learning and deep learning techniques namely: support vector machines, random forest and artificial neural network were also used to predict global solar irradiance in the same area. To assess the predictive power of these empirical and machine models, the estimated values for the global solar radiation was compared against the recorded data from the Mutale weather station. Based on the results that were found on this study, machine learning techniques tend to give better or good result compared with empirical models.

1. Introduction

There has been a consensus that several countries across the globe are confronted with significant energy crises. As energy consumption rises, the world will encounter a substantial shortage of fossil fuels in the future decades because such powers provide most of the world's energy [1]. The existing consumption energy across the globe is based almost completely on non-renewable resources like oil, gas, and coal. however in recent times, there has been a call to reduce carbon emmissions since these are not environmental free. The fundamental issue is the energy crises, particularly acute in emerging countries where there is a need to power families and industrial sectors [2]. It is estimated that in the next 30 years, the world will be overpopulated and the energy demand will also increase [12].

The economy of developing countries that are highly industrialized, for example, South Africa, has a high energy demand. As the primary power provider in South Africa, Eskom relies on fossil fuels as an energy source, which is harmful to the environment. Using fossil fuels increases greenhouse gas emissions, such as carbon dioxide, causing global warming and hurting the ecosystem and biodiversity. The best solution to this challenge is for the world to move away from non-renewable resources and use renewable energy resources. In any given location, the demand for renewable and sustainable energy has increased.
Solar radiation from the sun is quickly becoming a viable alternative source to other traditional energy sources. Solar energy looks to be the most popular alternative among the various forms of clean energy sources because of its endless and non-polluting nature. Precise solar radiation estimation tools are critical in the design of the solar system. However, solar irradiance forecast depends on the available data and the forecasting methodologies used. Empirical models have been developed and used in the past, but due to the availability of high-frequency data, intelligent algorithms have recently proven to have more prediction potential [5].

2. Study area

The study area of this project has been selected to be at Mutale area, Limpopo province South Africa. The province experiences high temperatures averaged to 25.2° in january while the coldest month is june at an average of 12.5°. The geographic coordinates of the selected station are tabulated in table 1.

Table 1. Geographical coordinates Mutale area.

Station	Latitude	Longitude	Altitude
Mutale	-22.73461	30.52188	550

3. Methodology

The study employed the historical data that was recorded by Agricultural Research Council (ARC). Two empirical models namely; Clemence and Hargreaves-Samani was employed to evaluate and estimate the global solar radiation for each location. Again, machine learning algorithms will also be utilized for further predictions. The empirical and machine learning models was compared against each other and the performances of empirical models and ML algorithms for estimating daily solar radiation will be further evaluated in different areas using statistical equations.

3.1. ***Hargreaves-Samani***

Hargreaves-Samani developed a temperature based model to predict global solar radiation [8]. Based on this principle, he recomended a simple equation defined as

$$H = H_o * (K_r \sqrt{\Delta T}), \tag{1}$$

where, H represents daily mean value of global solar radiation ($MJm^{-2}day^{-1}$), Ho is daily mean value of extraterrestrial radiation ($MJm^{-2}day^{-1}$), $\sqrt{\Delta T}$ is the difference between the maximum temperature (T_{max}) and the minimum temperature (T_{min}) in (^oC),K_r is the empirical coefficient, respectively and ΔT is the difference between the maximum temperature and the minimum temperature.

3.2. ***Clemence Model***

Clemence [9] has developed the temperature based equation for estimating global solar radiation given by

$$H = (1.233 * H_0 * \Delta T + 10.593 * T_{max} - 0.713 * T_{max} * \Delta T + 16.5480) * (0.04184), \quad (2)$$

where ΔT is the difference between the maximum temperature (T_{max}) and the minimum temperature (T_{min}).

3.3. ***Random Forest***

Random forest is a supervised ensemble machine learning technique that is utilized to solve classification and regression problems [13]. The random forests are a combination of several decision trees that were built using the bootstrapping technique, which involved selecting randomly at each node from samples in the learning dataset for the predictors. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers.

