Research on Prediction of Solar Power Considering the Methods of Statistical and Machine Learning – Based on the Data of Australian Solar Power Market

Puyang Zhao¹ and Wei Tian²
1 The University of Melbourne, Melbourne, Australia
2 The University of Melbourne, Melbourne, Australia
E-mail: puyangz@student.unimelb.edu.au

Abstract. In this paper, we use the methods of machine learning and traditional time series to predict solar power generation, which is based on the Australia Market Data. In the paper, we analyze Ausgrid’s Solar by using long and short-term memory (LSTM) methods and time series models (multiple regression models with related errors) to accurately estimate the parameters of photovoltaic (PV) array models, which is using the data of household electricity consumption from July 1, 2010, to June 30, 2013. The results show that the regression model with correlated errors is better than the machine learning-based LSTM algorithm, which is based on the differential MSE performance. The final prediction accuracy rate is as high as 98%, so the regression model can accurately predict solar power generation.

1. Introduction
Greenhouse gases are mainly released during the combustion of fossil fuels (coal, oil, etc.), and are the main factor causing global climate change [1,2]. At the same time, population and economic growth have led to increased energy demand and increased air pollutant emissions [3]. Therefore, how to reduce greenhouse gas emissions while ensuring the growth of energy demand is a problem that we urgently need to solve [4]. According to the report by the International Energy Agency, solar photovoltaic (PV) is one of the best renewable energy sources with almost no negative impact on the environment [5,6,7]. At the same time, solar energy also has many advantages such as high availability [8,9], low operation and maintenance costs [9,10,11,12]. Therefore, solar energy is one of the best choices for developing renewable and environmentally friendly alternatives.

Australia’s photovoltaic industry has the highest average solar radiation in the world [12-13], and many researchers focus on how to predict its maximum solar power generation capacity. Therefore, in Australia, the rational use of solar energy is of great significance to reducing air pollution and improving economic efficiency. The forecasting results of the solar energy in Australia can have instructive significance for the solar market of other countries. Other countries can learn how to manage and predict their solar market from Australia’s roaring rooftop Solar Market. The solar cell array is a complete power generation unit, which is electrically connected by multiple individual photovoltaic modules and panels [14]. A photovoltaic panel is a collection of individual solar cells connected [15].

The rest of the paper may be structured as follows: In Section 2, related papers about prediction and parameter estimation are summarized and analysed. Section 3 describes the methodology used in this project, including the Solar PV array model, MLR model with Autoregressive Integrated Moving Average (ARIMA) errors, and LSTM. Section 4 presents the experimental results and conclusion, including some figures and tables. Finally, Section 5 introduces the discussion and future research in this field, as well as the limitations of some articles and how to further improve.
2. Literature review
The prediction of solar irradiance and the estimation accuracy of photovoltaic cell parameter values are both critical to minimize energy costs. Accurate prediction of solar irradiance and photovoltaic power generation can reduce the impact of photovoltaic power generation uncertainty. The ARIMA method is usually used to forecast time series [16]. It can obtain useful statistical properties because they can use different sequence parameters to represent multiple different time series. With the latest development of computer computing power, more advanced machine learning algorithms have been developed to analyze and predict time series data [17]. Recently, newly developed algorithms for predicting time series data based on deep learning (such as LSTM) have received a lot of attention. As one of the most advanced recurrent neural networks, LSTM networks have shown many time-series learning tasks with remarkable results. In [18], LSTM is used to model hourly solar radiation. Therefore, in the prediction part, the ARIMA model and LSTM model are used to calculate and find the best prediction model for temperature and solar radiation prediction.

Extensive work has been done to address the problem of parameter estimation for PV cell models, and existing research has deterministic methods and metaheuristic methods. Some examples of deterministic methods involve methods such as least squares [19] and iterative curve fitting [20-21].

Apart from use solar power data, many researchers try to focus on modelling internal model parameters of solar cells with fixing external factors such as temperature. However, solar cells are not only related to the internal solar devices but also to the external lighting conditions. A recent study used Photovoltaic array model from [22] to estimate household level gross PV generation based on external lighting conditions by adjusting and evaluating the parameters of a PV-array model with ambient temperature data collected from the Australian Bureau of Meteorology, solar irradiance data extracted from satellite images, and historical PV generation data [23]. This paper aims to estimate the parameters of the solar PV model and make a prediction for PV generation based on historically relevant ambient temperature, solar irradiance, and historical PV gross generation. More specifically, the goal of this study is to develop an accurate predictive model to predict the solar PV power generation in the future by comparing the machine learning approach and the statistical approach.

