Saf Flatters

Data Scientist

Simulating Wind Speed for Wind Turbine Feasibility

A Computer Simulation Study

**Contents**

1. Introduction
2. Literature Review
3. Code
4. Data & Preprocessing
5. Distribution Modelling Methodology
   1. Selecting Distribution Model
   2. Distribution Modelling Results
6. SARIMA Forecasting Methodology
   1. Time Series Data & Stationarity
   2. SARIMA Model Selection
   3. Forecast Results & Interactive Forecast Plot
7. Dealing with Complex Seasonality
8. Limitations
9. Conclusion and Future Work
10. **Introduction**

Australia is at a pivotal stage in transitioning from fossil fuel-based electricity generation to renewable energy sources. Wind power is the fastest-growing renewable energy source in the country, with large-scale wind farms in various stages of development nationwide (Li et al., 2020). Since wind generation has a cubic relationship with wind speed, identifying suitable locations for wind turbines requires simulating wind speeds (Yatiyana et al., 2017).

This project aimed to simulate wind speeds in Ballarat, Victoria for use in feasibility studies where wind conditions are suitable for turbine operation. The observational data was sourced from the Ballarat Rowing Club, freely accessible through the City of Ballarat website.

Using R, this project involved two main components:

* Distribution Modelling - A Weibull distribution was fitted to the annual wind speed data using Maximum Likelihood Estimation (MLE). This was compared against a composite model in which separate Weibull distributions were fitted for each season (Autumn, Winter, Spring, and Summer) to assess whether seasonal modelling improved the simulation’s accuracy.
* Time Series Forecasting - Several SARIMA models were evaluated to forecast wind speeds. Model selection considered both Akaike Information Criterion values and residual diagnostics, with a focus on minimising AIC, a metric that estimates the relative amount of information loss to ensure reliable forecasts.

1. **Literature Review**

Wind speed simulations commonly employ statistical and time-series techniques widely used in wind energy research. According to Feijóo & Villanueva (2016), the Weibull distribution is one of the most commonly used distributions for modelling wind speeds. The probability density function of a Weibull random variable is:

where; is the wind speed, is the shape parameter and is the scale parameter.

To forecast future wind speeds, researchers use Seasonal ARIMA (AutoRegressive Integrated Moving Average). Karatepe & Corscadden (2013) note that SARIMA models are commonly applied for wind speed forecasting due to its ability to model both trend and seasonal components in time-series data. The ARIMA component uses Autoregression (AR) to incorporate past wind speeds, Differencing (I) to remove trends and achieve stationarity and Moving Averages (MA) to account for past error terms (Hyndman & Athanasopoulos, 2018). The seasonal extension of ARIMA enables the model to capture periodic fluctuations in wind speeds (Tyass et al., 2022). The equation for SARIMA is:

where; is the wind speed at time , is the backshift operator (, (and D for seasonal) is the number of differences required to achieve stationarity , (and is the autoregressive polynomial of order (representing the dependence of current wind speeds on previous values) and (or is the moving average polynomial of order (representing the influence of past errors on current values) and is the white noise error term, meaning it has no autocorrelation, constant variance, and a mean of zero.

This project used SARIMA to forecast wind speeds by comparing candidate models using the Akaike Information Criterion (AIC) and examining residuals for autocorrelation. The goal was to identify a model with low AIC and residuals that resembled white noise—ensuring that the underlying assumptions of the model were satisfied.

1. **Code**

To conduct this study, I used R and R Studio to implement the simulation and time series forecasting methods described below in the methodology. Code file is ‘Code\_STAT2005Project\_21827361.rmd’ and data is ‘wind-observations\_ballarat.csv’. The most important plot outputs are copied into this report. The forecasting process and model diagnostics were heavily guided by *Forecasting: Principles and Practice* by Hyndman and Athanasopoulos (2018), a resource that was especially instrumental in selecting and evaluating SARIMA models throughout the project.

