SEAI RD&D 2019

SMARTeBuses

Validation of Wind Power Prediction Models

Deliverable number: WP3-D2

This Project has received funding from the SEAI RD&D Programme.
Grand Award Number: 19/RDD/519

Project Acronym: SMARTeBuses
Project Full Title: SMART electric Buses
Grant Number: 19/RDD/519
Project URL: https://smartebuses.github.io/web/

Deliverable type:	Deliverable / Report
Dissemination level:	Public
Delivery Date:	April 30, 2021
Number of pages:	18
Keywords:	SMARTeBuses, Time Series, Artificial Intelligence, Deep Learning, ARIMA, SARIMA, LSTM
Authors:	Padraigh Jarvis, University College Cork Dr. Alejandro Arbelaez - University College Cork

Contents

List of Figures

List of Acronyms

SMARTeBuses SMART electric Buses
SEAI Sustainable Energy Authority of Ireland
AI Artificial Intelligence
ARIMA Autoregressive Integrated Moving Average
SARIMA Seasonal Autoregresssive Integrated Moving Average
LSTM Long-Short Term Memory
RMSE Root Mean Squared Error
MSE Mean Squared Error
MAE Mean Absolute Error

1 Introduction

This document corresponds to the delivery WP3-D2 "*Validation of Wind Power Models*"' of the SMART electric Buses (SMARTeBuses) project, funded by the Sustainable Energy Authority of Ireland (SEAI) RD&D Programme. This project is classified as Non-economic public Good Research under the European Union (EU) State Aid regulations and will exploit, combine and improve cutting-edge Artificial Intelligence (AI) technologies to develop and implement optimization models for the operation of electric buses in Ireland with operational constraints.

In this deliverable, we evaluate our previous models for time series prediction [1]. In particular, we focus our attention in the Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregresssive Integrated Moving Average (SARIMA) statistical models and the Long-Short Term Memory (LSTM) deep learning architecture. Furthermore, our experiments suggest that our predictions are slightly better than smart grid predictions.

This deliverable is structured as follows: Chapter 2 describes our evaluation methodology; Chapter 3 presents our experimental evaluations our models for October, November, and December 2020; and Chapter 4 provides some general conclusions.

2 Methodology

As previously outlined both the data for demand (hereby demand actual) and wind generated power (hereby wind actual) from the past seven years was used, for all of the models presented in this section the data was cleaned. This included the removal of any negative values that existed in the dataset and the interpolation of any values that are missing. Table 2.1 shows the metrics for both the demand and the wind generated power dataset. Let Std. Dev., Min., and Max. denote the standard deviation, the minimum, and maximum values of the dataset.

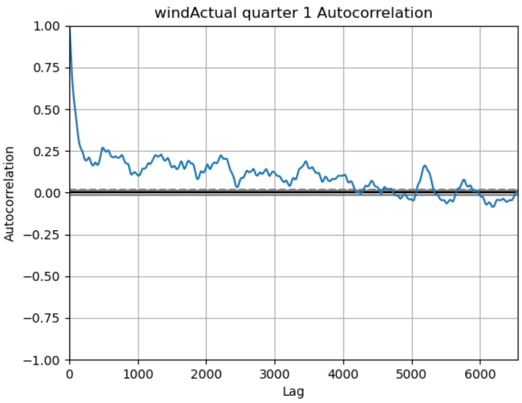
Dataset	Mean	Median	Std. Dev.	Upper quartile	Lower quartile	Max.	Min.
Wind Actual	839.79	679	668.91	1286	274	3337	0
Demand Actual	3141.03	3208	616.75	3595	2635	5033	1664

Table 2.1: Metrics for wind and demand datasets.

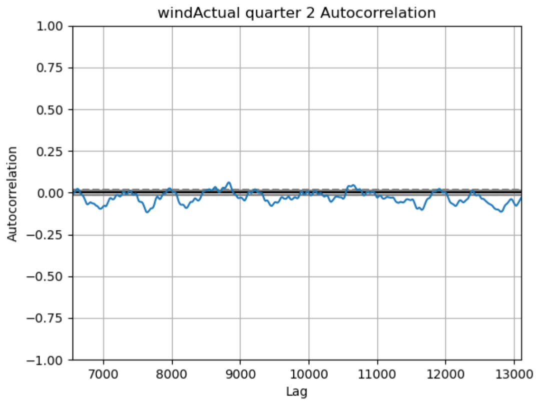
As can be seen from Table 2.1. The data for demand has less deviation than the wind data-set. This is further shown when comparing autocorrelation graphs for both data-sets. The autocorrelation graph shows the correlation between the observations in the time series at the current time step and previous steps [2].