3. Methodology
To estimate accurate solar PV generation, a model is produced relying on historical data. It focuses on the development of advanced models to improve the estimation of PV-array model parameters and proposes an accurate prediction model for solar irradiance.

3.1 Sample obtain
Since we lack half-hourly temperature data, we need to estimate half-hourly time data based on the maximum and minimum temperatures available. The hyperbolic tangent fitting model from [24] is used for estimating the temperature data. For solar irradiance data, we only have DNI data, so Global Horizontal Irradiance (GHI) and Diffuse Horizontal Irradiance (DHI) both need to be calculated.

3.2 Modeling approaches
3.2.1 hyperbolic tangent fitting method. For this model, data is collected from the Australian Bureau of Meteorology for the period from 1 July 2012 to 30 June 2013. The analysed data included the daily minimum and maximum temperature data. Data are used to estimate the half-hourly temperature data by Hyperbolic Tangent fitting model.

In the United States Air Force Environmental Technical Applications Center (USAFETAC) study [24], three different methods, simple cosine fit, hyperbolic tangent fit, and cosine fit with variable sunrise, were evaluated to fit the temperature profile given only the maximum and minimum temperatures. By comparing the root-mean-square error (RMSE), the hyperbolic tangent fitting method was the best and estimated the temperature by obtaining plateau rather than mountain peaks [24]. It turns out that the hyperbolic tangent fitting method is sufficiently accurate for estimating hourly temperature data.
Therefore, the hyperbolic tangent fitting method was used to estimate the half-hourly temperature data. The function of the hyperbolic tangent fitting method is then defined as [25]:

\[ T = \begin{cases} 
\frac{\text{max}_1 - \text{min}_1}{2} \times \tanh \left( \frac{\text{time} - 9}{5} \right) + \frac{\text{max}_1 + \text{min}_1}{2} & 0:00 - 9:00 \\
- \frac{\text{max}_1 - \text{min}_2}{2} \times \tanh \left( \frac{\text{time} - 33}{7} \right) + \frac{\text{max}_1 + \text{min}_2}{2} & 10:00 - 24:00 
\end{cases} \]  

(1)

3.2.2 Robledo-Soler model. GHI, DNI and DHI are all needed for estimating solar PV model. However, the data of GHI and DHI are missing. Therefore, the data of GHI and DHI should be estimated by DNI. Since we only have DNI data and zenith angle, we use Robledo-Soler (RS) model to calculate DHI and GHI using geometric calculations. When the propagation through the atmosphere, an outer perpendicular incident irradiance decay is a function of the zenith angle [26]. This model is based on the empirical correlation between the location and the measurement results of astronomical parameters. The higher the zenith angle and the greater the interaction between solar radiation and the atmosphere. Robledo-Soler model shows the relationship between GHI and zenith angle in Equation 2. The Erbs model describes the well-known relationship between the three components in Equation 3 [26-27].

\[ I_G = I_I + I_D \times \cos(Z) \]  

(2)

\[ I_G = 1159.24 \times \cos(Z) \times e^{(-0.0019 \times (90°-Z))} \]  

(3)

where Z is the solar zenith angle, \( I_D \) is DNI, \( I_G \) is GHI, \( I_I \) is DHI.

3.2.3 Multiple regression model with correlated errors. A regression model with more than one independent variable and one dependent variable is called multiple linear regression (MLR) model. The purpose of multiple regression analysis is to predict or control the value of another variable based on the value of several variables, and to know what accuracy this prediction or control can achieve.

3.2.4 Long Short-Term Memory (LSTM). LSTM was first proposed by Hochreiter and Schmidhuber in 1997. Long-short-term memory network (LSTM) is a special kind of Recurrent neural network (RNN). It is widely used in time series prediction problems.

3.2.5 Solar PV array model. A simple PV-array model is used in estimating solar power generation and is based on the simple isotropic model shown in [28]. Ambient temperature and solar irradiance are used in this model, which is based on the relationship between solar power generation and weather conditions (i.e., ambient temperature and solar irradiance). The output of the photovoltaic module is greatly affected by the temperature and solar radiation on the surface of solar panels. Figure 1 provides a structure of the PV-array model.