3.4. ***Support Vector Machine***

A support vector machine(SVM) learning is a supervised machine learning algorithm used for both classification and regression problems [10]. SVM models include a variety of fundamental kernel functions, including polynomial (Poly), Gaussian, exponential radial basis function (ERBF), radial basis function (RBF), sigmoid, and linear kernels. The SVR works by mapping the input space into a high-dimensional feature space and constructs the linear regression in it which can be expressed as

$$f(x) = w\phi(x) + b \quad (3)$$

where ω is the weight vector, $\phi(x)$ maps inputs x into a high dimensional feature space that is nonlinearly mapped from the input space x and b is the bias term. The main aim of using SVM is to minimize the weight. Comparing to the SV formulation for soft margin linear cassifiers, the linear regression fomulation is given by:

$$minimize \qquad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}(\zeta_i + \zeta_i^*) \quad (4)$$

$$Subject\ \ to \qquad y^i - \mathbf{w}^T\mathbf{x}^i - w_0 \leq \varepsilon + \zeta_i; \qquad i = 1, ..., N \quad (5)$$

$$\mathbf{w}^T\mathbf{x}^i - w_0 - y^i \leq \varepsilon + \zeta_i^*; \qquad i = 1, ..., N \quad (6)$$

$$\zeta_i, \zeta_i^* \geq 0 i = 1, ..., N \quad (7)$$

where C is the regularization, $||\mathbf{w}||$ is the weight, ζ^i and ζ^* is the errors.

3.5. ***Artificial Neural Network***

Artificial neural networks are the neural networks that is based on the design of a human neuron. In the domains of artificial intelligence, machine learning, and deep learning, neural networks enable computer programs to identify patterns and resolve common issues by recognising the behaviour of the human brain [11]. ANNs are comprised of a node layers, containing an input layer, one or more hidden layer and an output layer. Each node or artificial neuron is connected to others and has a weight and threshold that go along with it.For the general neural network model, we let N_l to denote the number of neurons in the l-th layerfor $l = 1, 2, ..., L$.

$$y_j^{(l)} = \sum_{j=0}^{N_{l-1}} \omega_{ij}^{(l)} a_j^{l-1}, \tag{8}$$

where $y_j^{(l)}$ denote activation, $\omega_{ij}^{(l)}$ is the weight and a_j^{l-1} represent the bias neuron. g is the activation function, in this study relu, sigmoid, softplus, softsign, tanh, selu, elu, exponential, LeakyReLU and relu were used and through crossvalidation, the best parameter was relu.

$$a_j^{(i)} = g\left(\sum_{j=0}^{N_{l-1}} \omega_{ij}^{(l)} a_j^{l-1}\right), i = 1, 2, ..., N_l, l = 2, 3, ..., L \tag{9}$$

The predicted output layer will be given by

$$\hat{y}(\mathbf{z},\mathbf{w}) = \sum_{j=0}^{N_{L-1}} \omega_{N_{Lj}}^{(L)} a_j^{L-1} \tag{10}$$

4. Model evaluation metrics

Statistical data analysis was used to test the accuracy and performance of the models. The following statistical equations namely: mean absolute error(MAE), coefficient of determination(R^2), root mean square error(RMSE) and mean square error(MSE) were employed to evaluate the results.

5. Results and discussion

This section presents the findings from the empirical and machine learning methods used to estimate the daily global solar radiation for the Mutale area. Python was used for all computations for this work, while Matlab was used to calibrate the empirical models when determining the model coefficients. The subsections below describe and display these results. Results for all methods that are used under this study and the comparison between the observed and estimated are shown and discussed under these subdections. Performance metrics results are also displayed under this section.

The figures below illustrate the observation and the estimated daily global solar radiation for the Clemence, Hargreaves-Samani, random forest, support vector machines and artificial neural network models. It can be observed that solar radiation reaches its peak and lowest points throughout the summer and winter seasons, respectively. Summer is the season with the largest levels of global solar radiation, which coincides with the highest temperatures. Figure 1 and 2 represent the global solar radiation of the observed and estimated data for the Clemence model and Hargreaves-Samani model. Figure 3, 2 and 5 represent global solar radiation for random forest and support vector machine and artificial neural network models,

The table below shows the statistical analysis for global solar radiation comparing observed and estimated data for the Clemence, Hargreaves-Samani, random forest, support vector machine and artificial neural networks models. It can be observed from the Table 2 that the values of MSE for all the models range from 0.027 to 23.38. RMSE values range from 0.52 to 4.84 which shows a good comparison because the values are close to zero since the values for RMSE ranges from 0 to infinity. The coefficient of determination R^2 ranges from 0.40 to 0.99 for all the models. This indicate that deep learning and machine learning techniques perform better for the Mutale area. And the values of MAE ranges from 0.50 to 3.84 for all the models.