Figures 2.1 and 2.2 show the autocorrelation of both the wind and demand data for 2020. These figures show the similarity between values in the same time-series data-set, where a autocorrelation of 1 shows a positive correlation and a value closer to -1 shows a negative correlation. Figure 2.1 shows no strong positive or negative correlation between values in the data-set. This means it might be difficult to create a predictive model for predicting wind values using traditional time series analysis methods. Figure 2.2 does show stronger correlation, especially in the first two quarters of 2020. From these figures we can estimate that ARIMA models will perform better when predicting grid demand as appose to wind generated power. This is expected, as wind power is weather dependent and demand is dependent on human patterns. This gives some motivation to exploring the effectiveness of a SARIMA model when it comes to predicting wind data, due to the inherent seasonality of the weather.

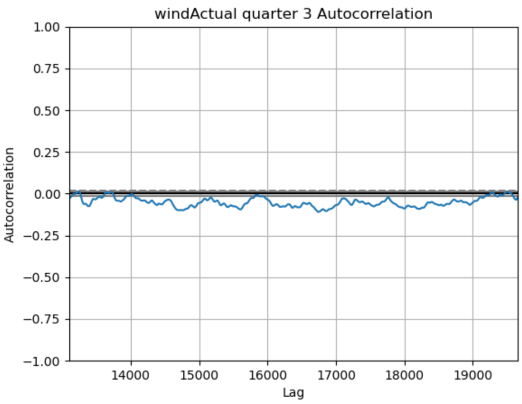
Figure 2.3 shows the partial autocorrelation of the wind and demand data for 2020. This figure is used to determine what values for p in autoregressive models can yield the best result, thus providing a starting point for the experimentation values of the ARIMA and SARIMA models. Figure 2.3 shows that the partial autocorrelation for both data-sets is 0 at a lag of 10. As a result a p value of 10 should be used in ARIMA and SARIMA models.