![Figure 1. The structure of PV-array model.](image)

The model estimates solar power generation by adjusting the parameter values of surrounding environmental factors (i.e., ambient temperature and solar radiation), using the following model in Equation 4:

\[ y = (-a \cdot T + b)(c \cdot I_D + g \cdot I_G + f \cdot I_I) + c_0 + \eta, \]  

(4)

where \( y \) is the actual value of PV gross generation, \( T \) is the ambient temperature, \( I_D \) is DNI, \( I_G \) is GHI, \( I_I \) is DHI, \( a, b, c, d, f \) are model parameters and \( \eta \) is the error term. It is mentioned in [24] that \( c, d \) and \( f \)
are related to the slope and direction of the PV array, while \( a \) and \( b \) are proportional to the surface area of the PV array. \( c_0 \) is the constant term. However, these parameter values are usually unknown, and the only available data are DNI and ambient temperature data. The goal of this project is to find a suitable model to find the suitable parameter values to accurately estimate the solar power generation.

4. Numerical results

This paper divides all the data set into training set and testing set. The training set is used to fit our model and testing set is used for testing the forecast results. We used regression with ARIMA error to model the total power generation data because it can be a potential model for autocorrelated time series data. Table 1 shows results from different regression model with ARIMA errors for three houses. The model with the lowest AICc will be chosen. Therefore, the regression with ARIMA (1, 0, 1) errors should be carefully selected for House 1. The regression with ARIMA (2, 0, 1) errors should be carefully selected for House 2 and 3. After selecting the model, the parameters of the model with ARIMA(1,0,1) is shown in Table 2, Table 3 and Table 4.

| ARIMA errors | House 1 AICc | House 2 AICc | House 3 AICc |
|--------------|--------------|--------------|--------------|
| ARIMA(1,0,1) | -1216.32     | -772.61      | -1372.80     |
| ARIMA(1,0,2) | -1214.87     | -771.99      | -1380.95     |
| ARIMA(3,0,1) | -1214.37     | -791.28      | -1380.43     |
| ARIMA(2,0,2) | -1214.34     | -791.28      | -1380.79     |
| ARIMA(2,0,1) | -1214.43     | -792.81      | -1380.95     |

Table 1. Results from different regression with ARIMA errors models for three houses

Table 2. Estimation of regression with ARIMA(1,0,1) errors parameters of House 1

| Variable     | ar1     | ma1     | intercept | GHI      | DHITemp |
|--------------|---------|---------|-----------|----------|---------|
| Coefficients| 0.7187  | 0.0904  | 0.3378    | 0.1778   | -0.0531 |
| s.e.         | 0.0448  | 0.0558  | 0.0148    | 0.0500   | 0.0337  |

| Variable     | DNI      | GHI      | DHI       | DNI      |
|--------------|----------|----------|-----------|----------|
| Coefficients| -0.3378  | -0.0189  | 0.0557    | 0.3700   |
| s.e.         | 0.0650   | 0.0482   | 0.0339    | 0.0628   |

Table 3. Estimation of regression with ARIMA(2,0,1) errors parameters of House 2

| Variable     | ar1     | ar2     | ma1     | intercept | GHI      | DHITemp |
|--------------|---------|---------|---------|-----------|----------|---------|
| Coefficients| 1.3532  | -0.5135 | -0.4214 | 0.4317    | 0.3451   |
| s.e.         | 0.1369  | 0.1078  | 0.1495  | 0.0137    | 0.0456   |

| Variable     | DNI      | DNI      | GHI      | DHI       | DNI      |
|--------------|----------|----------|----------|-----------|----------|
| Coefficients| 0.0332   | -0.0188  | -0.1646  | -0.0357   | 0.0603   |
| s.e.         | 0.0292   | 0.0523   | 0.0428   | 0.0285    | 0.0509   |
To check the stability of the selected model MLR with ARIMA errors models, the next step is to plot the inverse roots of AR polynomial and MA polynomial. MLR with ARIMA errors models are stable with all corresponding inverse roots of AR polynomial and MA polynomial are in a unit circle. After that, we need use Ljung-Box test to validate the selected model. The results of Ljung-Box test which returns a available p-value indicates the residuals are the white noise.

