MONASH University
Information Technology

HONOURS/MINOR THESIS SUBMISSION FORM – PRE EXAMINATION

You are to submit the following for examination purposes:

- *Honours/Minor Thesis Submission - Pre Examination* form, (electronically) signed by your supervisor.
- The declaration of originality which is embedded in the thesis (usually the page prior to the acknowledgements page) must be signed by you in all copies.

SECTION A: Personal Details

Student ID Number: 28397371

Surname: HSU Given Name(s): CHAO - KAI

Address: 1705 639 Lonsdale St. Melbourne 3000

E-mail: chsu0002@student.monash.edu

Telephone: (Business) 0481963541 (Home) ____________

Degree Title: Master of Data Science Course Code: C6004

Thesis Title: Spatio - temporal analysis of smart meters data

School: Monash University FIT

I have submitted a PhD scholarship application at Monash this year *(tick one of the boxes)* **TRUE** □ **FALSE** ☑

[✓ I agree] / do not agree (please circle as appropriate) that this thesis be made available for downloading on a university repository.

Signature: CHAO KAI HSU

SECTION B: Thesis Submission for Examination

Date of Submission: 9 / 11 / 2018

The thesis being submitted is worthy of examination.

Supervisor: Christoph Bergmeir Signature: C. Bergmeir

The information on this form is collected for the primary purpose of processing your thesis for examination. Other purposes of collection include proving your thesis to an internal examiner, attending to administrative matters, corresponding with you, and to prepare statistical analyses. If you choose not to complete all the questions on this form, it may not be possible for the Faculty of Information Technology to accept your thesis. Personal information may be also be disclosed to an external examiner You have a right to access personal information that Monash University holds about you, subject to any exceptions in relevant legislation. If you wish to seek access to your personal information or inquire about the handling of your personal information, please contact the University Privacy Officer on 9905 6011.

Office use only

Keyed: ____/____/____ Initials: ________ Receipt entered on Moodle: ____/____/_____ Variations Log:

Version: 15 February 2018

Monash University

SPATIO-TEMPORAL ANALYSIS OF SMART METERS DATA

This thesis is presented in partial fulfillment of the requirements for the degree of < Master of Data Science > at Monash University

By:
Chao Kai Hsu

Supervisors:
Christoph Bergmeir
Lachlan Andrew

Year:
2018

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the work of others has been acknowledged.

Signed by .... CHAO KAI HSU

Name CHAO KAI HSU
Date 9. 11. 2018

Solar power forecasting

SPATIO-TEMPORAL ANALYSIS OF SMART METERS DATA

Chao Kai Hsu

Faculty of Information Technology | Monash University

Table of Contents

0. Abstract.....3

1. Introduction.....3

2. Solar forecasting.....5

2.1 External factors that change the solar output.....5

2.2 Data used in solar forecasting.....6

2.2.1 System data.....6

2.2.2 Historical data.....6

2.2.3 Numerical weather prediction (NWP) data.....6

2.2.4 Sky imager and Satellite imaging for cloud tracking.....6

2.3 Types of forecasting model for solar power prediction.....7

2.3.1 Physical Method.....7

2.3.2 Time series statistical Method.....8

2.3.3 Hybrid Method.....9

2.4 Time horizon:.....10

2.4.1 Intra-hour forecasting:.....10

2.4.2 Intra-day forecasting:.....10

2.4.3 6 hours ahead or longer:.....11

2.5 Existing models.....11

3. Methodology.....12

3.1 Base methods.....12

3.1.1 ARIMA.....12

3.1.2 Random Forest.....14

3.1.3 Seasonal Naïve (Snaïve).....16

3.2 Modelling seasonality.....16

3.2.1 Dummy variables.....16

3.2.2 Fourier series.....17

3.3 Spatio-temporal analysis on cloud cover data.....18

3.3.1 Dynamic time warping.....18

3.3.2 DTW on cloud data.....21

3.4 The overall procedure.....24

3.4.1 In ARIMA model:.....24

3.4.2 In Random Forest model:.....25

4.Experimental study.....27

4.1 Datasets.....28

4.1.1 Solar output data:.....28

4.1.2 Cloud cover data:.....28

4.1.3 Final dataset:.....29

4.2 Error measures **.. 30**

4.2.1 Root Mean Square Error (RMSE): 30

4.2.2 Symmetric Mean Absolute Percent Error (sMAPE): 30

4.2.3 Mean Absolute Error (MASE): 31

4.2.4 Measure the method performances: 31

4.3 Results on the dataset **.. 32**

4.3.1 Results on 5 days testing: 32

4.3.2 Results on 15 days testing: 32

5. Conclusions 33

References 35

0. Abstract

To provide a reliable forecasting is crucial for this highly variable energy to be used in an efficient way, power supply company are collaborating with researcher that using types of approaches and sources of data to target a higher prediction accuracy for the solar generate forecasts, produce different time horizon forecasting for various aspects of managing of the solar power integrated into the electricity supply scheme.

The number of data sources have been adopted in solar forecasting to improve the prediction performance. The more data to be considered the higher the model's performance is expected. However, the model with required different sources of data as input values can also cause higher cost in the collection of the data and higher complexity the model has. Hence, to have a balance in the model’s performance and the cost is also important for a forecasting model to be implemented.

In this paper, we presented a method to bring similar cloud patterns from the neighbouring systems as input data to the time series forecasting model. A state-of-the-art approach uses Dynamic Time Warping on the cloud data which is estimated from the solar production data of households PV system without using other sources of data. The results show a consistent improvement of forecast accuracy both on the statistical model (ARIMA) and machine learning model (Random Forest) on 5 days testing for total 100 residential solar generation systems, while the improvement is not observed on Random Forest model in a 15-days testing.

1. Introduction

Accurate forecasting for solar generation in different temporal scales is essential to achieve the balance of the power grid with high penetrated solar generation systems. There are techniques to build predictive models for the solar power prediction with the different source of data as inputs to the model.

Solar photovoltaic (PV) system is a solar energy converting system which exhibits the photovoltaic effect [1]. The technique uses semiconducting materials to convert solar radiation into electrical energy [1]. Australia has the highest average solar irradiance of any continent in the world, making it a prime place for solar power development and utilisation [2]. It is predicted that Australia solar capacity will double by the end of 2018 with more and more residential rooftop PV system adopted by the citizen and more massive solar plant been constructed all over the country [2].

PV power production mainly depends on the solar irradiance with other meteorological data as potential factors such as temperature, wind speed, humidity and precipitation that bring different degrees of influence on the PV power generation. The amount of solar irradiance, however, is much more correlated to the actual PV output than other factors mentioned above. The production of solar power is highly variable mostly because of the variability of the cloud cover. As cloud moving in front of the sun can cause the change the solar irradiance a ground level, lead to the decreasing and increasing variation of the PV production of the system. The effect that the cloud brings to the PV system in the power production, depending on the volume and speed, can be intense and dramatic. Research has shown that by estimating the cloud motion and its structure improved PV output forecasting. Cros et al. [3] demonstrated a forecasting method that utilises satellite images to extract cloud motion vectors for solar power forecasting. Quesada-Ruiz et al. [4] used sky imaging, a novel way to track the cloud movement to forecast hourly Direct Normal Irradiance (DNI).

The rapid development of the distributed photovoltaic (PV) systems with the increasing deployment of the smart meters in residential and commercial buildings worldwide in recent years, make the solar generation data much more accessible and with finer temporal and space scales than it was previously. The availability of the spatial and temporal data collected from PV systems in the power grid networks makes it possible to identify the spatiotemporal patterns by using the Spatio-temporal analysis technique and be used for the forecast.

The Spatio-temporal analysis is used on the data that contains spatial and time properties, the benefit of the Spatio-temporal analysis over purely spatial or time series analysis as it investigates both time correlations and space correlations when modelling the data. The method has been commonly used in the study of different filed such as epidemic disease [5], environmental pollution [6] and city crime analysis [7].

In the research of solar power forecasting, we proposed a Spatiotemporal-like approach that uses Dynamic Time Warping (measuring the similarity between two times series and calculate optimal match path) on the cloud cover series to detect the similar pattern of the cloud movement. And to integrate those warping paths as external input data to improve the forecast.

The rest of the paper is structured as follows. In Section 2, we briefly discuss the existing models of solar forecasting. In Section 3, we present the proposed data-driven approach in detail. Section 4 shows the implementation with the results, and to conclude the research in Section 5.

2. Solar forecasting

2.1 External factors that change the solar output

The external factors that influence the production of solar power including Irradiance, Shading, Soiling and Temperature [8, 9], while solar irradiance is the critical factor among of the four. The solar irradiance is a measurement of the amount of sunlight that reaches to a particular region. And the solar panel received the reaching sunlight that converts it to the power. The higher the solar irradiance present, the more the solar power be generated.

The intensity of the solar irradiance depended on the sun's angle and position. It gives a strong seasonality for solar irradiance. Summer, on average, has the strongest and longer solar irradiance, followed by Spring and Autumn. It has the lowest and shortest value in Winter in general. Apart from the seasonal variation, solar irradiance has a daily seasonal pattern, and the present of the cloud and its moving cause the change of the amount of sunlight to be received by the solar panel, resulting in the fluctuation of power production within a day. The high variation of solar power due to the cloud is the most uncontrollable factor. [10] studied the influence of the cloud on the solar panel power production and conclude that the cloud in low level has the most detrimental effect on solar generation, and the observation of the cloud type and its density are also important on the determination the solar power production as it may bring substantially effect on the transmittance of sunlight through the cloud.

2.2 Data used in solar forecasting

2.2.1 System data

System data contains information of PV system's capacity, location, orientation, tilt and shading from nearby objects. The system data is essential for the physical model to be able to convert the irradiance prediction to system output of the solar productions based on its specification and related information to the system itself.

2.2.2 Historical data

Past observations of solar power generation, records of local measurements (temperature, humidity, weather condition, etc. from meteorological databases, provide a good source for statistical analysis/model to discover the patterns of seasonality and trend/cycle which are the main components in generating solar output forecasts.

2.2.3 Numerical weather prediction (NWP) data

The Numerical weather prediction is a process of creating mathematical models that describe the way the atmosphere and oceans change using sets of equations based on some physical principles to produce the weather forecasts. A simulation process is done through the help of the mathematical models with the computer to provide the prediction of the weather states based on current observations of the weather. The forecast output including temperature, humidity, wind, cloud and other meteorological elements. The NWP data is crucial for solar forecasting in time horizon above 6 hours to days ahead.

