Improving an accuracy of ANN-based mesoscale-microscale coupling model by data categorization: with application to wind forecast for offshore and complex terrain onshore wind farms.

Alla Sapronova[1*], Catherine Meissner[2], Matteo Mana[2]

[1] UniResearch Ltd, University of Bergen, Bergen, Norway
alla.sapronova@uni.no
[2] WinSim Ltd., Tonsberg, Norway
{catherine, matteo}@windsim.no

Abstract. The ANN-based mesoscale-microscale coupling model forecasts wind speed and wind direction with high accuracy for wind parks located in complex terrain onshore, yet some weather regimes remains unresolved and forecast of such events failing. The model's generalization improved significantly when categorization information added as an input. The improved model is able to resolve extreme events and converged faster with significantly smaller number of hidden neurons. The new model performed equally good on test data sets from both onshore and offshore wind park sites.

Keywords: wind speed forecast; data categorization; artificial neural network.

1 Introduction

It is important to forecast an accurate wind energy yield as the wind energy production in many countries became a large part of the grid supply. The variability of wind energy remains main challenge for grid engineers, wind park owners, and electricity market players: i.e. practitioners, operating in the field where accurate forecast is required (often obliged by governmental regulations) in the range from minutes to 72 hours ahead.

To forecast a wind speed and directions, the most important components for energy yield prediction, Numerical Weather Prediction (NWP) mesoscale models at a coarse resolution of tens of kilometers are commonly used. For site-specific accurate forecast a coupling between mesoscale model output and observation at wind park location is crucial. This coupling requires modeling of the wind flow near the ground, where terrain roughness and complexity affect the flow at microscale. That involves predictive modeling of nonlinear multivariable function in the environment where explicit physics-based models either have limited application or not available.

Recently a new mesoscale-microscale coupling model was proposed [1]. The model is based on artificial intelligence methods (i.e. unsupervised machine learning) and uses NWP forecasts on temperature, pressure, relative humidity, wind speed, and wind direction) to issue a site-specific forecast. Trained against historical observations of wind speed / wind direction, the model forecasts mean hourly wind speed /wind direction to be further utilized as an input for flow model (CFD) at finer scales where roughness and wake effects are taking into account [2].

On the test data sets the model predicted the wind speed in a very satisfactory manner with MSE=1.8 (m/s) for one-hour ahead prediction. Even though such accuracy found to be superior to that based on polynomial fittings as well as ARMA models, the detailed statistics on model performance shows that for some weather regimes MSE was significantly higher than others.

In this work, the model proposed in [1], is developed further to achieve the same level of forecast accuracy for the entire data set, including weather regimes that original model was not able to catch. Improved model uses data categorization approach, when information obtained from categorization of a single variable (wind speed) was supplied to ANN input in addition to above mentioned NWP variables. Improved model showed lower MSE =1.1 (m/s) for wind speed prediction on the test data sets, resolving all weather regimes at the same level of accuracy.

2 Model selection

Machine learning (ML) based models' generalization arising from model's ability to find similarity in training data that usually consists of continuous numeric data. Since numbers are rarely exactly the same from one example to the next, the model can fails in selecting the margins for identical properties. In this case, the generalization can be improved by classification. For example, a combination of ML methods like self organizing maps (SOM) and feed-forward neural networks (FF ANN) can lead to forecast improvement [3]. Unfortunately, if the training data either limited or incomplete bringing SOM for forecast improvement became a challenge itself.

Also, some examples in the training set, that normally treated equally can vary in reliability or carry less critical information about the target than the others, or can even carry wrong information. E.g. wind speed measurement device on site can fail for variety of reasons, and sometimes with greater error in specific wind speed range(s).

According to the intrinsic margin, physical nature of the problem, etc, training data-sets can be grouped into several discrete categories. The discrete categories will allow identical category values to be treated in the same manner.

One logical approach is to categorize numeric data, a wind speed in this case, similar to typical human concepts (e.g. «calm wind», «fresh breeze», «strong wind», etc) and then try to generalize. The task can be defined as trying to divide numeric fields into appropriate sets of categories.

While most of the researches done on learning from examples has not been concerned with categorizing numeric data, some experimental results [4] show that choice of methods for categorization is irrelevant to the generalization improvement. Methods, quite different by nature, give similar and reasonable results with real-world data-sets and all lead to improved generalization

3 Data

In this work the data from ERG Wind Farm located in Molise in central Italy was used. The farm location is in a wide area between the town of Ururi and Montelongo, situated in complex terrain of different ridges and is about 10 km wide. The wind farm layout is composed by 20 Vestas V90 turbines.