This paper uses the keras package to develop an LSTM model and uses R TensorFlow as the backend. We use the same training set and testing set as the regression model. We choose a batch size of 32, which has been adjusted according to code running time and balance loss, and predict next steps. Finally, we set epochs = 100, which has been adjusted according to balance loss. We select 'relu' as the activation function of the input layer and the hidden layer and we use 'linear' function for the output layer. Before building the model, we initialize the model as Sequential () in R. Since there is no final rule of thumb for how many nodes should be selected. We need to find the best parameters through trial and error.

| Variable     | ar1    | ar2    | ma1    | intercept | GHI   | Temp   |
|--------------|--------|--------|--------|-----------|-------|--------|
| Coefficients | 1.6610 | -0.7998| -0.8016| 0.1221    | 0.1110|
| s.e.         | 0.0470 | 0.0415 | 0.0510 | 0.0054    | 0.0411|

Table 4. Estimation of regression with ARIMA(2,0,1) errors parameters of House 2

To check the stability of the selected model MLR with ARIMA errors models, the next step is to plot the inverse roots of AR polynomial and MA polynomial. MLR with ARIMA errors models are stable with all corresponding inverse roots of AR polynomial and MA polynomial are in a unit circle. After that, we need use Ljung-Box test to validate the selected model. The results of Ljung-Box test which returns a available p-value indicates the residuals are the white noise.

This paper uses the keras package to develop an LSTM model and uses R TensorFlow as the backend. We use the same training set and testing set as the regression model. We choose a batch size of 32, which has been adjusted according to code running time and balance loss, and predict next steps. Finally, we set epochs = 100, which has been adjusted according to balance loss. We select 'relu' as the activation function of the input layer and the hidden layer and we use 'linear' function for the output layer. Before building the model, we initialize the model as Sequential () in R. Since there is no final rule of thumb for how many nodes should be selected. We need to find the best parameters through trial and error.

Through the results of Figure 2, we can conclude that for this data set, the prediction results of regression model with correlated errors are more accurate.

5. Conclusion
In this paper, we compare the prediction effectiveness of the machine learning method and the statistical method to forecast solar power generation. We estimate and tune the parameters of the PV array model based on irradiance, temperature, and actual historical PV power generation data to estimate the total amount of PV power generation at the household level. Therefore, we could give the conclusion which is that the statistical method is suitable for predicting the total PV power generation of each household.
in the future based only on weather (irradiance and temperature) forecasts. We compare the performance of LSTM and the MLR with the ARIMA errors model, from the simulation results, it was concluded that the regression model with correlated errors performs better for these data sets. Because our data has obvious seasonal trends and the time span is small, using traditional time series models can have better performance. And the overall amount of data is very small, so statistical models are more suitable for predicting solar power generation.

Since the model performs well, we can further apply it to solar forecasting in other countries. In the future, this may power the real-time total solar PV data stream and unlock accurate total PV forecasts. However, there are still some problems and limitations. Since we do not have real temperature data from other houses, we cannot try these methods on data from other houses. Since temperature, GHI and DHI data are artificial estimates, we are not sure whether this method is still applicable to real weather data. Therefore, we will solve the existing limitations and improve the accuracy of prediction in future research.

6. References

[1] Liu, Z., Feng, K., Davis, S. J., Guan, D., Chen, B., Hubacek, K., & Yan, J. (2016). Understanding the energy consumption and greenhouse gas emissions and the implication for achieving climate change mitigation targets. Applied Energy, 184, 737-741.

[2] Leung, D. Y., Caramanna, G., & Maroto-Valer, M. M. (2014). An overview of current status of carbon dioxide capture and storage technologies. Renewable and Sustainable Energy Reviews, 39, 426-443.

[3] Avtar, R., Tripathi, S., Aggarwal, A. K., & Kumar, P. (2019). Population–urbanization–energy nexus: a review. Resources, 8(3), 136.

[4] Dincer, I. (2000). Renewable energy and sustainable development: a crucial review. Renewable and sustainable energy reviews, 4(2), 157-175.

[5] Solangi, K. H., Islam, M. R., Saidur, R., Rahim, N. A., & Fayaz, H. (2011). A review on global solar energy policy: Renewable and Sustainable Energy Reviews.