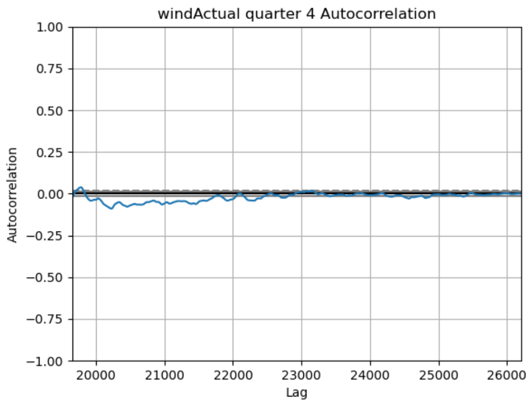
(a) Wind data quarter 1



(b) Wind data quarter 2

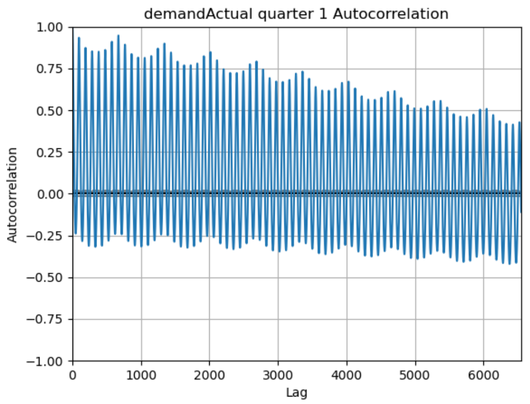


(c) Wind data quarter 3

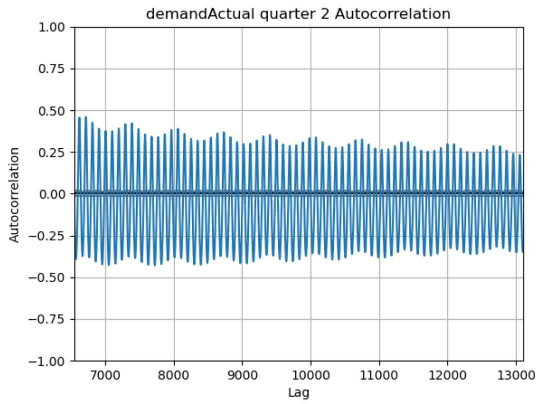


(d) Wind data quarter 4

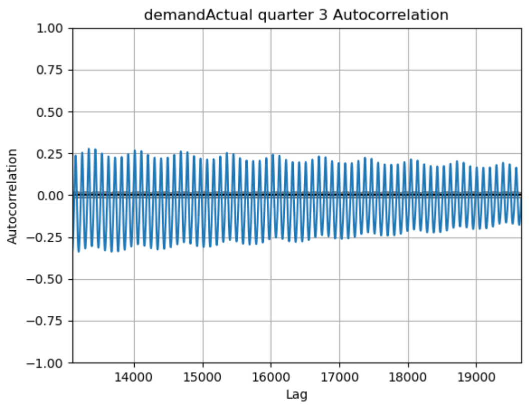
Figure 2.1: Autocorrelation for 2020 wind data



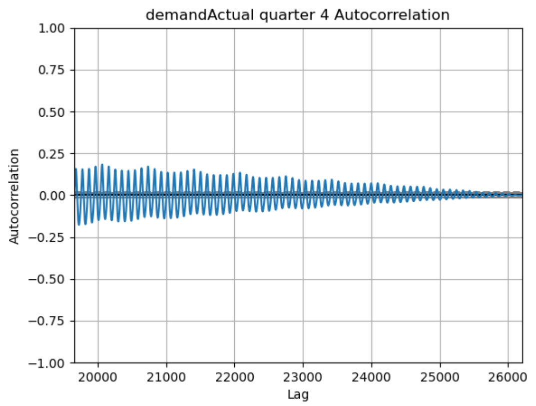
(a) Demand data quarter 1



(b) Demand data quarter 2

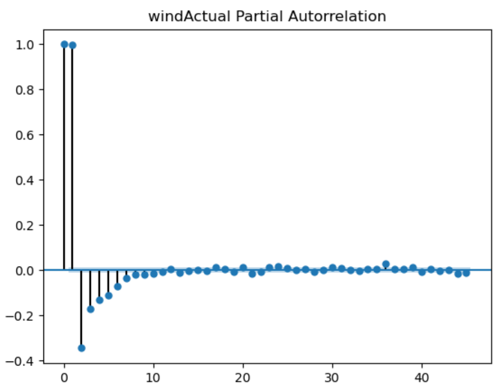


(c) Demand data quarter 3

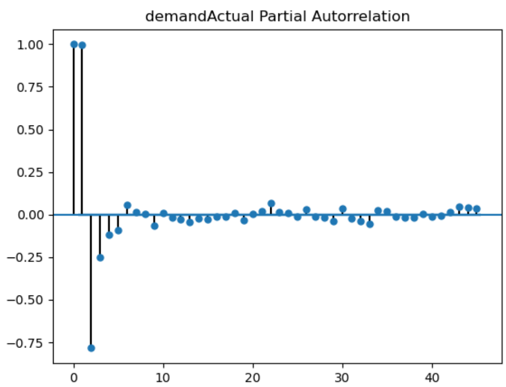


(d) Demand data quarter 4

Figure 2.2: Autocorrelation for 2020 Demand data



(a) Wind data



(b) Demand data

Figure 2.3: Partial Autocorrelation for 2020 data

2.1 Smartgrid baseline

In this project, we collect our reference data from the Smart grid dash board.[1] This dataset provides the actual demand and wind power generation in Ireland with reliable predictions. Additionally, we build our own models to estimate the wind power with flexible time windows.

In particular, we focus our attention in October, November, and December (up to December 7th) 2020 and use various metrics such as Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE) to evaluate the quality of the predictions at a time-step to the actual wind power [3].

Additionally, we would like to remark that the Smart grid predictions are made up to 4.5 days in advance. However every 6 hours the predictions are updated. As a result the predictions made by the baseline are only valid for the next 6 hours. This means for any prediction model to fairly compared to the baseline, the prediction models should only predict the subsequent 6 hours.

2.2 LSTM

As pointed out above in this deliverable we evaluate the performance of our LSTM model and there are a number of different variables which can effect the performance of a LSTM model. These include:

- Learning rate
- Learning rate decay
- Number of epochs
- Batch size
- Number of layers in the LSTM and their neurons
- Number of densely connected layers and their neurons
- The length for input sequence X

Experiments to find the best performing combinations of these values will take place. As mentioned in the previous section, the data is cleaned before use. In addition to this cleaning, the data is made stationary. This is done to remove any time related trends, allowing the data to more easily modeled. The data is then normalized so every value is between -1 and 1.