2.2.4 Sky imager and Satellite imaging for cloud tracking

Satellite imaging and sky imager are applied for cloud detection. The cloud location and motion vector are estimated by the analysis of the consecutive images taken from Satellite or sky imager to develop the forecast of cloud pattern which can be further used in the prediction of solar irradiation and PV output as the cloud pattern is an essential factor that influences the generation of the solar power.

The concepts of satellite imaging and sky imager are similar to each other. Sky imager, a ground-based digital camera device, obtains the high-quality images of the sky near the forecasting area. [11] demonstrated a novel method that uses the hybrid algorithm, making the use of correlation and local feature analysis from the sky images for the cloud motion estimation. [12] presented techniques for estimating cloud motion and quantifying the cloud stability by analysing the sky images to produce the forecast of cloud location up to 15 min ahead. Satellite imaging, a satellite-based sensor provides the photos of the sky in a broader spatial range. [13] built a model based on the use of satellite images with a proposed algorithm that computes the cloud speed and estimates the influence of the cloud cover to the solar irradiance and fed with artificial neural networks to generate global horizontal irradiance. [14] created an hourly solar irradiance forecast that applied the analysis of images from satellite to characterise the cloud cover index and the derived index was used as input to Exponential Smoothing State (ESS) space model to produce cloud cover index forecast in the next time step. Solar irradiance is estimated by utilising the predicted cloud cover index to a multi-layer perceptron (MLP) model. Tracking cloud from both satellite imaging and sky imager give the opportunity to generate instant estimation to monitor the cloud movement which is helpful for the solar power forecast in time horizon within 6 hours.

2.3 Types of forecasting model for solar power prediction

Solar forecasting technique can be mainly divided into three categories: physical methods, time series statistical methods and hybrid methods.

2.3.1 Physical Method

In the physical process, numerical weather prediction (NWP), cloud observations from the use of satellite or sky imager and the measurement of physical data such as temperature, wind, humidity and pressure are used with the analytical equations to model the solar system's performance on solar power production. Effort based on the physical methods is to study the interactions between the solar radiations and those physical data to generate accurate forecasts for solar irradiance. The predicted values of solar irradiance are converted into solar productions based on systems data of PV. The approach is denoted as ‘‘white box" method. The main advantage of the Physical method is that no historical data needed in the process of generating the prediction. However, the main disadvantage is that the physical model is highly depended on NWP data while the lack of sufficient spatiotemporal resolution has been reported as a cause of error in the prediction when using this approach [15].

2.3.2 Time series statistical Method

The time series statistical methods are based on historical and do not require information about the solar generation system. A data-driven approach which can extract the pattern from the past data and to predict the solar output performance by the use statistical method/model of time series analysis or the machine learning techniques automatically. As rely on the past data, the quality of those historical data become essential for accurate predictions. Statistical methods require large historical dataset for method/model be able to discover the pattern existing in the past data, especially for machine learning models to get well trained by feeding sufficient data for the model to learn from the data. This approach is called a ‘‘black box” method.

Below are some commonly used time series analysis techniques for forecasting:

- AutoRegressive Integrated Moving Average (ARIMA)

  ARIMA stands for AutoRegressive Integrated Moving Average, is a general statistical model for time series data analysing and forecasting. We use ARIMA model as one of our model bases of the proposed approach. Hence, we present the detail of the ARIMA model in Section 3.3.1.

- Exponential smoothing

  Exponential smoothing methods is another generally used statistical model in the time series forecasting. The forecasts are produced by using exponential smoothing methods with weighted averages of the past observations and the weights decaying exponentially as the observations get older. In other words, the recent observations are given relatively more weight than the older observations that contribute to the value as forecasts. Exponential smoothing is a time series method that bases on the description of the trend and seasonality in the data. They are different combinations of ways to handle the different types of Trend, Seasonality and Error component observed in the given time series. This framework generates a reliable forecast that can be used for a wide range of time series efficiently.

- Artificial Neural Networks (ANNs)

  ANNs is a data-driven model that is one of the main tools used in machine learning. The system is inspired by the operation of the neuron in the human brain that replicates the way that human learn. Neural networks consist of input and output layers, as well as the hidden layer(s). Each layer is consisting of units as the neuron with the parameters that transform input from the previous layer into some value for the next layer to use and generate the value in the output layers. In the process of learning from the data through the designed architecture of Neural networks, a well-trained system can be built that obtaining pattern from the input data which can be used in the prediction in a desired level of accuracy. This artificial intelligence technique has been proven useful in a vast variety of fields. The ANNs are one of the most used machine learning algorithms in the prediction of solar power with some modified ANNs method be invented [16].

- Support vector regression (SVR)

  Support vector regression is another commonly used method in the prediction of the time series and has been widely used in solar forecasting. Support vector regression (SVR) is extended from Support Vector Machines (SVM). SVM is a supervised learning algorithm in the application of classification. In general, it is used with binary classes even if they are not linearly separated in 2-dimensional space (2D), this is achieved by introducing a trick that transforms non-linear separated classes in 2D into a higher dimensional space where the classes become separable linearly. The SVR model uses the same principles as SVM, while SVM as a classifier performs classification, SVR performs regression analysis as a regressor, generating continuous variables as output in perdition tasks. This advanced machine learning technique has been proved to obtain good performance in prediction of series data.

2.3.3 Hybrid Method

The hybrid method also denoted as blended, and ensemble method is an approach to have combinations of two or more forecasting techniques to improve the accuracy of the forecast. The idea of using the hybrid method to build the prediction model is to overcome the deficiencies of a stand-alone procedure and to utilise the advantage of an individual approach by merging them to enhance the forecasting performance to reduce the forecast errors which cannot be achieved when using a stand-alone model. Studies have shown that integrating multiple methods can outperform an individual forecast method [16]. The model that has the combination of the physical process with the statistical method is called a “grey box” method. Note that a hybrid method can also be built with the combination of the statistical model without including physical way.

2.4 Time horizon:

The forecast horizon is the length of time which the forecast is generated. In solar power forecast, different lengths of forecasting horizon are made base on the type of management for power grid with the solar power generation systems integrated such as the management of the power grid including the maintenance of the grid stability, power supply scheduling or the decision making of electricity trading and system installation and so on. The forecast horizon can be classified into Intra-hour forecasting, Intra-day forecasting and 6 hours ahead or more extended forecasting.

2.4.1 Intra-hour forecasting:

Intra-hour forecasting also denoted as very short-term forecasting covers time horizons for the prediction from minutes to an hour. It is vital in the assurance of grid stability and quality and to better maintain the balance for the demand and supply in power grid which is crucial for a grid with lower quality of power supply where the highly penetrated solar systems are present. The main factor causing the change of solar production in this timeframe is the presence of the clouds. As mentioned above, cloud moving in front of the sun can create the change the solar irradiance at ground level, leads to the decreasing and increasing change of solar generation. Hence, the use of the technique and method to understand cloud movement and to give further prediction is essential for accurate intra-hour forecasting for solar power.

2.4.2 Intra-day forecasting:

Intra-day forecasting is also called short-term forecasting. It covers time horizons from 1 to 6 hours. This time frame of prediction is critical for grid operators to control different load zones efficiently and to trade power between different time zones to maximise the use of the generated power [17]. Clouds estimation and prediction also play an important role in producing reliable forecasting for solar in this timeframe.

2.4.3 6 hours ahead or longer:

6 hours to days ahead on hourly production is used to meet the need in the unit commitment and managing of the transmission and trading. Medium-term from 1 week to month on day production is for system optimisation and planning for the maintenance. Long-term provide predictions on monthly or annual output this can be used to select desirable site or area for government or investor to assess resources of solar power to deploy large scale of solar plants [18].

For hourly forecasting more than 6 hours ahead to day-ahead, information of the cloud prediction still plays an important role. Satellite image with NWP can provide a larger scale of cloud pattern tracking and forecasted data for the model as input to generate the forecast, whereas a smaller scale for the cloud tracking from sky imager shows its limitation of providing a reliable estimation of the cloud for the solar predictive model. For a longer timeframe of forecasting in weekly, monthly and annual basis, the historical output data and meteorological records are more suitable for use as an input variable of the prediction model.

2.5 Existing models

The study [19] demonstrated an approach of using the physical method to estimate the output of photovoltaic plants. [20] used Autoregressive Integrated Moving Average model with measured irradiance and observed cloud cover to produce next hour solar irradiance forecasts that show the advantages of considering cloud effects. [21] implemented an Artificial Neural Network – Multi-Layer Perceptron architecture to day-ahead solar irradiance forecasts with daily solar irradiance and daily air temperature as input variables. Seasonal- Autoregressive Integrated Moving Average and Artificial Neural Network with multiple inputs are also presented in [22] for a day, and intra-day PV output prediction shows a better performance of SARIMA model for an intra-day forecast than day-ahead forecast when previous day shows an irregular pattern. [23] created the hybrid model: PHANN which combined ANN with a physical clear sky solar radiation model for day-ahead hourly using NWP data including cloud cover to the solar forecasting. [24] carried out a physical modelling approach with the forecasted irradiance received from numerical weather prediction (NWP), and cloud motion data (CMV) to obtain PV power output forecasts and compared with a proposed hybrid model that used support vector regression.

3. Methodology

We describe the detail of our proposed methods in this section. Firstly, we provide a brief introduction to the selected forecast methods as our model base. Second, we illustrate of using Furrier series in the forecast model for modelling the seasonality of the solar generation data. Third, we present the proposed approach of generating the cloud pattern information by using DTW in detail and finally summarise the overall procedure of the implementation.

3.1 Base methods

3.1.1 ARIMA

ARIMA an acronym that stands for Autoregressive Integrated Moving Average where “AR” is Autoregression, “I” is Integration and “MA” is Moving Average. The model is a class of statistical model that widely used for time series analysis and forecasting [25].

For Autoregression (AR), it is a model that forecasts the variable against itself, unlike a multiple regression model that uses a combination of predictors to forecast variable. In another word, an autoregressive model forecast value of a time series by regressed on the previous values of the same time series. For example, if we want to predict y this year using the measurements of the global population in the previous three years (yt−1, yt−2 and yt-3) using the autoregressive model. It would be:

$$y_t = c + \phi 1 y_{t-1} + \phi 2 y_{t-2} + \phi 3 y_{yt-3} + \epsilon_t,$$

where ϵt is white noise that shows no autocorrelation on the time series data. The model is a third-order autoregression since the value at time t is forecasted as a linear combination of values at time t-1, t-2 and t-3. In general, the autoregressive model of order p can be written as:

$$y_t = c + \phi 1 y_{t-1} + \phi 2 y_{t-2} + \cdots + \phi p y_{t-p} + \epsilon_t,$$

where p is the number of the preceding observations in the series that are included in the used of the prediction of the value at present, defined as a multiple linear regression where value at any time t is a linear function of the value at times t−1, t−2, ⋯,t−p. We recognise to this as an AR(p) model, and hence, the preceding model is written as AR(3).

