SDWPF: A Dataset for Spatial Dynamic Wind Power Forecasting

Jingbo Zhou[1], Xinjiang Lu[1], Yixiong Xiao[1], Jiantao Su[3], Junfu Lyu[4], Yanjun Ma[2], Dejing Dou[1]
[1]Baidu Research [2]Baidu Inc. [3] Longyuan Power Group Corp. Ltd., [4]Tsinghua University
{zhoujingbo, luxinjiang, xiaoyixiong, mayanjun02, doudejing}@baidu.com
[3]12091329@chnenergy.com.cn, [4]lvjf@mail.tsinghua.edu.cn,

ABSTRACT

The variability of wind power supply can present substantial challenges to incorporating wind power into a grid system. Thus, Wind Power Forecasting (WPF) has been widely recognized as one of the most critical issues in wind power integration and operation. There has been an explosion of studies on wind power forecasting problems in the past decades. Nevertheless, how to well handle the WPF problem is still challenging, since high prediction accuracy is always demanded to ensure grid stability and security of supply. We present a unique Spatial Dynamic Wind Power Forecasting dataset: SDWPF, which includes the spatial distribution of wind turbines, as well as the dynamic context factors. Whereas, most of the existing datasets have only a small number of wind turbines without knowing the locations and context information of wind turbines at a fine-grained time scale. By contrast, SDWPF provides the wind power data of 134 wind turbines from a wind farm over half a year with their relative positions and internal statuses. We use this dataset to launch the Baidu KDD Cup 2022 to examine the limit of current WPF solutions. The dataset is released at https://aistudio.baidu.com/aistudio/competition/detail/152/0/datasets.

ACM Reference Format:

Jingbo Zhou[1], Xinjiang Lu[1], Yixiong Xiao[1], Jiantao Su[3], Junfu Lyu[4], Yanjun Ma[2], Dejing Dou[1] . 2022. SDWPF: A Dataset for Spatial Dynamic Wind Power Forecasting . In *Proceedings of The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Baidu KDD Cup 2022).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Wind Power Forecasting (WPF) aims to accurately estimate the wind power supply of a wind farm at different time scales. Wind power is a kind of clean and safe source of renewable energy, but cannot be produced consistently, leading to high variability. Such variability can present substantial challenges to incorporating wind power into a grid system. To maintain the balance between electricity generation and consumption, the fluctuation of wind power requires power substitution from other sources that might not be available at short notice (for example, usually it takes at least 6 hours to fire up a coal plant). Thus, WPF has been widely recognized as one of the most critical issues in wind power integration and operation. There has been an explosion of studies on wind power forecasting problems appearing in the data mining and machine learning community. Nevertheless, how to well handle the WPF problem is still challenging, since high prediction accuracy is always demanded to ensure grid stability and security of supply.

We present a unique Spatial Dynamic Wind Power Forecasting dataset: SDWPF, which includes the spatial distribution of wind turbines, as well as the dynamic context factors like temperature, weather, and turbine internal status. Whereas, many existing datasets and competitions treat WPF as a time series prediction problem without knowing the locations and context information of wind turbines.

SDWPF is obtained from the real-world data from Longyuan Power Group Corp. Ltd. (the largest wind power producer in China and Asia). There are two unique features for this competition task different from previous WPF competition settings: 1) Spatial distribution: this competition provides the relative location of all wind turbines given a wind farm for modeling the spatial correlation among wind turbines. 2) Dynamic context: the weather situations and turbine internal status detected by each wind turbine are provided to facilitate the forecasting task.

With aiming to examine the limit of WFP methods, we use the SDWPF dataset to launch the Baidu KDD Cup 2022 Challenge. SDWPF contains the wind power data obtained from the Supervisory Control And Data Acquisition (SCADA) system of a wind farm which has 134 wind turbines. The dataset provide the information about the wind, temperature, turbine angle and historical wind power. The time range of the dataset is over half a year. We also provide a baseline for this dataset[1]. The introduction about the challenge can found in the Baidu KDD Cup 2022 website[2] and the dataset can be down after registration[3].