1. **Data & Preprocessing**

The data was obtained from Ballarat Rowing Club. Plotted as a time-series, it had large gaps so I filtered it to where there were no missing values (April 2023 to April 2024). It was also plotted to determine how consistently the measurements were recorded over time. The wind speed measurements were between 1 – 5 observations an hour so observations were averaged by the hour using the floor of each timestamp to create regular, evenly spaced time- series – suitable for modelling. A 7-day moving average plot was generated to explore potential seasonality and trends; however, no clear patterns were identified.

1. **Distribution Modelling Methodology**
   1. **Selecting Distribution Model**

As discussed, the Weibull Distribution was used to model observed hourly wind speed using MLE. The weibullness wp.test indicated the data followed the Weibull distribution (correlation = 0.9973, p-value = 0.1133). A visual assessment of the data’s density plot (blue) with the theoretical Weibull PDF (red) (*Figure 1*) confirmed this. The Kolomogorov-Smirnov (K-S) test was initially proposed to test goodness of fit, however it was unsuitable due to the presence of ties which violates the tests’ assumptions. After exploring alternative distributions – the density and Q-Q plots of the Weibull was used to demonstrate a good fit. The model however, underpredicts high winds suggesting rare extreme weather events may be overlooked (as seen in the upper quartiles in the Q-Q plot (*Figure 2*).

A graph of a wind speed

AI-generated content may be incorrect. A graph with a line

AI-generated content may be incorrect.

*Figure 2*

*Figure 1*

To investigate seasonal variation, the data was segmented into four meteorological seasons and fitted separate Weibull distributions for each (*Figure 3).* Seasonal Q–Q plots (*Figure 4)* revealed similar underestimation of high wind speeds across all seasons except winter. Interestingly even though the seasonal models underestimated the upper-tail, when combine to use for simulation, it over estimated the upper tail. This may be due to added variation from sampling across multiple distributions.

A graph of different seasons

AI-generated content may be incorrect. A graph of different colored lines

AI-generated content may be incorrect.

*Figure 4*

*Figure 3*

* 1. **Distribution Modelling Results**

The seasonal Weibull model fit achieved a higher total log-likelihood (−19015.87) and a lower AIC (38047.74) compared to the annual model fit (log-likelihood = −19215.11, AIC = 38434.22), indicating that modelling wind speeds separately by season provides a significantly better fit to the data despite the increased complexity. These simulated values closely matched the overall shape of the distribution when plotted (*Figure 5)*. Further improvement could include moving seasonal boundaries to refine the fit.

A graph of a graph showing a graph of a wind speed distribution

AI-generated content may be incorrect.The final fit is:

*Figure 5*

1. **SARIMA Forecasting Methodology**
   1. **Time Series Data & Stationarity**

A graph of a wind speed

AI-generated content may be incorrect.To convert the data into a time series object for SARIMA modelling, a frequency of 24 was specified, which assumes an underlying daily cycle in the hourly wind speed data. To validate this assumption, an Autocorrelation Function (ACF) analysis was performed. The ACF plot (*Figure 6)* showed strong, regularly spaced spikes at lags 24, 48, 72, and beyond, confirming the presence of daily cyclic behaviour and supporting the chosen frequency. This confirmed the presence of seasonality and justified skipping using a non-seasonal ARIMA model for a Seasonal ARIMA model for forecasting.

*Figure 6*

The time-series was tested for Stationary using the Augmented Dickey-Fuller Test which rejected the null hypothesis (Dickey-Fuller = -13.2913, p-value = 0.01) indicating that the time series is stationary and does not require differencing (subtracting the previous observation from the current one (Yadav, 2021). It is important to note that, when assessing the core assumptions of ARIMA, tests were conducted to determine whether differencing the time series manually would affect the outcome. The results showed no impact as SARIMA incorporates differencing within its modelling process.

* 1. **SARIMA Model Selection**

The aim of the SARIMA modelling process was to select a model that minimised AIC while also producing residuals that were as close to white noise as possible (independently and identically distributed with no significant autocorrelation) (Bandyopadhyay, 2020). In R, the model outputs as ARIMA(p, d, q)(P, D, Q)[m] where lowercase p, d, q refers to the non-seasonal parameters and uppercase to the seasonal parameters. [m] denotes the seasonal period.