...

For Moving average (MA), the model uses the combination of the past errors of the forecast as a regression-like model to predict the value of the present time [26], rather than using the previous values of the time series data when doing in the forecast as the AR model does. Take the same example stated in the preceding AR model, the prediction of global population y at time t with a 3rd order moving average model is calculated from a linear combination of forecast errors at time t-1,t-2 and t-3, and can be written as

$$y_t = c + \epsilon t + \beta 1 \epsilon_{t-1} + \beta 2 \epsilon_{t-2} + \beta 3 \epsilon_{t-3},$$

where ϵt is white noise. The model can be denoted by MA(3). Thus, the Moving average model of order q can be written as MA(q):

$$y_t = c + \epsilon t + \beta 1 \epsilon_{t-1} + \beta 2 \epsilon_{t-2} + \cdots + \beta q \epsilon_{t-q},$$

We can think of the model as a weighted moving average of the few past errors of the forecast.

...

For Integration (I), it is the use of computing the differences between consecutive observations on the time series data to make the series stationary, which is known as differencing. A stationary time series is the one with the constant of mean, variance and autocorrelation all over time. Thus, it has the properties of no trend, no seasonality as it does not depend on the time when the series is observed. Modelling on a stationarised time series make the prediction task relatively easy as the stationarised series can be predicted by rely on its statistical properties as it will remain the same as they have in the past. It is important to stationarise the time series when fitting an ARIMA model [25].

...

A non-seasonal ARIMA model is obtained when we are combining differencing with the autoregressive model and the moving average model. And the model can be written as:

$$y_t = c + \phi 1 y_{t-1} + \phi 2 y_{t-2} + \cdots + \phi p y_{t-p} + \beta 1 \epsilon_{t-1} + \beta 2 \epsilon_{t-2} + \cdots + \beta q \epsilon_{t-q} + \epsilon_t,$$

where y_t is the differenced series, and the lagged values of y_t and the lagged values of forecast errors as predictors on the right-hand side of the equation. The model is denoted by ARIMA(p,d,q), where

- p: the order of the lagged values been considered in the autoregressive part.
- d: the number of differencing used on the original time series to make it stationary, which is called the degree of differencing.
- q: the order of the lagged forecast errors been included in the moving average part.

The ARIMA model can be seasonal. In the time series data, the seasonality is a regular pattern time with the same period that repeats over time. The period can be yearly or less than a year such as quarterly, monthly, weekly, so on and so forth. For example, in a monthly time series, there is a yearly seasonality that particular months tend to have a higher value than the rest other months. The Seasonal ARIMA hence can be written as follow [26]:

$$\text{ARIMA}(p,d,q)(P,D,Q)_m$$

where (p,d,q) is Non-seasonal part of the model, and (P,D,Q)m is the Seasonal part of the model. And m is the span of the periodic as the seasonal behaviour. For example, both m= 4 in quarterly data and m=12 in the monthly data represent the yearly seasonality. The seasonal ARIMA is to include the modelling of the seasonality that presents in the series, which is the use of the same time step but in the previous period(s) to predict the time step in present period when doing the forecast.

3.1.2 Random Forest

Random Forest is a machine learning algorithm. Instead of using the form of the mathematical equations to formalise the relationships between variables in the forecast as the statistical method, machine learning is a method in the data analysis to build the analytical model that learn from data automatically. It is a branch of artificial intelligence technique based on the idea of making systems learn from the data with minimal human intervention to identify the patterns and make decisions [27].

The Random Forest, one of the most effective and powerful models in machine learning for the predictive task, is a supervised learning technique that learning data through a set of input and output data which are called the training set and to get a learned function that maps input to output on the test data. It is generally used in both classification task and regression task, and also can be used in the prediction of the time series data.

A random forest model can be written as:

$$g(x)= f_0(x)+ f_1(x)+ f_2(x) +\cdots+ f_k(x)$$

Model in training step output the class which is the mode of the classes generated from the individual decision trees for the classification and the mean of the predictions made from the trees for the regression tasks.

The technique of using multiple models in the purpose of obtaining better performance of the prediction is called ensemble learning method to overcome the low predictive power from a simple Decision tree. The ensemble learning method has another advantage which is the avoid/reduce model's overfitting which the overfitting happened when the model that fits well in the training data set but perform badly in the test data set. The one difference in the random forest to other ensemble learning methods such as bagging is that each base model is constructed independently by only a random selection of predictor of the input data. The process provides a flexible and efficient way that produces excellent results that tend to be the most accurate machine learning tool in the prediction task [28].

The Random Forest model in R package [29] is set as:
randomForest(y= , x= , ntree = , mtry = , nodesize =)

where y is the response, x is the predictors. And there are Hyperparameters for the model including

- ntree: number of trees
- mtry: number of variables sampled in each split
- nodesize: minimum size in terminal nodes

3.1.3 Seasonal Naïve (Snaïve)

In time series forecasting, it is critical to use persistence or naïve method to provide lower-bound of the prediction performance as a baseline that give an idea of how well the proposed models perform on the dataset and to ensure that we are not wasting time on the more sophisticated models that are not predictive on the dataset when it performs worse than the straightforward method by just use the last observation as the forecast value.

Nevertheless, it does not make sense to use the persistence or naïve forecast when on the time series data that has the seasonal component. It is better to use the Seasonal Naïve method as a baseline model when forecasting on a highly seasonal time series data. The Seasonal Naïve give a prediction to be equal to the last observation of the same season. For example, given a time series data of Quarterly beer production in Australia, using Seasonal Naïve for the prediction of the beer production for Q1 2018 (where Q1 means the first quarter), the forecast value is equal to the last observed value of the same season which is Q1 in 2017. With monthly data, the prediction of the all future value of May is equal to the last observed May value. The formal notation of the model that forecast for time t+h can be written as [30]:

$$y_{t+h|t} = y_{t+h-km},$$

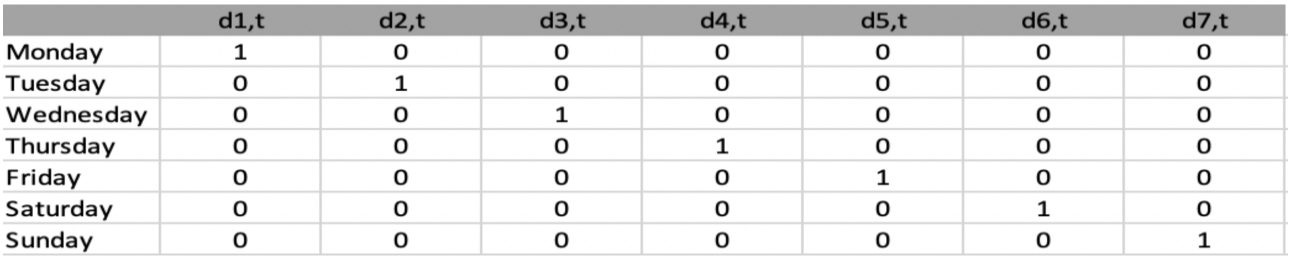
where m is the span of the seasonal period and k is the smallest integer that greater than $\frac{h-1}{m}$.

3.2 Modelling seasonality

3.2.1 Dummy variables

The dummy variable is used when the predictor is a categorical variable that taking only two values. For example, when forecasting daily sale of the beer in Melbourne city, Australia, we would like to take into account of whether the day is a public holiday or not for the model to distinguish the difference between the day that people work, and the day people don’t work. However, takes the value of “yes” or “no” on where the day is a public holiday or not brings the categorical value which cannot directly be used as a predictor in the forecast model. Thus, this situation will be talked by using the dummy variable which takes value "1" that corresponding to “yes” and value "0" corresponding to “no” instead, to replace the categorical variable. The dummy variable is also called indicator variable. If the categorical variable as a predictor that taking more than two values, it is required to use multiple dummy variables where the total number of variables to handle this is one fewer than the total number of the values takes in the predictor.

Takes dummy variable for modelling the seasonality as called seasonal dummy as an example, suppose we want to forecast the daily data that count the effect of the day in a week as a predictor, as there 7 days in a week, this categorical variable can be replaced with 6 dummy variables that as a combination of value of 1 and value 0 to represent the 7 different days in a week [Table 1]. And those dummy variables then can be input as regressors to the model to model the weekly seasonality of the time series data.



	d1,t	d2,t	d3,t	d4,t	d5,t	d6,t	d7,t
Monday	1	0	0	0	0	0	0
Tuesday	0	1	0	0	0	0	0
Wednesday	0	0	1	0	0	0	0
Thursday	0	0	0	1	0	0	0
Friday	0	0	0	0	1	0	0
Saturday	0	0	0	0	0	1	0
Sunday	0	0	0	0	0	0	1

Table 1: dummy variables to represent the day of week.

3.2.2 Fourier series

Use the Fourier term is an alternative way to the dummy variables to modelling the seasonality, especially for a long seasonal period. Fourier term is also known as Fourier series. A series of sine and cosine terms of the right frequencies has been shown to be able to approximate any periodic pattern which can be used for the seasonal patterns [31].

Fourier term for modelling seasonality where m is the seasonal period are given by sequentially as below:

$x1,t = \sin(\frac{2\pi t}{m})$, $x2,t = \cos(\frac{2\pi t}{m})$, $x3,t = \sin(\frac{4\pi t}{m})$, $x4,t = \cos(\frac{4\pi t}{m})$ and so on.

If we have annual seasonality for monthly data, we will get the same forecast as 11 dummy variables do as using the first 11 of these Fourier series as the predictor variables. The advantage of using Fourier term rather than dummy variables in the modelling of the seasonality is that in general, seasonality can be represented by using fewer predictors of the Fourier series than using dummy variables, especially when the m is large. It brings the benefit of not increasing the computation complexity as fewer variables to be used. It is often to use Fourier term with a regression model which is called a harmonic regression approach [32]. The seasonal pattern is handled by using the Fourier term.