As outlined in [1] an input sequence X must be provided to the LSTM model. This sequence represents the historic data up to the time-step which is to be predicted. As previously mentioned experiments will be conducted to find an optimal length of historical values for this variable, the range of values tested include 24 hours, and 36 hours. This is broken down into the individual time-steps depending on what time-step aggregation is selected. For example is a time-step aggregation of 1 hour is used then 24 and 36 are used as the values for t. However if 15 minutes time-step aggregation is used then the values for t become 96 and 144 respectively.

In addition to the experiments to find the optimal value for t additional experiments took place to find the best method of prediction for an LSTM model. Two types of predicting future times steps where explored, the first is the recursive approach. In such models the only prediction made is for the subsequent time-step $t+1$. However the value for $t+1$ then becomes element t in the input sequence

[1] http://smartgriddashboard.eirgrid.com/

X. This way a recursive LSTM model can be used to predict as many time-steps in the future without having to be retrained.

The second method is multi-step ahead, here the LSTM model predicts the next n time-steps. These types of models are less flexible as they require retraining if the value for n is changed, however preliminary experiments show promising results when compared to recursive models. For the purpose of comparing the results of the LSTM model to the baseline a n value of $6*(60/\text{time-step aggregation})$ is used for these experiments, (i.e. a time-step aggregation of 15 minutes results in a n value if 24).

2.3 ARIMA and SARIMA

In addition to the experiments to find the most effective time-step aggregation, experiments also took place to find the optimal value for p, d, and q for ARIMA as outlined in [1]. For SARIMA the addition of P, D, Q, and s are included to determine patterns in seasonality. As outlined in 2 the use of partial autocorrelation shows a value of 10 for p should provide optimal results. In addition to this, an additional value of 5 was used to obtain a fair comparison for the impact of variable p on the results produced.

A range for these values are determined based on Figures 2.1, 2.2, and 2.3. These ranges are:

- p: 5, 10
- d: 2
- q: 0, 1
- P: 1
- D: 2
- Q: 0, 1
- s: 24

Much like the data being used in LSTM, the data for ARIMA and SARIMA was cleaned. However the data is not made stationary or normalized. The value for d and D mentioned above is responsible for making any data stationary when using ARIMA and SARIMA.

Much like LSTM, the ARIMA and SARIMA models are used to predict the next 6 hours. After these 6 hours are predicted a new model needs to be create using the new data. After some initial experiments it was found that this is a time consuming process which lead to memory related errors. This was due to the amount of data increasing as the time-step aggregation gets smaller, the size of the prediction model becomes to large to use if using all available data. To work around this the data from 2018-2020 was used for 1 hour time-step aggregation, smaller time-step aggregation could only use data from 2020 to run, this mean the loss of some seasonal data. Because of this the decision was made to not run experiments for time-step aggregation 15, and 30 minutes.

In addition to the memory issues, the time to create was found to be prohibitive for running experiments. To work around this only the first week of October was predicted with ARIMA and SARIMA models.

2.4 Environment

The environment which these experiments took place on was a single computer with an AMD 8-core 3700x processer, this computer also featured 16gb DDR4 RAM running at 3600 Mhz. Experiments for LSTM prediction models made use of a GPU that was installed in the computer, this was a RTX 2070 Super with 8GB of video memory.

3 Evaluation

In this chapter we present a general evaluation of the LSMT deep learning model and the our reference statistical models, i.e., ARIMA and SARIMA.

3.1 LSTM results

While there where a number of experiments which took place to find an optimized LSTM model, one of the important factors was the time-step aggregation. Table 3.1 shows the performance of the two best performing LSTM models for both the demand and wind data-sets with time-step aggregation for 1 hour and 15 minutes. From this table it was found that lower time-step aggregations create more accurate models, this is mainly due to the fact that when time-step aggregation is low the length of the input sequence X is longer.