2 RELATED WORK

Wind power forecasting (WPF) has been extensively investigated over the past decades [4, 5, 13, 15]. According to the spatial scale of the wind power, the problem can be categorised as a single wind turbine, a wind farm and a group of wind farms [9]. The dataset of this challenge belongs to the wind farm scale. A few of delicate models have been specially designed for WPF problem with variant of spatial and temporal scales based on statistic models [12, 13], machine learning methods [8, 17] and deep learning methods [6, 14]. Many advanced time series prediction methods like [7, 10, 11, 16, 18, 19] also have great potential to tackle this problem.

Though there are a few of public WPF datasets, they usually have only a limited number of wind turbines and do not provide the spatial information of each turbine. For example, the Penmanshiel dataset has only 14 turbines [1], and the Kaggle dataset has only 1

Baidu KDD Cup 2022, Mar. 16 – Jul. 17, 2022,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1]https://github.com/PaddlePaddle/PaddleSpatial/tree/main/apps/wpf_baseline_gru
[2]https://aistudio.baidu.com/aistudio/competition/detail/152/0/introduction
[3]https://aistudio.baidu.com/aistudio/competition/detail/152/0/datasets

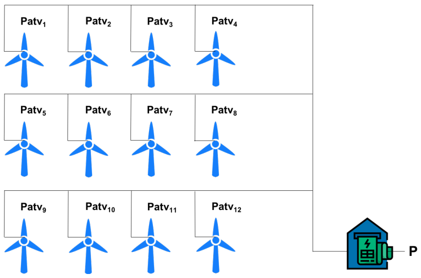

Figure 1: An illustration of a wind farm.

turbine [2]. We leave a comprehensive discussion about the WPF methods and WPF datasets as a future work.

3 SDWPF DATASET

In this section, we provide an brief introduction of the SDWPF dataset, including data source, overall statistics, schema and spatial distribution. The SDWPF dataset is collected from the Supervisory Control and Data Acquisition (SCADA) system of a wind farm. The SCADA data are sampled **every 10 minutes** from each wind turbine in the wind farm which consists of 134 wind turbines. The statistics of the important information of the SDWPF dataset is shown in the Table 1.

The dataset includes critical external features, such as wind speed, wind direction and external temperature, that influence the wind power generation; as well as essential internal features, such as the inside temperature, nacelle direction and Pitch angle of blades, which can indicate the operating status of each wind turbine.

Each wind turbine can generate the wind power $Patv_i$ separately, and the outcome power of the wind farm is the sum of all the wind turbines. In other words, at time t, the output power of the wind farm is $P = \sum_i Patv_i$. An illustration of a wind farm is shown in Figure 1. We also provide a detailed introduction about the main attributes of the data in Table 2. Please refer to Wikipedia for more details about components of wind turbines[4].

The relative position of all wind turbines in the wind farm is also released to characterize the spatial correlation between wind turbines. An illustration of the spatial distribution of the totally 134 wind turbines are shown in Figure 2. The units of x and y are meter.

4 EVALUATION

The Baidu KDD Cup 2022 requires to address the Spatial Dynamic Wind Power Forecasting ahead of 48 hours. For example, given at 6:00 A.M. today, it is required to effectively forecast the wind power generation beginning from 6:00 A.M. on this day to 5:50 AM on the day after tomorrow, given a series of historical records of the wind farm and the related wind turbines. It is required to output the predicted values every 10 minutes. To be specific, at one time point, it is required to predict a future length-288 wind power supply time-series. The average of RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) is used as the main evaluation score. Formally, at a time step t_0, it is required to predict a time series of

[4]We suggest to refer to https://en.wikipedia.org/wiki/Wind_turbine#Horizontal_axis and https://en.wikipedia.org/wiki/Wind_turbine#Components

