Data Article

Title: A Hybrid Linear--Gaussian Process Framework with Adaptive Covariance Selection for Spatio-Temporal Wind Speed Forecasting

Authors: [1]Thinawanga Hangwani Tshisikhawe, [1]Caston Sigauke, [2]Timotheous Brian Darikwa, [3]Saralees Nadarajah

Affiliations:

[1]Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou, 0950, South Africa.
[2]Department of Statistics and Operations Research, University of Limpopo, South Africa
[3]Department of Mathematics, Manchester University, United Kingdom

Contact email: caston.sigauke@univen.ac.za

Abstract

This research demonstrates the importance of employing regime-specific hybrid models to forecast wind speeds accurately. The authors adopted a linear regression method coupled with Gaussian Process (GP) residual modelling after clustering the sites geographically and by elevation. However, the results showed that they depended significantly on the regimes considered. For regional sites characterised by higher elevations (Cluster 2), employing the GP model with a Matern covariance function significantly improved forecasting, reducing errors by 16.3%. Conversely, for sites near the coast characterised by relatively lower elevations (Cluster 1), using GP for modelling the error did not improve the results but was detrimental to forecasting. The data are stored in an Excel file.

Specifications Table

Subject area	*Renewable Energy Modelling*
More specific subject area	*Predictive modelling of wind energy*
Type of data	*Excel file*
How the data was acquired	*Downloaded using the* https://wasadata.csir.co.za/wasa1/WASAData
Data format	*Raw and processed*
Experimental factors	*N/A*
Experimental features	*N/A*
Data source location	*Thohoyandou, South Africa (Latitude −22.95°, Longitude 30.49°)*
Data accessibility	*Publicly available from* https://wasadata.csir.co.za/wasa1/WASAData
Related research article	A Hybrid Linear--Gaussian Process Framework with Adaptive Covariance Selection for Spatio-Temporal Wind Speed Forecasting

Value of the Data

Its value lies in its high resolution and multidimensionality. It presents critical spatiotemporal features and meteorological information for robust wind-speed modelling. Extensive data preprocessing has made the data reliable and comparable. This enables the creation and verification of advanced hybrid forecasting models. The dataset is essential for developing regime-aware wind prediction techniques.

Data

The data used in this research are high-resolution wind data from the Wind Atlas South Africa (WASA), available at https://wasadata.csir.co.za/wasa1/WASAData. It was obtained from ten meteorological stations and includes wind speed at 62m hub height, along with other important predictors such as temperature, barometric pressure, relative humidity, and wind direction. Crucially, it captures the temporal (time dimension) and spatial (coordinates) dimensions, as well as meteorological dimensions; measurements are recorded every 10 minutes. The thorough preprocessing pipeline included the following steps: imputing missing values with feature-wise means and standardising all numeric features to have a mean of 0 and a variance of 1 for model stability and comparability. Further, spatial coordinates were normalised to remove scaling bias between latitude and longitude.

The stations were clustered using hierarchical clustering based on the agglomerative method. The optimal number of clusters was found to be two. This thorough preparation was necessary for the proper training of both the Linear Regression and Gaussian Process models used in the subsequent hybrid forecasting approach.

The analytic data can be accessed from https://github.com/csigauke?tab=repositories

Experimental Design, Materials, and Methods

The data used in the study are from the Wind Atlas of South Africa. The study compares the predictive abilities of a Hybrid Gaussian Process Regression (GPR) model combined with a linear regression model and an individual Gaussian Process (GP) applied to the residuals, enabling interpretable trend estimation and flexible modelling of spatio-temporal dependencies.

Acknowledgements

The authors acknowledge WASA for providing the data and the Universities of Venda, Limpopo, and Manchester for their use of resources in carrying out this study.

Funding sources

This research received no external funding

References

WASA https://wasadata.csir.co.za/wasa1/WASAData
(Accessed on 16 March 2025).