The anemometer used as a reference, located in a central position of the wind farm at 30 m height.

The measurement data sets (anemometer registered values) start 1/04/2010 and end 1/03/2012. The time series almost two years long thus allow the entire year of data to be used in the training set and the rest period is used to validate the performance of the trained network.

The provided by Meteogroup, NWP data is a forecast of five days delivered four times a day. The time series start 1/04/2010 and end 6/05/2012 to keep a concurrent period with measured data. The data forecasted at two different heights: 80m and 30m; at both heights the variables forecasted are wind speed, wind direction, temperature, pressure, density and relative humidity.

Prior usage, the anemometer registered values were cleaned from invalid data and all the events with wind speed under 0.5 m/s were excluded. Then the measured and forecasted values were pre-processed and normalized. Finally, the data was fed into feed-forward neural network as described in Figure 1: the wind data of the anemometer was used as target for FF ANN training against NWP.

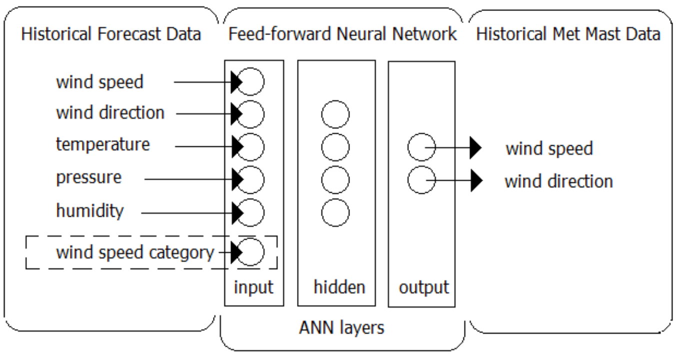

Fig. 1. Data usage for FF ANN training and validation

This model output further utilized for a correction that connects the raw forecast to the measured wind data as showed in Figure 2. This correction is used for the energy yield calculation in WindSim software.

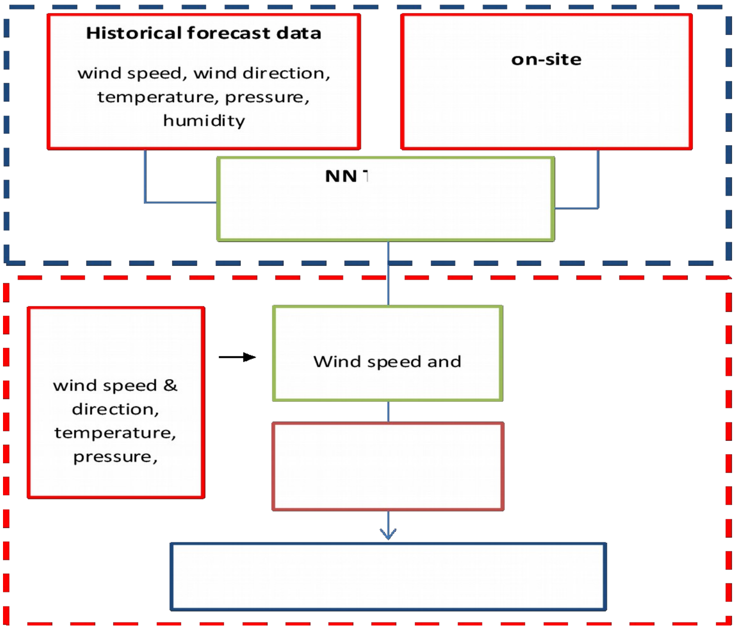

Fig. 2. The schema illustrates the employment of trained FF ANN for improved energy yield forecast

4 Results

The artificial ANN receives NWP data as an input to predict the wind speed inside the wind farm for one hour ahead. The wind speed prediction is done in two steps: data categorization and wind speed forecast.

To split the mesoscale NWP data into categories, a non-linear scale (to some extend similar to Beaufort scale) for wind speed has been used as shown in Table 1.

This type of categorization also allowed implicit data pre-processing, so that failed wind speed values with wind speed below zero combined into a separate category.

Obtained numerical attribute of the category has been fed as additional input to feed-forward ANN. For the training of ANN the wind speed category, wind direction, temperature, and pressure at 80m for one hour ahead NWP have been used as inputs.

