Predicting Wind Energy Output with Machine Learning Methods for Wind Farms in Texas

Data Science for Energy Transition, Cohort 3

Santiago Gonzalez Ramos*, Ramesh Niraula*, and Hatice Yazici**

*Sam Houston State University, **University of Houston

Abstract

This research investigates the efficacy of machine learning approaches for predicting wind speed across Texas using ERA5 reanalysis data. The study compares different forecasting methods including linear regression, first-order autoregressive models (AR1), and more advanced techniques like random forest with temporal lag features and neural networks. Using historical data from 2018-2020, we develop predictive models for 2021 wind patterns. Our findings demonstrate that incorporating lag features significantly improves prediction accuracy, with random forest models outperforming traditional approaches. The results have important implications for renewable energy planning and integration in Texas's power grid.

Background

Wind energy constitutes a significant portion of Texas's renewable energy portfolio. Accurate forecasting of wind speeds is essential for effective grid management and energy planning. The ERA5 dataset, part of the PLUSWIND data repository, provides high-quality reanalysis data that can be leveraged for wind speed prediction.

Key challenges in wind forecasting include:

- High temporal and spatial variability
- Complex terrain influences
- Integration of multiple meteorological factors
- Need for site-specific calibration

This study addresses these challenges by evaluating machine learning approaches for ERA5-based wind forecasting in Texas, focusing on temporal dependencies and spatial characteristics.

Data Description

We utilized ERA5 reanalysis data for Texas from 2018-2020 for model training, with 2021 data reserved for validation. ERA5 is the fifth generation ECMWF atmospheric reanalysis dataset, offering several advantages for wind modeling:

- Hourly meteorological estimates with global coverage
- 31 km spatial resolution (higher than previous reanalysis datasets)
- 137 vertical levels from surface to 80 km altitude
- Incorporates observations through 4D-Var data assimilation

Our dataset includes:

- Location identifiers (id, lat, lon)
- Height measurements
- Temporal information (year, day)
- ERA5 wind speed measurements (target variable)

Methods

Neural Network Design

- **Input Layer:** Accepts scaled features
- **Hidden Layer 1:** 64 neurons, ReLU activation for non-linearity
- **Hidden Layer 2:** 32 neurons, ReLU activation
- **Output Layer:** 1 neuron for predicting continuous ERA5 wind speed

Linear Regression Model

Initial baseline approach using temporal and spatial features:

- Temporal features: day of year, month
- Spatial features: latitude, longitude, height
- Target variable: ERA5 wind speed

Autoregressive Model (AR1)

Time series approach incorporating one-period lag:

- Models current wind speed as function of previous day
- Accounts for temporal autocorrelation
- Separate models fitted for each location

Random Forest - Basic

Ensemble approach without temporal dependencies:

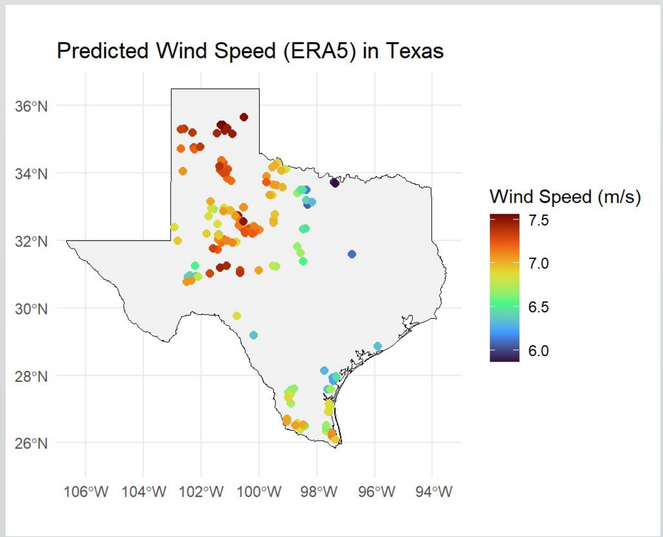
- Features: day, month, lat, lon, height
- 30 decision trees with bootstrap sampling
- Maximum depth of 10 to prevent overfitting

Random Forest with Lag Features

Advanced model incorporating temporal memory:

- Base features: day, month, lat, lon, height
- Added lag features: 1-day lag, 7-day lag
- Rolling prediction approach for 2021

Spatial Mapping based on Latitude and Longitude



Results

Model Evaluation

Metrics used to assess model performance:

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- Coefficient of Determination (R^2)
- Both overall and location-specific evaluation

Model	RMSE (m/s)	MAE (m/s)	R^2
Linear Regression	1.89	1.52	0.32
AR1	1.65	1.31	0.45
Random Forest	1.42	1.15	0.58
Random Forest w/ Lag	1.18	0.94	0.71
Neural network	4.46	1.70	0.09

Discussion

Our findings highlight several important aspects of wind speed prediction using ERA5 data. The significant improvement observed when incorporating lag features underscores the importance of temporal memory in wind prediction. This aligns with the physical understanding of wind patterns, which often exhibit strong autocorrelation over short time periods.

Conclusion

This study demonstrates that machine learning approaches, particularly random forests with lag features, can effectively predict wind speeds in Texas using ERA5 reanalysis data. Key conclusions include:

- Incorporation of temporal dependencies significantly improves prediction accuracy
- Random forest models outperform traditional statistical approaches
- ERA5 data provides a valuable foundation for wind modeling in Texas performance

References

1. Millstein, D., Cosgrove, J., et al. (2023). A database of hourly wind speed and modeled generation for US wind plants based on three meteorological models. Scientific Data, 10(1), 722.
2. Copernicus Climate Change Service (C3S). (2017). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate.
3. Davidson, M. R., & Millstein, D. (2022). Limitations of reanalysis data for wind power applications. Wind Energy, 25(11), 1646-1653.
4. Draxl, C., et al. (2015). The Wind Integration National Dataset (WIND) Toolkit. Applied Energy, 151, 355-366.