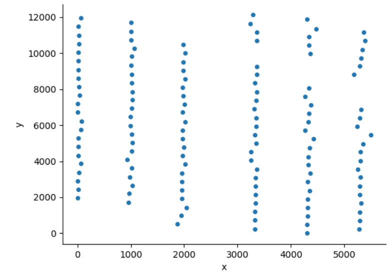

Figure 2: Spatial distribution of all wind turbines (x and y are with the meter unit).

wind power of the wind farm $P = \{p_{t_0+1}, p_{t_0+2}, \cdots, p_{t_0+288}\}$. The evaluation score s_{t_0} at the time step t_0 is defined as:

$$S_{t_0} = \frac{1}{2}\left(\sqrt{\frac{\sum_{i=1}^{288}(p_{t_0+i} - \hat{p}_{t_0+i})^2}{288}} + \frac{\sum_{i=1}^{288}|p_{t_0+i} - \hat{p}_{t_0+i}|}{288}\right) \quad (1)$$

where p_{t_0+i} is the actual power and $\hat{p}_{t_0+i}$ is the predicted power of the wind farm at time step t_0. A length-L_x-length-288 prediction window is adopted to roll the whole test set with stride 1 (i.e. 10 minutes), and the averaged evaluation score is reported. Note that, L_x denotes the length of input time-series. For example, if one aims to forecast the wind power in future seven days, the evaluation score should be calculated as follows:

$$score = \sum_{k=0}^{7*24*6} S_{t_0+k*10min} \quad (2)$$

The code to calculate the score is available in our sample code.

Note in our settings, we aim to forecast the power generated by a wind farm with the SCADA data and spatial data on top of the spatiotemporal modeling paradigm without knowing the future meteorological data (wind speed, temperature, etc.). During the Baidu KDD Cup 2022 challenge, except the released data of 184 days, we still privately hold data of several months to evaluate the submitted models by participants.

4.1 Caveats about the data

Here we introduce a few of caveats when to use this data to train and evaluate the models.

Zero values. There are some active power and reactive power which are smaller than zeros. We simply treat all the values which are smaller than 0 as 0, i.e. if $Patv < 0$, then $Patv = 0$.

Missing values. Note that due to some reasons, some values at some time are not collected from the SCADA system. These missing values will not be used for evaluating the model. In other word, if p_{t_0+j} is a missing value, we set $|p_{t_0+j} - \hat{p}_{t_0+j}| = 0$ regardless of the actual predicted value of $\hat{p}_{t_0+j}$.

Unknown values. In some time, the wind turbines are stopped to generate power by external reasons such as wind turbine renovation and/or actively scheduling the powering to avoid overloading the grid. In these cases, the actual generated power of the wind

# of days	Time interval	# of columns	# of turbines	# of records
184	10 minutes	13	134	3,550,465

Table 1: Statistics of the SDWPF data.

Column	Column Name	Specification
1	TurbID	Wind turbine ID
2	Day	Day of the record
3	Tmstamp	Created time of the record
4	Wspd (m/s)	The wind speed recorded by the anemometer
5	Wdir(°)	The angle between the wind direction and the position of turbine nacelle
6	Etmp (°C)	Temperature of the suronding environment
7	Itmp (°C)	Temperature inside the turbine nacelle
8	Ndir (°)	Nacelle direction, i.e., the yaw angle of the nacelle
9	Pab1 (°)	Pitch angle of blade 1
10	Pab2 (°)	Pitch angle of blade 2
11	Pab3 (°)	Pitch angle of blade 3
12	Prtv (kW)	Reactive power
13	Patv (kW)	Active power (target variable)

Table 2: Column names and their specifications of the SDWPF data.

turbine is unknown. These unknown values will also not be used for evaluating the model. Similarly with the missing values, if p_{t_0+j} is a unknown value, we always set $|p_{t_0+j} - \hat{p}_{t_0+j}| = 0$. Here we introduce two conditions to determine whether the target variable is unknown:

- If at time t, $Patv \leq 0$ and $Wspd > 2.5$, then the actual active power $Patv$ of this wind turbine at time t is unknown;
- If at time t, $Prab1 > 89^\circ$ or $Prab2 > 89^\circ$ or $Prab3 > 89^\circ$, then the actual active power $Patv$ of this wind turbine at time t is unknown.