[6] Peng, J., Lu, L., & Yang, H. (2013). Review on life cycle assessment of energy payback and greenhouse gas emission of solar photovoltaic systems. Renewable and sustainable energy reviews, 19, 255-274.

[7] Kim, H., Park, E., Kwon, S. J., Ohm, J. Y., & Chang, H. J. (2014). An integrated adoption model of solar energy technologies in South Korea. Renewable energy, 66, 523-531.

[8] Bhattacharya, T., Chakraborty, A. K., & Pal, K. (2014). Effects of ambient temperature and wind speed on performance of monocrystalline solar photovoltaic module in Tripura, India. Journal of Solar Energy, 2014.

[9] Silveira, J. L., Tuna, C. E., & de Queiroz Lamas, W. (2013). The need of subsidy for the implementation of photovoltaic solar energy as supporting of decentralized electrical power generation in Brazil. Renewable and Sustainable Energy Reviews, 20, 133-141.

[10] Mundo-Hernández, J., de Celis Alonso, B., Hernández-Álvarez, J., & de Celis-Carrillo, B. (2014). An overview of solar photovoltaic energy in Mexico and Germany. Renewable and Sustainable Energy Reviews, 31, 639-649.

[11] Hosennuzzaman, M., Rahim, N. A., Selvaraj, J., Hasanuzzaman, M., Malek, A. A., & Nahar, A. (2015). Global prospects, progress, policies, and environmental impact of solar photovoltaic power generation. Renewable and Sustainable Energy Reviews, 41, 284-297.

[12] Kabir, E., Kumar, P., Kumar, S., Adelodun, A. A., & Kim, K. H. (2018). Solar energy: Potential and future prospects. Renewable and Sustainable Energy Reviews, 82, 894-900.

[13] Azad, A. K., Rasul, M. G., Khan, M. M. K., Ahasan, T., & Ahmed, S. F. (2014). Energy scenarios: production, consumption and prospect of renewable energy in Australia. Journal of power and Energy Engineering, 2(4), 19-25.

[14] Bahadori, A., & Nwaoha, C. (2013). A review on solar energy utilisation in Australia. Renewable and Sustainable Energy Reviews, 18, 1-5.
[15] Attila Tamas Vekony. Find solar panels in the uk why solar panels in the uk are a great option, 2021. https://www.greenmatch.co.uk/solar-energy/solar-panels.
[16] Hassan, J. (2014). ARIMA and regression models for prediction of daily and monthly clearness index. Renewable Energy, 68, 421-427.
[17] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1394-1401). IEEE.
[18] Qing, X., & Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. Energy, 148, 461-468.
[19] Easwarakhanthan, T., Bottin, J., Bouhouch, I., & Boutrit, C. (1986). Nonlinear minimization algorithm for determining the solar cell parameters with microcomputers. International journal of solar energy, 4(1), 1-12.
[20] Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. Renewable Energy, 105, 569-582.
[21] Karaboga, D., & Basturk, B. (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. Journal of global optimization, 39(3), 459-471.
[22] Mahdavi, N., Braslavsky, J., & Perfumo, C. (2016). Inference of temperature-dependent loads and solar generation capacity from aggregate demand data.
[23] L.O’Neil. (2018). Panel data analytics for estimating the gross solar output of households (Tech. Rep.). CSIRO.
[24] Schaub Jr, W. R. (1991). A method for estimating missing hourly temperatures using daily maximum and minimum temperatures. AIR FORCE ENVIRONMENTAL TECHNICAL APPLICATIONS CENTER SCOTT AFB IL.
[25] Maxwell, E. L. (1987). A quasi-physical model for converting hourly global horizontal to direct normal insolation. Unknow.
[26] Reno, M. J., Hansen, C. W., & Stein, J. S. (2012). Global horizontal irradiance clear sky models: Implementation and analysis. SANDIA report SAND2012-2389.
[27] Erbs, D. G., Klein, S. A., & Duffie, J. A. (1982). Estimation of the diffuse radiation fraction for hourly, daily and monthly-average global radiation. Solar energy, 28(4), 293-302.
[28] Duffie, J. A., Beckman, W. A., & Worek, W. M. (1994). Solar engineering of thermal processes.