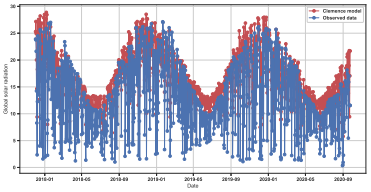

Figure 1. Clemence model

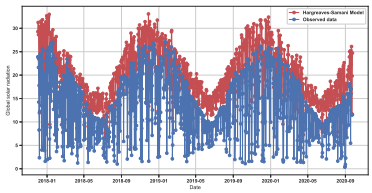

Figure 2. Hargreaves-Samani

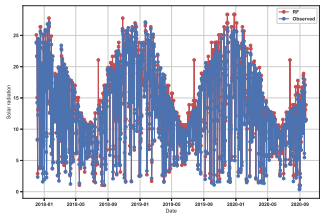

Figure 3. Random forest

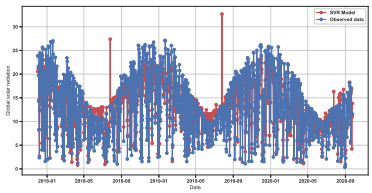

Figure 4. Support vector machine

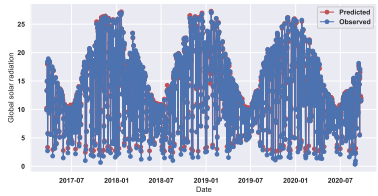

Figure 5. Artificial neural network

Table 2. Statistical errors obtained from different models

Models	MSE	R^2	RMSE	MAE
Clemence	23.38	0.44	4.84	3.84
Hargreaves-Samani	51.47	0.40	7.17	6.45
Random Forest	1.15	0.97	1.07	0.72
Support vector machine	7.87	0.81	2.81	2.12
Artificial neural network	0.27	0.99	0.52	0.50

6. Conclusion

In this study, empirical models, machine and deep learning techniques were employed to estimate the global solar radiation for Mutale area in Limpopo province. Statistical analysis was also utilized to determine the performance measures of the models. It can then been observed that the performance of the empirical models is low compared to machine learning and deep learning. Random forest and support vector machines together with artificial neural networks seems to be the best model for Mutale area since the coeeficient of determination ranges form from 0.81, 0.97 and 0.99. If the values of R^2 is close to 1 then the model is said to be a perfect fit.

7. Acknowledgments

I would like to thank God for giving me strength and wisdom during this study. I wound like to extend my acknowledgement to my supervisor Dr TS Mulaudzi and my co-supervisor Mr N Mphephu and Dr NE Maluta for all their untirelessly effort and encouragement. I would also like to thank university of vends and ARC for providing with the data to carry out this study. Lastly, i would thank my family and friends for all their endless support.

8. References

[1] Dorian, J., Franssen, H. & Simbeck, D. Global challenges in energy. *Energy Policy*. **34**, 1984-1991 (2006)
[2] Kessides, I. Chaos in power: Pakistan's electricity crisis. *Energy Policy*. **55** pp. 271-285 (2013)
[3] Panwar, N., Kaushik, S. & Kothari, S. Role of renewable energy sources in environmental protection: A review. *Renewable And Sustainable Energy Reviews*. **15**, 1513-1524 (2011)
[4] Khambalkar, V., Katkhede, S., Dahatonde, S., Korpe, N. & Nage, S. Renewable energy: an assessment of public awareness. *International Journal Of Ambient Energy*. **31**, 133-142 (2010)
[5] Kearns, M. & Nevmyvaka, Y. Machine learning for market microstructure and high frequency trading. *High Frequency Trading: New Realities For Traders, Markets, And Regulators*. (2013)
[6] Ampratwum, D. & Dorvlo, A. Estimation of solar radiation from the number of sunshine hours. *Applied Energy*. **63**, 161-167 (1999)
[7] Aad, G., Abat, E., Abdallah, J., Abdelalim, A., Abdesselam, A., Abi, B., Abolins, M., Abramowicz, H., Acerbi, E., Acharya, B. & Others The ATLAS experiment at the CERN large hadron collider. *Journal Of Instrumentation*. **3** pp. S08003 (2008)
[8] Hargreaves, G. & Samani, Z. Reference crop evapotranspiration from temperature. *Applied Engineering In Agriculture*. **1**, 96-99 (1985)
[9] Clemence, A. & Mitrofanis, J. Cytoarchitectonic heterogeneities in the thalamic reticular nucleus of cats and ferrets. *Journal Of Comparative Neurology*. **322**, 167-180 (1992)
[10] Shmilovici, A. Support vector machines. *Data Mining And Knowledge Discovery Handbook*. pp. 231-247 (2009)
[11] Tyagi, A. & Chahal, P. Artificial intelligence and machine learning algorithms. *Research Anthology On Machine Learning Techniques, Methods, And Applications*. pp. 421-446 (2022)
[12] Ehrlich, P. & Holdren, J. Impact of Population Growth: Complacency concerning this component of man's predicament is unjustified and counterproductive.. *Science*. **171**, 1212-1217 (1971)
[13] Ren, Y., Zhang, L. & Suganthan, P. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*. **11**, 41-53 (2016)