First, I used auto.arima() with default settings, which utilises stepwise selection and approximation and it resulted in ARIMA(0,1,1)(0,0,2)[24], incorporated one level of non-seasonal differencing and two seasonal moving average terms based on a daily cycle. This model returned an AIC of 22403.47. However, the Ljung–Box test on the residuals (Q\* = 209.24, p-value < 2.2e-16) strongly rejected the null hypothesis of independence, indicating significant autocorrelation remained.

A graph of a graph of a graph

AI-generated content may be incorrect.Next, I disabled both stepwise and approximation to perform a more exhaustive grid search for an optimal model. This returned ARIMA(2,1,0)(2,0,0)[24] with a slightly improved AIC of 22395.72. Despite this, residual autocorrelation was still present.

Acknowledging automatic methods were insufficient, I proceeded with manual experimentation. I used ACF and PACF plots to determine starter combinations. I tested multiple configurations with the best results being ARIMA(2,1,2)(1,0,1)[24] and ARIMA(2,1,5)(1,1,2)[24]. The first model yielded the lowest AIC (21978.78), while the second showed the best residual independence (Ljung–Box p = 0.006582), though still not fully satisfying the white noise assumption as seen in the residual ACF plot (*Figure 7)*.

*Figure 7*

I selected ARIMA(2,1,2)(1,0,1)[24] for forecasting due to its lower AIC despite failing Ljung-Box, following Hyndman (2018), who notes that “residual tests are often imperfect and that practical fit often outweighs strict assumptions”. Final SARIMA model (using the coefficients from the R output):

The model includes two non-seasonal autoregressive terms, indicating dependence on the previous two time points, and one non-seasonal differencing operation to achieve stationarity. It also includes two non-seasonal which models the influence of recent forecast errors on current values. Additionally, the model incorporates a seasonal autoregressive term and one seasonal moving average term at lag 24 to capture daily cyclic behaviour, but no seasonal differencing is applied.

* 1. **Forecast Results & Interactive Forecast Plot**

To forecast the next 48 hours of wind speed, as per on of the project aims, I used the forecast() function with the selected SARIMA model. The resulting plot (*Figure 8)* displays the 80% prediction interval in dark blue and the 95% interval in light blue, with the observed data shown in teal and the forecast mean in tomato red. Since prediction intervals extended below zero, I modified the code to prevent negative values from appearing, which are not physically meaningful for wind speed. I focused on the final 10 days (240 hourly observations) of the dataset and re-generated the 48-hour forecast based on that window for plotting. This was visualised using ggplotly, enabling users to interact with the plot by hovering to view exact values, zooming in and out, and dynamically comparing different time segments (Yu, 2020).

A graph with blue and red lines

AI-generated content may be incorrect.

*Figure 8*

1. **Dealing with Complex Seasonality**

Although short-term forecasts appear reasonable, the residuals indicate significant structure that remains unexplained, suggesting the current model fails to fully capture the intricacies of the wind speed data.

To address this, I experimented with incorporating seasonal effects explicitly. A set of seasonal dummy variables were created representing Summer, Autumn, Winter, and Spring and explored two approaches:

* Separate Meteorological Season-specific SARIMA models – see code file
* ARIMAX incorporating seasons as a covariate (Additional Marks) - I conducted further research and experimented with the ARIMAX model guided by Hyndman (2014). This model can incorporate seasonal dummies (summer being the reference to avoid multicollinearity) as a covariate (known as exogenous regressors). The fitted ARIMAX model found with auto.arima was ARIMA(2,0,1)(2,0,0)[24]. The final ARIMAX model (in regression format) is:

In this equation; is wind speed at time and are wind speeds at the two previous time points (short-term autocorrelation), and are seasonal lags (at 24 and 48 hours), is the random error (white noise) term and the seasonal dummy coefficients measure the influence of each season relative to Summer.