In the implementation of the solar forecast, the daily seasonality presented in the solar data that has a long seasonal period which the m = 48. As the long period, seasonal ARIMA model may not be suitable as it is designed for shorter periods such as m=12 on monthly data, m=4 on quarterly data and not provide the efficient on the longer periods. Thus, on this situation, we use harmonic regression approach that uses Fourier term with the ARIMA model where the seasonal pattern is modelled by the Fourier term and the short-term dynamic time series are handled by ARIMA model itself. It is the method preferred by researchers when handling the long seasonal period with ARIMA model. The model can be written as [33]:

$$y_t = \alpha + \sum_{k=1}^{K}[\alpha_k \ \sin\left(\frac{2\pi t}{m}\right) + \ \beta_k \cos\left(\frac{2\pi t}{m}\right)] + N_t \ ,$$

where N_t is an ARIMA process**.** The value of K can be chosen by minimising the information criterion (such as AIC, BIC and AICc) [34] which measuring the quality of statistical model that fitted to the given set of data [35].

3.3 Spatio-temporal analysis on cloud cover data

3.3.1 Dynamic time warping

In time series analysis, Dynamic time warping (DTW) is an algorithm that measures the similarity of the two temporal sequences that may vary in speed, first proposed by [36]. This nonlinear distance measure technique allows the sequence of the two time series to be stretched along the time to minimise the distance between them. Through the distance measure of the time sequence data, the DTW algorithm returns the points matching that helps to quantify the similarity of the time series and can be used to perform classification and discover the corresponding pattern between the two time series.

As the DTW algorithm that is capable of measuring the similarity of the two sequence that differs in time to detect the similar pattern between, it has been widely applied in different fields for the propose of pattern matching for decades despite its high computational complexity which is O(n^2). Some of the well-known applications including Automatic Speech Recognition to tackle the pattern in different speaking speeds [37, 38], Handwriting Recognition and Gesture Recognition tasks [39: 41]. Other applications contain biometrics, medicine, entomology, anthropology, finance and so on. For data that can be transformed into the linear sequence in temporal as a time series data can be analysed by DTW.

Consider we have two time series:
X = (x1, x2, ..., xn) of length n;
Y = (y1 , y2, ..., ym) of length m.

A distance measured by DTW D(x, y) gives the similarity between the time series X and Y. The two series X and Y can be arranged on the sides of a grid, with one on the top and the other on the left-hand side. The sequence of both time series starts on the bottom left of the grid. The grid can be seen as an n by m matrix D, where D(i, j) = d(xi, yj), where d(xi, yj) = $\sqrt{(xi - yj)^2}$ is the Euclidean distance of the two numerical values. (note that the d(xi, yj) can be any distance measurement).

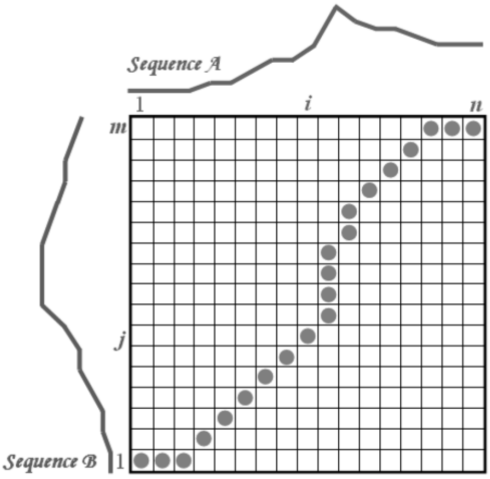


Figure 1. Example of Dynamic Time Warping (DTW) on two time series

A warping path is the path of sequence match of the two time series that may be aligned in time. The best match path of the two sequences can be retrieved by finding an optimal path through the matrix where minimised the cost that cumulated through the warping path [Figure 1]. It can be computed by dynamic programming using the recursion method. The alignment of the optimal warping path D(x,y) is obtained by tracing backwards from D(m, n) with points machining for the two time series. The optimal path generated form DTW has the minimum distance between the two series, it is used as a measure of similarity between the two sequences, and the optimal warping can be used to discover similar pattern the two sequences may possess. (Detail can be found in [42]).

1. The mapping path starts from (1,1) to (i,j); where D(i,j) is the DTW distance between x(1:i) and y(1:j)
2. Recursion:

$$D(i,j) = d(xi - yj\ ) + \min \begin{Bmatrix} D(i-1, & j &) \\ D(i-1, & j-1) \\ D(& i & , j-1) \end{Bmatrix}, \text{ where } D(1,1) = 0$$

3. Optimal path: D(m,n)

Since it is a high complexity in the computation where the algorithm consider all the possible paths through the grid when in a long time series, several well-known condition or constraints have been applied as optimisations to faster the algorithm that lower the complexity by reducing the number of paths that to be considered to reduce the computation complexity during the process of finding the optimal path in the cost matrix. The constraints and conditions include:

- **Monotonic**: the path does not turn back in “time” index, both i and j indexes are either stay at the same or increase on each step of warping and never decrease. The Monotonic constrains guarantees the alignment in the warping path does not repeat [43].

- **Continuity**: the path moves forward one step at a time, both i and j indexes if increase can only one step in the “time” along the path. The condition is to ensure that each point in both the sequences is presented in the warping path [43].

- **Boundary**: Of the grid, the path starts on the left-hand side of the lowest position and ends at the right-hand side of the highest position. That is to make sure to have a complete warping path that does not consider partially matching on the sequences [43].

- **Warping window**: This is a constrains derived from the concept that a good path is not likely to meander far away from the diagonal of the grid. The window width is the distance that the path is allowed to ramble when finding the optimal path. The two most well-known and commonly used constraints are Sakoe-Chuba Band [37] and Itakura Parallelogram [44] shown in Figure 2, [45]. The DTW algorithm finds the optimal wrap path inside the constrained window (indicted as the shaded area in the figure). The constrains method works well when the optimal path is close to the linear wrap as a relative diagonal through the cost matrix. However, it works poorly when the similarity pattern of the time series varies widely in time/ speed which requires DTW to evaluate all cells in the unconstrained cost matrix to obtain the optimal path.

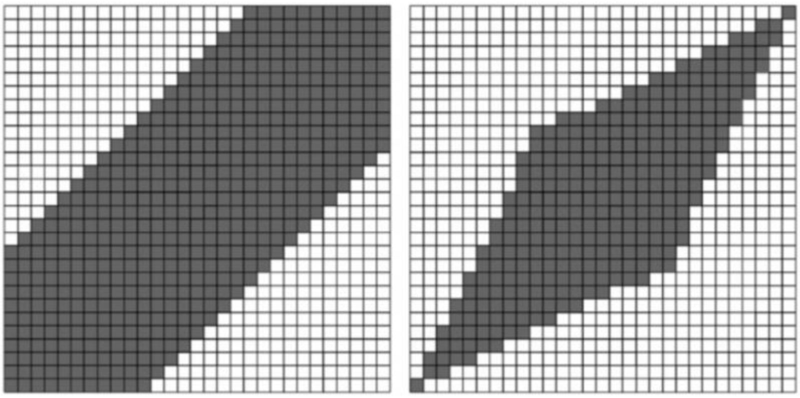


Figure 2. Sakoe-Chuba Band on left, and Itakura Parallelogram on right [45]

3.3.2 DTW on cloud data

A cloud pattern that a location observed can be discovered in its neighbouring area but may vary in time as the change of the cloud speed and direction. This characteristic the cloud process, give us an idea that uses DTW to find the optimal match between a cloud data time series and the neighbouring cloud cover time sequences to obtain the similar pattern that may vary in time and use those patterns as external data to enhance the forecast performance. This approach can be regarded as feature extraction to generate the valuable input data that to improve the forecast accuracy.

We demonstrate the steps of generating data from the DTW warping on the cloud data that represent the similar pattern as input features to the forecast model.

- **Step 1**: Measuring the similarity of cloud data with neighbouring cloud data. We pair each cloud cover data with one of the neighbouring cloud cover data at a time, use DTW to obtain the distance between the two as the similarity of the two time series. In the cloud time series, provide the cloud data is available only on a fixed period which is the measurement 10th to measurement 40th corresponding to the solar output data. Hence, a sequence is period from measurement 10th to 40th in a day.

  The time series is normalised as inputs to DTW. It is stated that the results of DTW would be inadequate without normalisation on the time series [46] and the standardisation on the time series need to on sequence basis. And z-normalising that transform elements to a sequence of values whose mean is approximately 0 while the standard deviation is close to 1, in the most cases has been proved to be the best normalisation method for DTW to generate the best result [47].

- **Step 2**: Neighbouring postcodes that has the cloud cover series that similar to the area of prediction will be used to generate warping paths. For each warping path, only the warping from alignments of the cloud cover sequences that align a point time of predicted area to a point time in past on its neighbouring will be kept (alignments in solid line, green colour as shown in Figure 3). It is because those warping paths show a cloud pattern from a neighbouring district to the predicted area that can be used as historical data for the forecast model [Figure 3].