Time-step aggregation	Data set	MAE	MSE	RMSE
1 hour	Demand	369.62	206431.41	454.34
	Demand	431.01	293954.08	542.17
	Wind	231.98	119059.76	345.05
	Wind	259.47	138165.43	371.70
15 minutes	Demand	**68.51**	**8598.39**	**92.72**
	Demand	69.60	8753.96	93.56
	Wind	**202.72**	**89518.56**	**299.19**
	Wind	229.11	109538.53	330.96

Table 3.1: LSTM 1 hour vs 15 minute time-step aggregation

This would imply that the length of the input sequence X is a important factor on the accuracy of the LSTM model. Further enforcing that fact the experiments where conducted to find how many hours worth of data should be passed to the LSTM model via input sequence X showed a positive correlation between the length of X and accuracy. Table 3.2 shows the accuracy of the best LSTM model which uses 36 hours worth of time-step data compared to the same model but using 24 hours worth of time-step data. Here we see the accuracy of the model increases when a time-steps from the past 36 hours is used as the input sequence. Subsequent experiments using 46 hours failed however, due to issues with the implementation of LSTM used for these experiments.

The results for the best models for both the demand and wind data-set where taken from a range of different LSTM models using a variety of values for the variables outlined in Chapter 2. The values for the variables used in the best performing models outlined in tables 3.1 and 3.2 can be found below. As previously mentioned metrics for LSTM where gathered by predicting the demand and wind data for October, November, and up to December 7th. The predictions for these the first week of December can be seen in Figures 3.1 and 3.2. In these graphs the yellow series represents the actual value for

Input sequence length	Data set	MAE	MSE	RMSE
36 hours	Demand	**68.51**	**8598.39**	**92.72**
	Wind	**202.72**	**89518.56**	**299.19**
24 hours	Demand	70.92	9727.66	98.62
	Wind	205.17	92228.05	303.69

Table 3.2: LSTM 36 vs 24 hour input sequence length

demand and wind values, the blue series represents the predictions made by the LSTM model. The green series in Figure 3.2 represents the predictions made by the baseline model.

- Learning rate: 0.0015
- Learning rate decay: 1e-06
- Number of epochs: 600
- Batch size: 512
- LSTM layers: 1
- LSTM layer neurons: 50
- Densely connected layers: 0
- Length of X: 144 time-steps

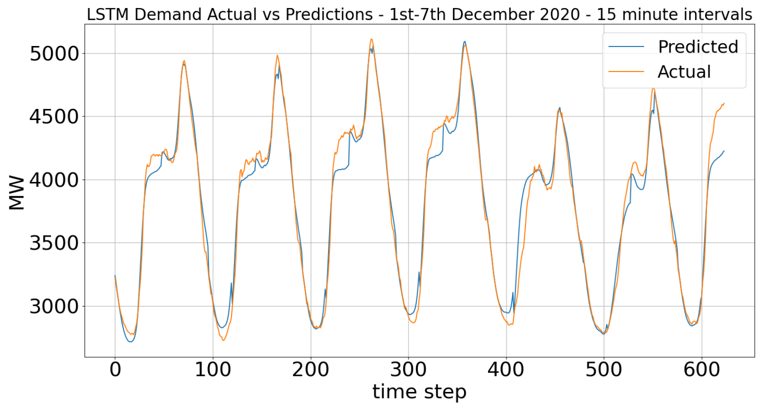


Figure 3.1: Demand for the first week of December.

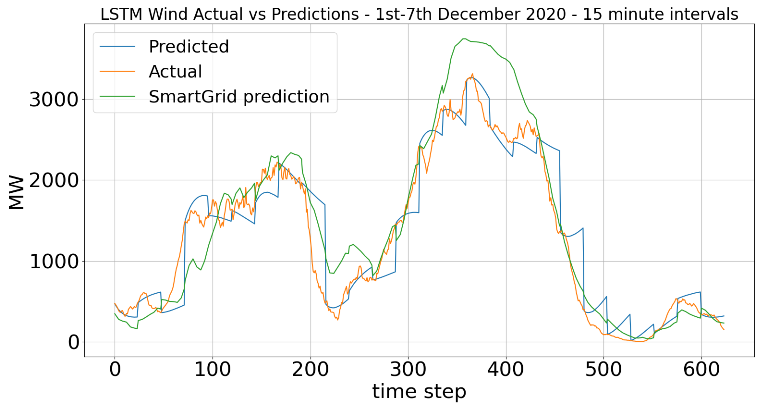


Figure 3.2: Wind prediction for the first week of December.