On-site registered wind speed and direction at forecasted time have been used as desired output for training or for model test and validation.

Table 1. Wind speed ranges and corresponding numerical attributes used for categorization.

wind speed range, m/s	<0.5	0.5-1	1-3	3-7	7-10	10-15	15-20	20-25	>25
category	failed record	wind calm	light breeze	gentle breeze	fresh breeze	strong breeze	near gale	gale	cut off speed
numerical attribute	1	2	3	4	5	6	7	8	9

The performance of proposed model containing categorization values has been compared to a) non-categorization model and b) classification model, where input variables (excluding category variable) have been enhanced by classification identification obtained from various SOM. (In latest case, NWP parameters, like temperature, wind speed, wind directions, have been quantized by SOM and a separate ANN was used to find the correlation between inputs and desired outputs). Best SOM classification has been achieved by 2x3 matrix.

For non-categorization approach, two models were created (referred as model I and model II). In model I two inputs (wind speed and wind direction) from latest NWP run were used for ANN input. In model II 54 inputs, containing 6 variables (temperature, humidity, stability, pressure, wind speed, wind direction) issued by latest 6 NWP runs (1-72 hours ago) were used for ANN input. The summary for performance of all the above mentioned models shown in Table 2.

The proposed categorization approach has been tested on data from offshore wind farm: the new FF ANN was trained on the data a wind park consisting of 90 turbines and located in North Sea. The data contains hourly met mast reading and NWPs for years 2005-2007. The categorization model performance was compared with non-categorization model (both models had four inputs: wind speed category/wind speed NWP data, wind direction, temperature, and pressure and were trained against measured data). The comparison shows that categorization model performs equally good for offshore and onshore wind forecast: significant improvement in wind speed forecast for categorization model (RMSPE 3.2% vs 5.2%) was observed with number of hidden neurons lowered from 15 in non-categorization model to 7 in categorization model.

Table 2. Comparision of different FF ANN models' performance

Model	With categorization	Non-categorization I	Non-categorization II	Classification
ANN architecture with winning perfor-	4 input, 9 hidden, 1 output neurons	2 input, 30 hidden, 1 output neurons	36 input, 60 hidden, 1 output neurons	4 input, 30 hidden, 1 output neurons

mance				
Training time required to reach 0.01% training error, number of iterations rounded to thousands	20,000	240,000	2,185,000	438,000
Mean absolute percentage error	2.4%	4.6%	4.1%	5.0%
Root mean square percentage error (RMSPE)	5.4%	9.8%	8.8%	10.4%
Coefficient of determination (R^2)	0.83	0.74	0.77	0.69
Correlation	0.87	0.62	0.79	0.67

5 Conclusions

It is shown that ANN, that was previously used successfully for mesoscale-microscale models coupling can be improved significantly by adding categorization information.

It is observed that model with added categorization information has nearly twice lower RMSPE than regular model (5.4% vs 9.8%). Surprisingly, adding SOM classification output to the model input slightly lowered the generalization ability of the network (10.4% vs 9.8%). This can be explained in the way that after SOM quantized the entire data-set into 2x3 classes less examples became available for ANN training associated with each class. Assuming that enough data samples can be obtained to successfully train SOM-produced ANNs, it is important to note, that categorization model requires almost three times less hidden neurons (9 vs 30). Therefore data categorization more successfully lowers the data dimensions comparing to classification made by SOM. While, data classification is also valuable and shall not be underestimated. E.g in the current work, the visualization of SOM output provides additional information on data patterns and can lead to more successful choice of categories. Also classification patterns evolution can be used to study the dynamics of wind flow inside the wind farm.

Acknowledgment

This work is sponsored by Norwegian Research Council, project ENERGIX, 2013-2014.

References

1. A. Sapronova, C. Meissner, M. Mana, Mesoscale-microscale coupled model based on artificial neural network techniques for wind power forecast., Poster at EWEA Offshore 2014, PO ID 385.
2. M.Mana, Short-term Forecasting of wind energy production using CFD simulations., Poster at EWEA 2013.
3. E. Toth, Combined use of SOM-classification and Feed-Forward Networks for multinetwork streamflow forecasting., Geophysical Research Abstracts, Vol. 11, EGU2009-11962, 2009.
4. Ling Li, Amrit Pratap, Hsuan-Tien Lin, and Yaser S. Abu-Mostafa, Improving Generalization by Data Categorization., Lecture Notes in Computer Science Vol. 3721, 2005, pp 157-168.