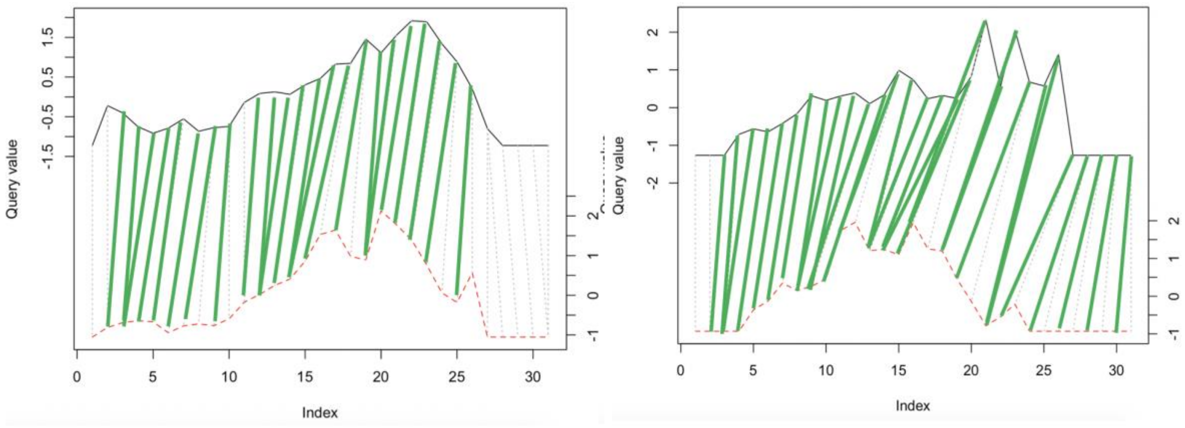
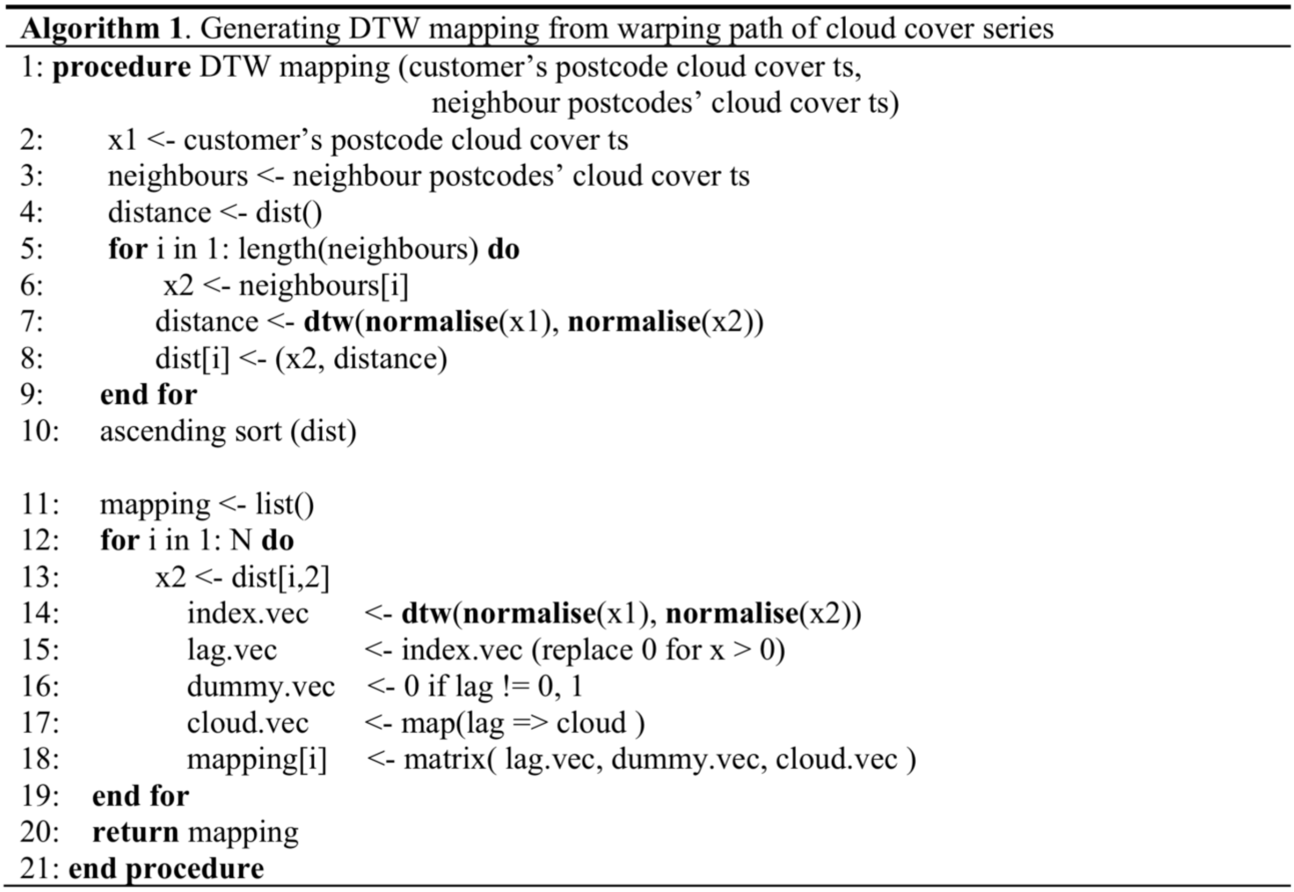


Figure 3. Two examples of DTW on cloud cover data.

- **Step 3**: As the obtained warping path are the time indexes will be used as a predictor with other two predictors us the generated input variables to the forecast model. One of these predictors is the neighbouring cloud cover data that the indexes of the warping path map to. The concept of giving an extract variable apart from the warping path is that warping path can only provide information of the time the similar pattern from on the neighbouring. Hence, this additional variable is to give information of how much the cloud amount was on a cloud pattern was mapped from its neighbour. And another predictor is a dummy variable that is created to take account of the warping path of the cloud is provided or not, which takes value of 1 corresponding to “no” and value of 0 corresponding to “yes”. Combining the three variables provides a cloud pattern from a neighbour with similar cloud pattern that we assume the data can be used as a form of input data that presenting the change of the cloud volume and the speed.

The procedure of generating DTW mapping on cloud data is given in [Algorithm 1].



Algorithm 1. Generating DTW mapping from warping path of cloud cover series

```
procedure DTW mapping (customer’s postcode cloud cover ts,
                       neighbour postcodes’ cloud cover ts)
    x1 <- customer’s postcode cloud cover ts
    neighbours <- neighbour postcodes’ cloud cover ts
    distance <- dist()
    for i in 1: length(neighbours) do
        x2 <- neighbours[i]
        distance <- dtw(normalise(x1), normalise(x2))
        dist[i] <- (x2, distance)
    end for
    ascending sort (dist)

    mapping <- list()
    for i in 1: N do
        x2 <- dist[i,2]
            index.vec     <- dtw(normalise(x1), normalise(x2))
            lag.vec       <- index.vec (replace 0 for x > 0)
            dummy.vec     <- 0 if lag != 0, 1
            cloud.vec     <- map(lag => cloud )
            mapping[i]    <- matrix( lag.vec, dummy.vec, cloud.vec )
    end for
    return mapping
end procedure
```

Providing the forecast is used the DTW warping on the cloud data that to include the matched pattern of the cloud to represent the cloud movement and the cloud cover data are estimated from the solar power output, we choose 2 hours forecast horizon which falls into a category of Intra-day forecasting. The reason of not testing the forecast performance for more than 2 hours is that because of analysis the cloud motions from the neighbouring postcodes and do not study the similarity of the postcodes far away from the predicted area. In other words, to be able to give a longer span forecast horizon, methods of tracking the cloud motion on a large scale in both time and space is needed.

3.4 The overall procedure

In this section, we summarise the process used in model fitting and the forecast with DTW warping on the cloud as inputs. The different numbers of DTW warping are given to the model as input to see the variation of model performance. Besides, we use the same input data with the same parameter setting on both ARIMA model and Random Forest model to compare performance between the linear model (ARIMA) and the non-linear model (Random Forest).

3.4.1 In ARIMA model:

We use 30 days long of the solar data as training data set with the Fourier series as a regressor to fit a model and forecast 2hours ahead. Firstly, we vary K, the number of Fourier sin and cos pairs from K= 1 to K=6 on the ARIMA model to find a K that best represent the daily seasonal pattern of the given solar data. The AICc value is minimised when K=5, with a significant jump observed on which the value of K going from 4 to 5; hence the value of 5 would be the ones used in generating the Fourier series as for the forecast model.

auto.arima() is a function in R. It helps in finding the best fitted ARIMA model by using the AICc information criterion automatically [48]. In the function, the solar data is the input data y and Fourier series as a matrix of external regressors, which must have the same number of rows as y. Both Maximum value of p and Maximum value of q [point to up section] is set to 5. The stepwise is set to True to perform a stepwise selection to find the best model. In the forecast step after the model is fitted, h = 4 is the number of periods ahead to forecast, which give 2 hours forecast time horizon for a half hourly data. In forecast step, Fourier term of the forecast period is given as regressors [Figure 4].

We add DTW mapping as external input data to the model as the proposed approach. ARIMA + 2 DTW model takes two DTW mapping data generated from the most two similar measures [Figure 5], and ARIMA + 4 DTW model takes four DTW mapping data generated from the most four similar measures [Figure 5].

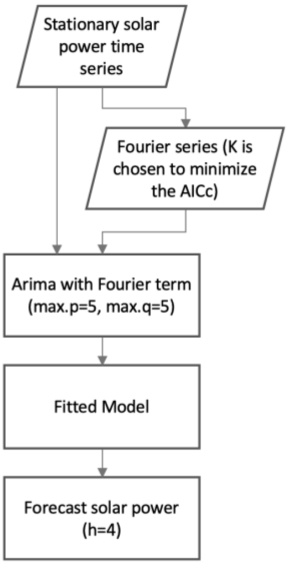

Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
Arima with Fourier term (max.p=5, max.q=5)
Fitted Model
Forecast solar power (h=4)

Figure 4. ARIMA method

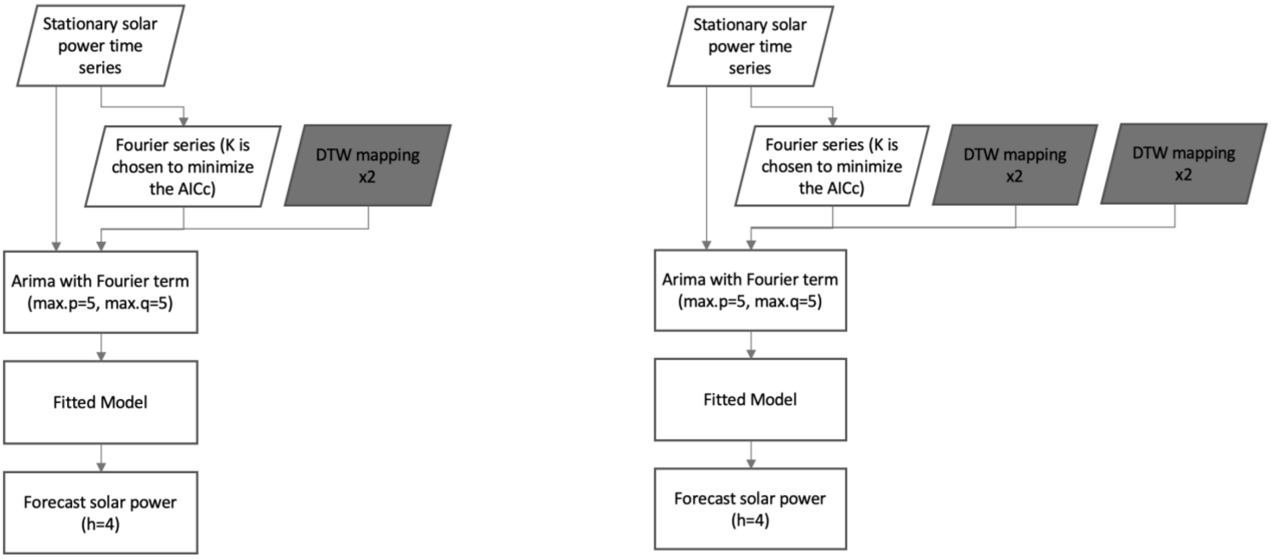

Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
DTW mapping x2
Arima with Fourier term (max.p=5, max.q=5)
Fitted Model
Forecast solar power (h=4)
Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
DTW mapping x2
DTW mapping x2
Arima with Fourier term (max.p=5, max.q=5)
Fitted Model
Forecast solar power (h=4)

Figure 5. Left: ARIMA with 2 DTW. Right: ARIMA with 4 DTW.