3.2 ARIMA and SARIMA results

During the experiments to create ARIMA and SARIMA models there where a number of difficulties encountered. Many of these are already outlined in 2, one additional obstacle was that some combination of values for pdq and $pdqPDQM$ resulted in crashes. As a result a number of experiments to find values for the aforementioned variables had to be skipped. In tables 3.3 and 3.4 the results of the ARIMA and SARIMA experiments can be seen. Noticeably there missing experiments in table 3.4, this was due to previously mentioned issue that occurred during the execution of the experiments. However from these results a number of conclusions can be drawn on the performance of ARIMA vs SARIMA.

These tables show the performance of each model for both of the data-sets used in our experiments. These include the variable settings used and the MAE, MSE, and RMSE of these models. Here we can see that SARIMA out performs ARIMA in both data-sets. This was expected due to the seasonality of the data-sets, this is especially prevalent in the results for power demand. The patterns for this data-set are well established, as a result there is a large jump in accuracy of 93% compared to wind power generation where there is a accuracy increase of 72% between the RMSE of the best performing model for both models.

The predictions of the best performing SARIMA model for both data-sets can be seen in Figures 3.3 and 3.4. These figures showcase the predictions for the first week of October compared to the actual values, the yellow series represents the actual value at of ether system demand or wind power generation. The blue series represents the prediction value from the SARIMA model, in Figure 3.4 a green series is also present, this represents the predictions of the baseline model. From these Figures it can be observed that, while SARIMA outperforms ARIMA and obtains good results for system demand, the model's prediction for wind generated power is not as accurate as the baselines predictions.

Data set	pdq	MAE	MSE	RMSE
Demand	10,1,0	3422.16	12052278.47	3471.63
	10,1,1	**3416.18**	**12027856.51**	**3468.12**
	10,2,0	3421.87	12055361.32	3472.08
	10,2,1	3421.43	12086323.92	3476.53
	5,1,0	3423.68	12048110.01	3471.03
	5,1,1	3423.92	12083696.05	3476.16
	5,2,0	3424.17	12063240.42	3473.21
	5,2,1	3421.69	12080833.24	3475.75
Wind	10,1,0	1368.23	2455858.02	1567.11
	10,1,1	1371.54	2491356.45	1578.40
	10,2,0	1384.57	2497093.36	1580.21
	10,2,1	1386.37	2513899.39	1585.52
	5,1,0	**1369.70**	**2446322.48**	**1564.07**
	5,1,1	1379.32	2511144.82	1584.65
	5,2,0	1383.16	2495083.67	1579.58
	5,2,1	1386.51	2512435.96	1585.06

Table 3.3: Metrics for ARIMA models for both data sets

Data set	pdqPDQM	MAE	MSE	RMSE
Demand	10,2,0,1,2,0,24	217.63	84265.18	290.28
	10,2,0,1,2,1,24	146.80	39546.22	198.86
	10,2,1,1,2,0,24	151.76	47450.98	217.83
	5,2,0,1,2,0,24	191.80	68141.65	261.03
	5,2,0,1,2,1,24	133.74	37278.28	193.07
	5,2,1,1,2,0,24	158.58	50536.87	224.80
	5,2,1,1,2,1,24	**105.77**	**24873.98**	**157.71**
Wind	10,2,0,1,2,0,24	444.89	351057.52	592.50
	10,2,0,1,2,1,24	**334.46**	**199732.71**	**446.91**
	10,2,1,1,2,0,24	494.07	570872.16	755.56
	5,2,0,1,2,0,24	438.48	377382.11	614.31
	5,2,0,1,2,1,24	336.23	212486.57	460.96
	5,2,1,1,2,0,24	486.38	571058.29	755.68

Table 3.4: Metrics for SARIMA models for both data sets

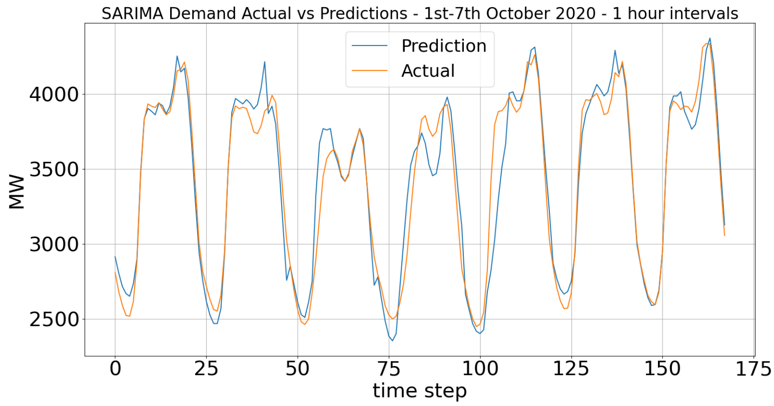


Figure 3.3: SARIMA demand prediction for the first week of October.