This model captured both temporal autocorrelation and seasonal influences on wind speed, however the AIC=21995.21 showed no improvement on the simpler SARIMA model.

While these could offer incremental improvements with refinement, the residual diagnostics suggested more sophisticated modelling would be required (see Future Work section).

1. **Limitations**

There were many limitations discovered in this study:

* Tied values in continuous data (due to rounding) meant the K-S test could not be used for Goodness of Fit
* Parameter estimation bias may have occurred due to ties and rounding in the data as well (D’Agostino & Stephens, 1986).
* ARIMA model residuals showed autocorrelation, suggesting the presence of complex seasonality and non-linear patterns not captured by the model. ARIMA can not capture multiple seasonalities within a single model.
* Reliance on auto.arima for initial model selection was a limitation as manually selected models performed better in some cases. Manual selection of SARIMA configurations was time consuming and subjective (looking at both AIC and Ljung-Box test).

1. **Conclusion & Future Work**

This project successfully simulated and forecasted wind speeds in Ballarat using a combination of statistical distribution modelling and time series forecasting in R. By fitting a Weibull distribution using Maximum Likelihood Estimation, I was able to model wind speed behaviour, with results indicating that seasonal modelling provided a significantly better fit than a single annual distribution. Time series forecasting using SARIMA models allowed for short-term wind speed predictions, with the chosen model ARIMA(2,1,2)(1,0,1)[24] selected for its low AIC despite remaining autocorrelation in the residuals. While both approaches demonstrated reasonable accuracy, limitations related to tied data and unexplained residual patterns suggest that further refinement is possible.

Future work could explore more advanced methods such as multiple seasonal decomposition, harmonic regression, or neural networks to better capture the complexities of wind behaviour and improve forecasting reliability.

**References**

Bandyopadhyay, P. (2020, May 19). *SARIMA models in R*. Medium. <https://medium.com/biased-algorithms/sarima-models-in-r-6cbfcfe57f09>

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques*. CRC Press.

Feijóo, A., & Villanueva, D. (2016). Assessing wind speed simulation methods. *Renewable and Sustainable Energy Reviews*, *56*, 473–483. <https://doi.org/10.1016/j.rser.2015.11.094>

‌Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2

Hyndman, R. J. (2016, July 18). *Ljung-Box test and degrees of freedom*. Hyndsight. <https://robjhyndman.com/hyndsight/ljung_box_df.html>

Hyndman, R. J. (2014, August 19). *ARIMAX models*. Hyndsight.

Li, H. X., Edwards, D. J., Hosseini, M. R., & Costin, G. P. (2020). A review on renewable energy transition in Australia: An updated depiction. *Journal of Cleaner Production*, *242*(118475), 118475. <https://doi.org/10.1016/j.jclepro.2019.118475>

Karatepe, S., & Corscadden, K. W. (2013). Wind Speed Estimation: Incorporating Seasonal Data Using Markov Chain Models. *ISRN Renewable Energy*, *2013*, 1–9. <https://doi.org/10.1155/2013/657437>

Tyass, I., Bellat, A., Raihani, A., Mansouri, K., & Khalili, T. (2022). Wind Speed Prediction Based on Seasonal ARIMA model. *E3S Web of Conferences*, *336*, 00034. <https://doi.org/10.1051/e3sconf/202233600034>

Yadav, A. (2021, January 21). *Time series analysis for dummies*. Medium. <https://medium.com/@amit25173/time-series-analysis-for-dummies-821a9cb4b779>

Yatiyana, E., Rajakaruna, S., & Ghosh, A. (2017). *Wind speed and direction forecasting for wind power generation using ARIMA model*. <https://doi.org/10.1109/aupec.2017.8282494>

Yu, S. (2020, June 10). *Time series in 5 minutes, part 1: Visualization with the time plot*. R-bloggers. <https://www.r-bloggers.com/2020/06/time-series-in-5-minutes-part-1-visualization-with-the-time-plot/>