3.4.2 In Random Forest model:

The training data set is a matrix that each row of the matrix represents a sequence of past observations from the lag of 5 to lag of 1, combining with the Fourier series for the forecast time step which is 2 hours ahead. In another word, we train a Random Forest model by a 30 days' time series data that transformed a number of vectors that each vector is data of the y_{t-5} to y_{t-1}, where y is the solar data time series with Fourier series f_{t+3} as an input and output of y_{t+3} as predict value. We use the default setting in an R package, [29] for the regression task at which

- ntree = 500
- mtry = the number predictors divided by 3.
- nodesize = 5.

The trained model can then be used to generate forecast of y_{t+3} (2 hours ahead) when a vector of solar data from y_{t-5} to y_{t-1} with Fourier series f_{t+3} is given measures [Figure 6].

And we add DTW mapping as external input data to the model as the proposed approach. The DTW mapping data is added as vectors that each vector containing the DTW warping path from the cloud cover data in a time window of y_{t-5} to y_{t-1} to the original input vector to form new vectors in both training data and testing data. Where Random Forest + 2 DTW takes two DTW mapping data generated from the most two similar measures [Figure 7], and Random Forest + 4 DTW model takes four DTW mapping data generated from the most four similar measures [Figure 7].

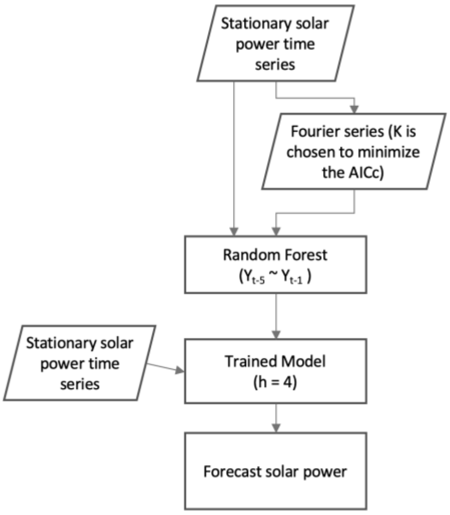

Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
Random Forest (Y_{t-5} ~ Y_{t-1})
Stationary solar power time series
Trained Model (h = 4)
Forecast solar power

Figure 6: Random Forest method

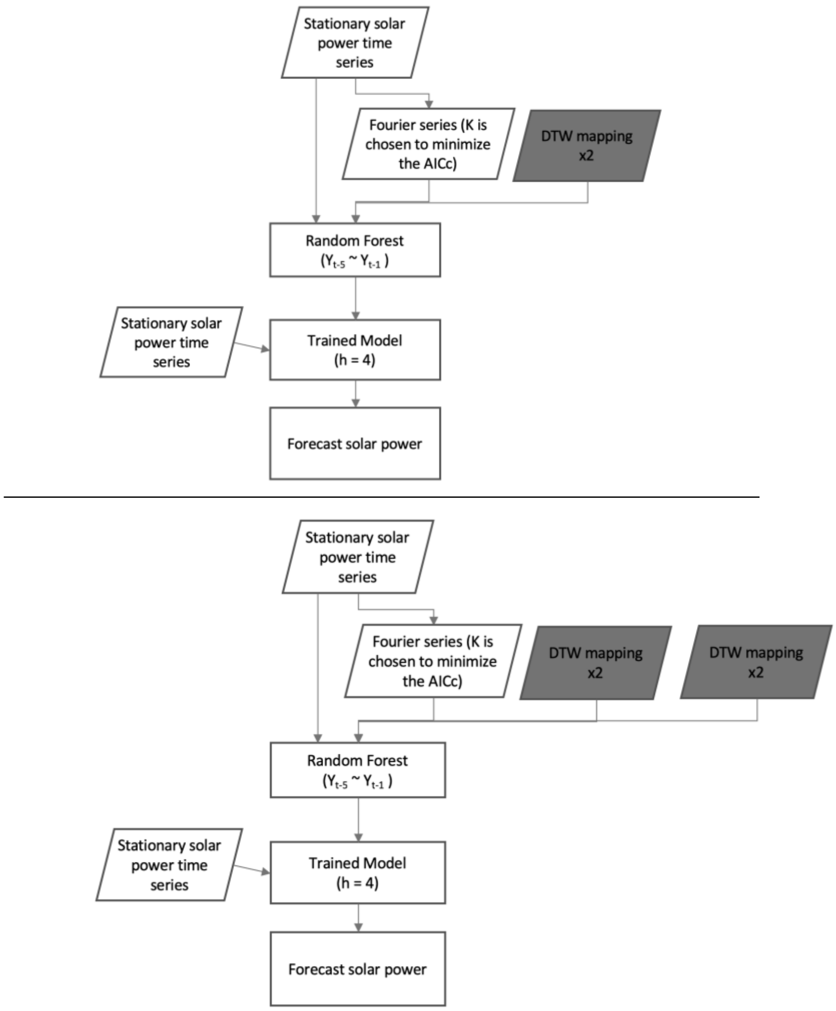

Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
DTW mapping x2
Random Forest (Y_{t-5} ~ Y_{t-1})
Stationary solar power time series
Trained Model (h = 4)
Forecast solar power
Stationary solar power time series
Fourier series (K is chosen to minimize the AICc)
DTW mapping x2
DTW mapping x2
Random Forest (Y_{t-5} ~ Y_{t-1})
Stationary solar power time series
Trained Model (h = 4)
Forecast solar power

Figure 7. Top: Random Forest with 2 DTW. Bottom: Random Forest with 4 DTW.

4.Experimental study

In this section, we evaluate the proposed procedure on a solar dataset collected from distributed PV systems. We describe the two datasets used including solar dataset and the cloud cover dataset that the cloud cover is estimated from the solar data, error metrics used to measure the model performance and the results.

4.1 Datasets

4.1.1 Solar output data:

Ausgrid dataset [49] contains smart meter data that record the energy consumption and the PV generation that from 300 randomly selected residential customers over the period 1 July 2010 to 30 June 2013. The smart meters record the electricity load and solar generation in the gross metering mode for each residential system, in which the measurements of the power flow are recorded in a single direction.

To protect the customers' privacy, the dataset is de-identified with all the personally identifiable information removed. Consequently, the postcode of the customer remained as the only context of geographical spread. There is a total of 100 unique postcodes in the dataset. The time format in the dataset, by default, is Australian Eastern Standard Time (AEST) except for the summer period (around October -April) which the Australian Eastern Daylight Savings Time (AEDT) is used. The load and generation data were recorded at 0.5-hour interval represents the amount sum of the daily half-hour interval, resulting in 48 entries on each day. The PV generation (in kWh), referred to as Gross Generation (GG) over the period 1 Jan 2011 to 31 Dec 2011 in the dataset as the data used in our research project.

4.1.2 Cloud cover data:

Cloud cover data is generated by estimating the PV generation data from the solar data selected from Ausgrid dataset. An understudying physical model that taking the solar generation data with the location provided (in postcode scale) is used[1]. The model output estimated cloud cover data for total 68 unique postcodes in a 0.5 h interval over the year 2011 while removing 32 unique postcodes as the presence of shallow effect on those residential solar panel located under the covered area of the postcodes as it influences the model performance on the estimating the cloud cover. From the overall cloud cover data, we further select 38 unique postcodes covering area of the centres in Sydney and part of regional NSW, AUS as depicted in [Figure 8] with 100 residential customers covered within this area for our implementation.

[1] A currently conducting research under Lachlan Andrew of Monash University (Australia)

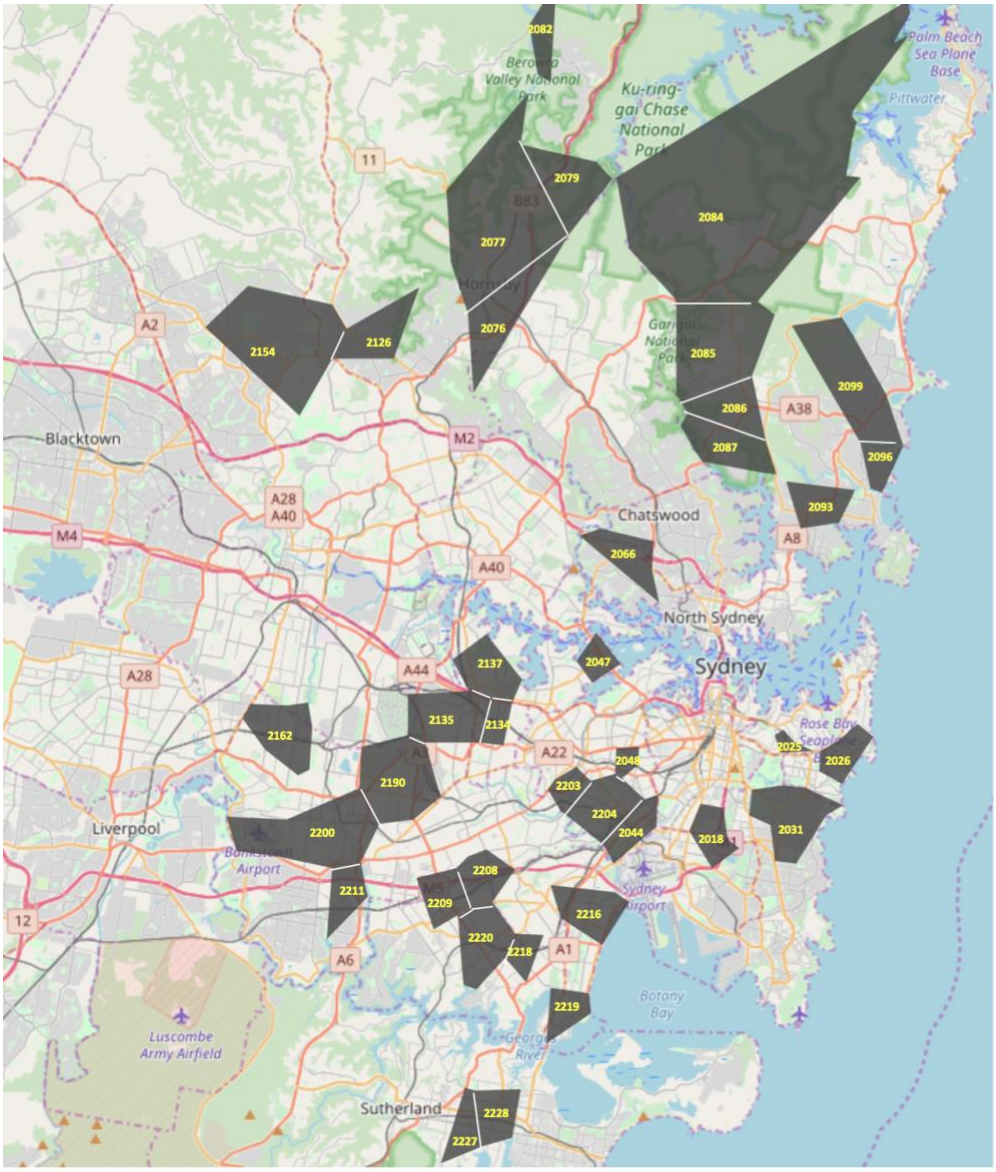


Figure 8. Outline of cloud cover data that covers the centres in Sydney and part of regional NSW, AUS. The shaded regions correspond to the postcode areas with 100 residential customers' PV system covered within. Map data @ Leaflet

4.1.3 Final dataset:

After the prepossessing of the dataset, we finally obtain a dataset consist of solar generation data over the whole year of 2011 from a total number of 100 residential customers and postcodes' cloud cover data that covering the area of those 100 households located.