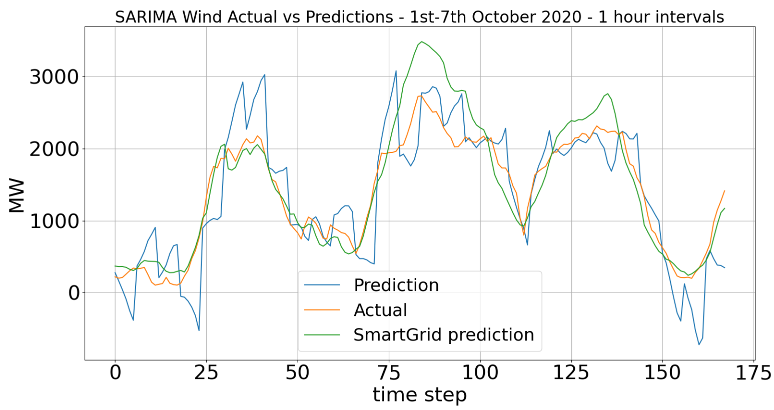


Figure 3.4: SARIMA wind prediction for the first week of October.

3.3 Comparison

Tables 3.5 and 3.6 showcase the metrics of the best performing models from all three types of categories. Here we can see that the LSTM predictions are notable more accurate compared to ARIMA, SARIMA, and the baseline. When directly comparing the RMSE of LSTM to the next best performing model there is a 20.6% increase in prediction accuracy for wind data and 24.4% increase for the demand data-set.

It should be noted that the SARIMA model uses 1 hour time-step aggregation and the baseline and LSTM models uses 15 minute time-step aggregation. When comparing the 1 hour time-step aggregation for the LSTM model found in table 3.1 with the best SARIMA results its observed that SARIMA does perform better. However the training time should be taken into consideration for SARIMA and ARIMA models.

LSTM models only need to be trained once, to predict the demand or wind generated power for a six hour period only the input sequence X needs to be provided. With the sizes of X explored in these experiments this process was instantaneous, however for SARIMA and ARIMA the model must be retrained for every six hour prediction. In the case of the best performing SARIMA model for predicting wind data the time to train a new model takes 19 minutes on average. This is a costly investment with 76 minutes of computation time a day being required to make new predictions.

Model	MAE	MSE	RMSE
Baseline	260.66	141992.69	376.82
ARIMA	1377.11	2478503.38	1404.79
SARIMA	334.46	199732.72	399.57
LSTM	**202.72**	**89518.57**	**299.2**

Table 3.5: Metrics of the best performing models for predicting wind data

Model	MAE	MSE	RMSE
ARIMA	3416.18	12027856.52	3430.36
SARIMA	105.77	24873.99	122.05
LSTM	**68.52**	**8598.39**	**92.73**

Table 3.6: Metrics of the best performing models for predicting demand data

4 Conclusions

In this deliverable we have evaluated the performance of our models to estimate the electricity demand and amount production of clean energy in Ireland. In particular, we focussed our attention in two statistical models, i.e., ARIMA and SARIMA and the LSTM deep learning architecture. Our empirical evaluation suggest that our deep learning model is consistently better than the statistical models and the Smart grid predictions. Furthermore, unlike the Smart grid predictions that only provide estimations up 6 hours ahead, our models are capable and flexible to provide predictions up to any predefined time window. In this context, we provide estimations up to 1 hour ahead for the electricity demand in Ireland.

We plan to use the outcome of this deliverable in WP3-D2 *“Online Model to Detect Abnormal Wind Power Supply”* and WP5-D3 *“Evaluation of the Scheduling Algorithms”* and WP5-D4 *“Guide lines for the fleet-size”*.

Bibliography

[1] P. Jarvis and A. Arbelaez, "Deliverable WP3-D1: Forecasting models for power prediction," Jan. 2021, tech. Rep. WP3-D1 SMARTeBuses project.

[2] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. Wiley, 2008.

[3] F. M. Dekking, C. Kraaikamp, H. P. Lopuhaä, and L. E. Meester, *Efficiency and mean squared error*. London: Springer London, 2005, pp. 299–311.

SMARTeBuses

SMART electric Buses

April 30, 2021