5 BASELINE CODE

We have released a simple baseline code with Gated Recurrent Unit [3] in PaddleSpatial[5]. In our experiment, we sum the active power of all the wind turbines to form a wind power time series. And then we use the data of first 153 days as training data, the middle 16 days as validation data, and the last 15 days as test data. According to our statistics, the score of our baseline over the tested time series of 15 days is: RMSE: 50.045929, MAE: 39.774162, and the overall score is 44.910046.

[5]https://github.com/PaddlePaddle/PaddleSpatial/tree/main/apps/wpf_baseline_gru

REFERENCES

[1] 2022. Penmanshiel (United-Kingdom) dataset. https://www.thewindpower.net/windfarm_en_23147_penmanshiel.php. Online; accessed 06 April 2022.
[2] 2022. Wind Power Forecasting (Kaggle). https://www.kaggle.com/datasets/theforcecoder/wind-power-forecasting. Online; accessed 06 April 2022.
[3] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111.
[4] Xing Deng, Haijian Shao, Chunlong Hu, Dengbiao Jiang, and Yingtao Jiang. 2020. Wind power forecasting methods based on deep learning: A survey. *Computer Modeling in Engineering and Sciences* 122, 1 (2020), 273.
[5] Aoife M Foley, Paul G Leahy, Antonino Marvuglia, and Eamon J McKeogh. 2012. Current methods and advances in forecasting of wind power generation. *Renewable energy* 37, 1 (2012), 1–8.
[6] Ying-Yi Hong and Christian Lian Paulo P Rioflorido. 2019. A hybrid deep learning-based neural network for 24-h ahead wind power forecasting. *Applied Energy* 250 (2019), 530–539.
[7] Jun Hu and Wendong Zheng. 2020. Multistage attention network for multivariate time series prediction. *Neurocomputing* 383 (2020), 122–137.
[8] Qinghua Hu, Shiguang Zhang, Man Yu, and Zongxia Xie. 2015. Short-term wind speed or power forecasting with heteroscedastic support vector regression. *IEEE Transactions on Sustainable Energy* 7, 1 (2015), 241–249.
[9] Zhao-Yu Jiang, Qing-Shan Jia, and XH Guan. 2019. A review of multi-temporal-and-spatial-scale wind power forecasting method. *Acta Automatica Sinica* 45, 1 (2019), 51–71.
[10] Ting Li, Junbo Zhang, Kainan Bao, Yuxuan Liang, Yexin Li, and Yu Zheng. 2020. Autost: Efficient neural architecture search for spatio-temporal prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 794–802.
[11] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. 2018. Geoman: Multi-level attention networks for geo-sensory time series prediction.. In *IJCAI*, Vol. 2018. 3428–3434.
[12] M Milligan, M Schwartz, and Yih-huei Wan. 2003. *Statistical wind power forecasting models: Results for US wind farms.* Technical Report. National Renewable Energy Lab.(NREL), Golden, CO (United States).
[13] George Sideratos and Nikos D Hatziargyriou. 2007. An advanced statistical method for wind power forecasting. *IEEE Transactions on power systems* 22, 1 (2007), 258–265.
[14] Huai-zhi Wang, Gang-qiang Li, Gui-bin Wang, Jian-chun Peng, Hui Jiang, and Yi-tao Liu. 2017. Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied energy* 188 (2017), 56–70.
[15] Xiaochen Wang, Peng Guo, and Xiaobin Huang. 2011. A review of wind power forecasting models. *Energy procedia* 12 (2011), 770–778.
[16] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems* 34 (2021).
[17] Jianwu Zeng and Wei Qiao. 2011. Support vector machine-based short-term wind power forecasting. In *2011 IEEE/PES Power Systems Conference and Exposition*. IEEE, 1–8.
[18] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI*.
[19] Jingbo Zhou and Anthony KH Tung. 2015. Smiler: A semi-lazy time series prediction system for sensors. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 1871–1886.