On testing the model performance of the forecasting, we firstly randomly pick 5 days (5 Mar 2011, 16 Nov 2011, 14 Apr 2011, 12 Oct 2011, 22 July 2011 and 7 Apr 2011) as the first testing data and secondly choose 15 days (7 Apr 2011, 24 Sept 2011, 2 Feb 2011, 17 Oct 2011, 2 Dec 2011, 5 Feb 2011, 6 July 2011, 13 Oct 2011, 19 Nov 2011 and Apr 15 2011 with the original 5 days in first testing data) as the second testing set. And the training set is the last 30 days of each day picked in the testing set. The models will be fitted/trained on the training set and test on the testing set.

4.2 Error measures

To compare the proposed method of including the DTW warping path of the cloud data as the input to the forecast model against the base forecast model that solely rely on the solar data itself in order to see if there is an improvement on the model performance and to examine if the models are better than merely using the baseline method. We choose three error metrics including Root Mean Square Error (RMSE), mean absolute scaled error (MASE) and Symmetric Mean Absolute Percent Error (sMAPE).

4.2.1 Root Mean Square Error (RMSE):

$$= \sqrt{\mathrm{mean}(\mathrm{e}_\mathrm{t}^2)}$$

It is a scale-dependent error measure as the error is on the same scale as the data. When comparing model's performance applied to time series with the same units, RMSE is only of the widely used error metric.

4.2.2 Symmetric Mean Absolute Percent Error (sMAPE):

$$= \mathrm{mean}\ (200\frac{|y_t - \hat{y}_t|}{(y_t + \hat{y}_t)}).\,where$$

Consider of that even the solar data come from the same unit, and each solar system may give a different scale of solar output because of the difference of system capacity. Thus, we also use sMAPE, a Percentage errors measure which is a unit-free measure that is commonly used to measure the performances on forecast between data sets [50].

However, there is a disadvantage of using sMAPE when the data is very close to zero, and the error will be being infinite which make the measure not stable. And what is more, the sMAPE doesn’t work for time series has any data point that sum of the prediction value, and true value is equal to zero as the error metric it as denominator. In the solar generation time series there is time with no output of solar power, for example, the night time on which the sun not presents, make the sMAPE not works. To overcome this, we use sale moving up on the time series data for the time series on both the series and the forecast value by adding a value of 1 before calculating the sMAPE. This scale-up approach can be used as there is no negative value for the solar output time series.

4.2.3 Mean Absolute Error (MASE):

$$= \text{mean}\ (|y_t - \hat{y}_t|)\ /\ (\text{mean}\ (|\text{training error}|)),$$

where the errors are scaled by the training mean absolute error (MAE) of seasonal naïve forecasts. Hence, to make the errors scale independent as the value on both numerator and denominator both are on the scale of the original data. The scaled error is less than the value of 1 when the forecast is better than the average of forecasts the seasonal naïve makes on training data. In contrast, the forecast is worse than the average of forecasts the seasonal naïve computes on training data when the scaled error is greater than the value of 1 [51].

The Scaled errors were proposed by Hyndman and Koehler [51]. It is an alternative method to percentage error method when dealing with comparing forecast prediction performance across the series that the series is in different units.

4.2.4 Measure the method performances:

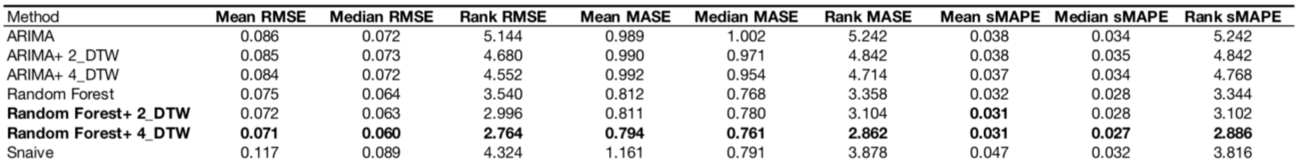
We use the above three different error metrics on every series and obtain the mean of them. And to void outlier of single time series may present in the data set that brings dominated influence on the evaluation, we also compute the median and rank of each error metric across all series. As a result, we provide the following error measures in the comparison of the model performances: Mean of the RMSE (Mean RMSE), Median of the RMSE (Median RMSE), Mean of the RMSE ranks of each series (Rank RMSE), Mean of the MASE (Mean MASE), Median of the MASE (Median MASE), Mean of the MASE ranks of each series (Rank MASE), Mean of the sMAPE (Mean sMAPE), Median of the sMAPE (Median sMAPE), Mean of the sMAPE ranks of each series (Rank sMAPE).

4.3 Results on the dataset

To test if the proposed method is better than the base models, we compare the model's performance on 100 solar system's time series data test the forecast error on 5 randomly select days and 15 randomly selected days in 2011 separately.

4.3.1 Results on 5 days testing:

On the results of 5 days of testing [Table 2], regarding all the error measures, the proposed approach outperforms the benchmark methods. In the ARIMA model, improvement of the model performance is observed when including the DTW warping path. Also, Model with 4 DTW warping paths outperforms than the model with 2 DTW warping paths. Improvement has been discovered in the Random forest model as well. And it has shown that the 4 DTW warping paths perform better and 2 DTW warping paths as well. In particular, the model Random forecast with 4 DTW warping path is the best method of the 7 methods including Arima models with and without DTW warping path and the seasonal naïve method.

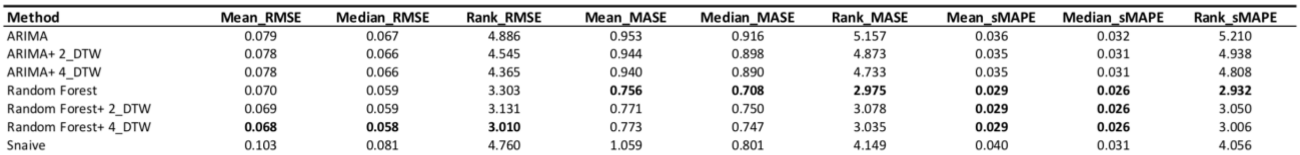


Method	Mean RMSE	Median RMSE	Rank RMSE	Mean MASE	Median MASE	Rank MASE	Mean sMAPE	Median sMAPE	Rank sMAPE
ARIMA	0.086	0.072	5.144	0.989	1.002	5.242	0.038	0.034	5.242
ARIMA+ 2_DTW	0.085	0.073	4.680	0.990	0.971	4.842	0.038	0.035	4.842
ARIMA+ 4_DTW	0.084	0.072	4.552	0.992	0.954	4.714	0.037	0.034	4.768
Random Forest	0.075	0.064	3.540	0.812	0.768	3.358	0.032	0.028	3.344
Random Forest+ 2_DTW	0.072	0.063	2.996	0.811	0.780	3.104	**0.031**	0.028	3.102
Random Forest+ 4_DTW	**0.071**	**0.060**	**2.764**	**0.794**	**0.761**	**2.862**	**0.031**	**0.027**	**2.886**
Snaive	0.117	0.089	4.324	1.161	0.791	3.878	0.047	0.032	3.816

Table 2: Results for the series of solar power generation taken from total 100 household PV systems in NSW, AU over 5 randomly selected days in 2011. For each column, the best performing method(s) are marked in bold.

4.3.2 Results on 15 days testing:

We further test methods performance of forecast on more time series data. On the results of 15 days testing [Table 3], the ARIMA method still shows an improvement on the proposed approach, the ARIMA with 4 DTW has the best performance, followed by ARIMA with 2 DTW and where the ARIMA without DTW come to last. However, in the Random Forest Model, the proposed methods show no improvement and the decreasing of the performance is observed on the percentage error measure sMAPE and scaled independent error measure MASE.



Method	Mean_RMSE	Median_RMSE	Rank_RMSE	Mean_MASE	Median_MASE	Rank_MASE	Mean_sMAPE	Median_sMAPE	Rank_sMAPE
ARIMA	0.079	0.067	4.886	0.953	0.916	5.157	0.036	0.032	5.210
ARIMA+ 2_DTW	0.078	0.066	4.545	0.944	0.898	4.873	0.035	0.031	4.938
ARIMA+ 4_DTW	0.078	0.066	4.365	0.940	0.890	4.733	0.035	0.031	4.808
Random Forest	0.070	0.059	3.303	**0.756**	**0.708**	**2.975**	**0.029**	**0.026**	**2.932**
Random Forest+ 2_DTW	0.069	0.059	3.131	0.771	0.750	3.078	**0.029**	**0.026**	3.050
Random Forest+ 4_DTW	**0.068**	**0.058**	**3.010**	0.773	0.747	3.035	**0.029**	**0.026**	3.006
Snaive	0.103	0.081	4.760	1.059	0.801	4.149	0.040	0.031	4.056

Table 3: Results for the series of solar power generation taken from total 100 household PV systems in NSW, AU over 15 randomly selected days in 2011. For each column, the best performing method(s) are marked in bold.

5. Conclusions

The proposed approach on the chosen dataset, respect to the evaluation metrics used, shows a slight improvement over model with only the solar data as input on both ARIMA model and Random Forest model when testing on 5 randomly selected days, while the increase of the prediction accuracy is not presented on Random Forest model when examining on 15 randomly selected days. The inconsistently chance of the model performance may due to the limitations as follow:

- **The performance of the physical model used for estimating the cloud cover data:** As we use cloud cover data is calculated from the solar generation data instead of the measurement of the real cloud cover data. An underestimated or overestimated cloud cover data may influence the result of a similar pattern that matched from the cloud cover data and led to the change of the forecast.
- **The selection of similar patterns:** In the proposed method, we used similar patterns matched from the nearby postcodes by DTW. The cloud cover data that has the most similar matched to the area of prediction is mostly from the nearest neighbours. Map the two cloud cover time series that are similar to each other in time result in a DTW warping path contain many zero data [Section 3.3.2 Step 2]. A warping from a less similar cloud cover series from other postcodes may hold some similar cloud pattern that varies in time and seed. This cloud pattern provides the cloud information process in a past period that can help the model to learn and detect the correlation of the cloud movement to the solar power generation to increase the forecast performance. However, at the same time, the warping of the cloud pattern may also contain a matched sequence which is not similar to the cloud series of the prediction area. It brings the noise data both on training time and test time that result in a deterioration on the forecast.

The above limitations open new possibilities for optimising the current approach. To have less noise of data as input to the forecast model, further research should focus on establishing a process of refining the warping path to obtain the cloud profile that is truly similar while removing the less similar pattern that produced by the DTW. Besides, other future works may include evaluating the proposed approach with Artificial Neural Networks models (RNNs, LSTM) given the amount of the data is large enough for the deep learning model to get well trained. Last but not least, the proposed method can be tested on other forecasting applications such as the wind power forecast and so on.

References

[1] How do Photovoltaics Work? | Science Mission Directorate [Internet]. Available from: https://science.nasa.gov/science-news/science-at-nasa/2002/solarcells

[2] Solar energy - Australian Renewable Energy Agency [Internet]. Available from: https://arena.gov.au/about/what-is-renewable-energy/solar-energy/

[3] Cros S, Liandrat O, Sébastien N, Schmutz N. Extracting cloud motion vectors from satellite images for solar power forecasting. Geoscience and Remote Sensing Symposium. 2014;.

[4] Quesada-Ruiz S, Chu Y, Tovar-Pescador J, Pedro H, Coimbra C. Cloud-tracking methodology for intra-hour DNI forecasting. Solar Energy. 2014;102:267-275.

[5] Backer J, Wallinga J. Spatiotemporal Analysis of the 2014 Ebola Epidemic in West Africa. PLOS Computational Biology. 2016;12(12):e1005210.

[6] Matějíček L, Engst P, Jaňour Z. A GIS-based approach to spatio-temporal analysis of environmental pollution in urban areas: A case study of Prague's environment extended by LIDAR data. Ecological Modelling. 2006;199(3):261-277.

[7] Yar P, Nasir J. GIS Based Spatial and Temporal Analysis of Crimes, a Case Study of Mardan City, Pakistan. International Journal of Geosciences. 2016;07(03):325-334.

[8] "Simple Factors Affecting Solar Performance worth Knowing | EnergyzedWorld", Energyzedworld.com, 2018. [Online]. Available: http://energyzedworld.com/index.php/2017/06/15/simple-factors-affecting-solar-performance-worth- knowing/

[9] "Factors Affecting Solar Performance", Green Convergence, 2018. [Online]. Available: https://support.greenconvergence.com/customer/en/portal/articles/2233072-factors-affecting-solar- performance.

[10] P. Chrobak, J. Skovajsa and M. Zalesak, "Effect of cloudiness on the production of electricity by photovoltaic panels", MATEC Web of Conferences, vol. 76, p. 02010, 2016.

[11] H. Huang, S. Yoo, D. Yu and D. Huang, "Correlation and Local Feature Based Cloud Motion Estimation", 2012.

[12] C. Chow, S. Belongie and J. Kleissl, "Cloud motion and stability estimation for intra-hour solar forecasting", Solar Energy, vol. 115, pp. 645-655, 2015.

[13] R. Marquez, H. Pedro and C. Coimbra, "Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to ANNs", Solar Energy, vol. 92, pp. 176-188, 2013.

[14] Z. Dong, D. Yang, T. Reindl and W. Walsh, "Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics", Energy Conversion and Management, vol. 79, pp. 66-73, 2014.

[15] A. Dolara, S. Leva and G. Manzolini, "Comparison of different physical models for PV power output prediction", Solar Energy, vol. 119, pp. 83-99, 2015.

[16] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. Martinez-de-Pison and F. Antonanzas-Torres, "Review of photovoltaic power forecasting", Solar Energy, vol. 136, pp. 78-111, 2016.

[17] H. Pedro and C. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs", Solar Energy, vol. 86, no. 7, pp. 2017-2028, 2012.

[18] M. Zamo, O. Mestre, P. Arbogast and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production", Solar Energy, vol. 105, pp. 792-803, 2014.

[19] G. Graditi, S. Ferlito, G. Adinolfi, G. Tina and C. Ventura, "Energy yield estimation of thin-film photovoltaic plants by using physical approach and artificial neural networks", Solar Energy, vol. 130, pp. 232-243, 2016.

[20] D. Yang, P. Jirutitijaroen and W. Walsh, "Hourly solar irradiance time series forecasting using cloud cover index", Solar Energy, vol. 86, no. 12, pp. 3531-3543, 2012.

[21] A. Mellit and A. Pavan, "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, vol. 84, no. 5, pp. 807-821, 2010.

[22] E. G. Kardakos, M. C. Alexiadis, S. I. Vagropoulos, C. K. Simoglou, P. N. Biskas and A. G. Bakirtzis, "Application of time series and artificial neural network models in short-term forecasting of PV power generation", 2013.

[23] A. Gandelli, F. Grimaccia, S. Leva, M. Mussetta and E. Ogliari, "Hybrid model analysis and validation for PV energy production forecasting", 2014.

[24] B. Wolff, J. Kühnert, E. Lorenz, O. Kramer and D. Heinemann, "Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data", Solar Energy, vol. 135, pp. 197-208, 2016.

[25] "Introduction to ARIMA models", People.duke.edu, 2018. [Online]. Available: https://people.duke.edu/~rnau/411arim.htm

[26] "Forecasting: Principles and Practice", Otexts.org, 2018. [Online]. Available: https://otexts.org/fpp2/seasonal-arima.html

[27] S. Insights, "Machine Learning: What it is and why it matters", Sas.com, 2018. [Online]. Available: https://www.sas.com/en_au/insights/analytics/machine-learning.html

[28] K. Nordhausen, "Ensemble Methods: Foundations and Algorithms by Zhi-Hua Zhou", International Statistical Review, vol. 81, no. 3, pp. 470-470, 2013.

[29] Cran.r-project.org, 2018. [Online]. Available: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[30] "Forecasting: Principles and Practice", Otexts.org, 2018. [Online]. Available: https://otexts.org/fpp2/simple-methods.html

[31] I. Iwok, "Seasonal Modelling of Fourier Series with Linear Trend", International Journal of Statistics and Probability, vol. 5, no. 6, p. 65, 2016.

[32] "Forecasting: Principles and Practice", Otexts.org, 2018. [Online]. Available: https://otexts.org/fpp2/dhr.html

[33] "Forecasting with long seasonal periods | Rob J Hyndman", Robjhyndman.com, 2018. [Online]. Available: https://robjhyndman.com/hyndsight/longseasonality/

[34] M. Brewer, A. Butler and S. Cooksley, "The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity", Methods in Ecology and Evolution, vol. 7, no. 6, pp. 679-692, 2016.

[35] Content.pivotal.io, 2018. [Online]. Available: https://content.pivotal.io/blog/forecasting-time-series-data-with-multiple-seasonal-periods

[36] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series", 1994.

[37] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43- 49, 1978.

[38] X. Sun, Y. Miyanaga and B. Sai, "Dynamic Time Warping for Speech Recognition with Training Part to Reduce the Computation", Journal of Signal Processing, vol. 18, no. 2, pp. 89-96, 2014.

[39] L. Livingston, P. Deepika and M. Benisha, "An Inertial Pen with Dynamic Time Warping Recognizer for Handwriting and Gesture Recognition", International Journal of Engineering Trends and Technology, vol. 35, no. 11, pp. 506-510, 2016.

[40] A. Hernández-Vela, M. Bautista, X. Perez-Sala, V. Ponce-López, S. Escalera, X. Baró, O. Pujol and C. Angulo, "Probability-based Dynamic Time Warping and Bag-of-Visual-and-Depth-Words for Human Gesture Recognition in RGB-D", Pattern Recognition Letters, vol. 50, pp. 112-121, 2014.

[41] S. Lee, H. Um and H. Kwon, "Feature-Strengthened Gesture Recognition Model Based on Dynamic Time Warping for Multi-Users", KIPS Transactions on Software and Data Engineering, vol. 5, no. 10, pp. 503-510, 2016.

[42] B. Giao and D. Anh, "Similarity search for numerous patterns over multiple time series streams under dynamic time warping which supports data normalization", Vietnam Journal of Computer Science, vol. 3, no. 3, pp. 181-196, 2016.

[43] "11.2. Dynamic Time Warping", Web.science.mq.edu.au, 2018. [Online]. Available: http://web.science.mq.edu.au/~cassidy/comp449/html/ch11s02.html

[44] F. Itakura, "Minimum prediction residual principle applied to speech recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pp. 67-72, 1975.

[45] S. Salvador and P. Chan, "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", 2007.

[46] O. Henniger and S. Muller, "Effects of Time Normalization on the Accuracy of Dynamic Time Warping", 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems, no. 9825521, 2007.

[47] A. Mueen and E. Keogh, "Extracting Optimal Performance from Dynamic Time Warping", KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2129-2130, 2016.

[48] Cran.r-project.org, 2018. [Online]. Available: https://cran.r-project.org/web/packages/forecast/forecast.pdf

[49] "Solar home electricity data - Ausgrid", Ausgrid.com.au, 2018. [Online]. Available: https://www.ausgrid.com.au/Industry/Innovation-and-research/Data-to-share/Solar-home-electricity-data

[50] "Errors on percentage errors | Rob J Hyndman", Robjhyndman.com, 2018. [Online]. Available: https://robjhyndman.com/hyndsight/smape/

[51] R. Hyndman, "Another Look at Forecast Accuracy Metrics for Intermittent Demand", Foresight: The International Journal of Applied Forecasting, pp. 43-46, 2006.