

SDWPF: KDD 杯 2022 空间动态风力发电预测挑战的数据集

Jingbo Zhou^{1*}, Xinjiang Lu^{1*}, Yixiong Xiao^{1*}, Jiantao Su³, Junfu Lyu⁴, Yanjun Ma², Dejing Dou¹

¹Baidu Research ²Baidu Inc. ³Longyuan Power Group Corp. Ltd., ⁴Tsinghua University

{zhoujingbo, luxinjiang, xiaoyixiong, mayanjun02, doudejing}@baidu.com

³12091329@chnenergy.com.cn, ⁴lvjf@mail.tsinghua.edu.cn,

摘要

风力发电供应的波动性给将风力发电整合到电网系统中带来了重大挑战。因此，风电预测（WPF）已被广泛认为是风能集成和运行中最关键的问题之一。过去几十年来，关于风电预测问题的研究呈爆炸式增长。然而，如何妥善处理 WPF 问题仍然是一个挑战，因为始终需要高精度的预测以确保电网稳定性和供电安全。我们提出了一套独特的空间动态风电预测数据集：SD-WPF，其中包括风力涡轮机的空间分布以及动态环境因素。相比之下，现有的大多数数据集只有少量的风力涡轮机，并且不了解这些风力涡轮机的位置和环境信息在细粒度的时间尺度。相反，SDWPF 提供了来自一个风电场 134 台风力涡轮机半年的发电数据及其相对位置和内部状态。我们使用此数据集启动了百度 KDD 杯 2022 比赛以测试当前 WPF 解决方案的极

限。数据集已发布于 <https://aistudio.baidu.com/aistudio/competition/detail/152/0/datasets>。

ACM Reference Format:

Jingbo Zhou^{1*}, Xinjiang Lu^{1*}, Yixiong Xiao^{1*}, Jiantao Su³, Junfu Lyu⁴, Yanjun Ma², Dejing Dou¹. 2025. SDWPF: KDD 杯 2022 空间动态风力发电预测挑战的数据集. In *Proceedings of The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Baidu KDD Cup 2022)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 介绍

风力发电预测（WPF）旨在准确估计风电场在不同时间尺度上的风力发电量。风能是一种清洁和安全的可再生能源，但不能稳定生产，导致其具有高度变异性。这种变异性可能给将风力发电纳入电网系统带来重大挑战。为了保持电力生产和消费之间的平衡，风力发电的波动需要其他来源的电力替代，而这些来源可能无法在短时间内提供（例如，通常至少需要 6 小时才能启动一个燃煤电厂）。因此，WPF 已被广泛认可为风电集成和运营中最关键的问题之一。关于风力发电预测问题的研究在过去出现了爆炸性增长，在数据挖掘和机器学习领域尤为明显。然而，如何妥善处理 WPF 问题仍然是一个挑战，因为始终需要高精度的预测来确保电网稳定性和供电安全。

我们提供了一个独特的空间动态风电预测数据集：SDWPF，其中包括风力涡轮机的空间分布以及诸如温度、天气和涡轮机内部状态等动态上下文因素。而许

* Equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2025 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

多现有的数据集和竞赛将风电预测视为一个时间序列预测问题,而不考虑风力涡轮机的位置和上下文信息。

SDWPF 是从龙源电力集团股份有限公司(中国及亚洲最大的风电生产商)的真实世界数据中获得的。本次竞赛任务与其他 WPF 竞赛设置相比有两个独特的特征: 1) 空间分布: 此次比赛提供了给定风电场内所有风力涡轮机的相对位置, 以建模风力涡轮机之间的空间相关性。2) 动态环境: 每个风力涡轮机检测到的天气情况和内部状态被提供出来以促进预测任务。

旨在考察 WPF 方法的极限, 我们使用 SDWPF 数据集发起了百度 KDD 杯 2022 挑战赛。SDWPF 包含来自一个拥有 134 台风力涡轮机的风电场的监视控制和数据采集(SCADA)系统的风力发电数据。该数据集提供了关于风速、温度、涡轮角度及历史风力发电量的信息。数据集的时间范围超过半年。我们还为该数据集提供了一个基线¹。有关挑战赛的介绍可以在百度 KDD 杯 2022 网站²找到, 注册后可以下载数据集³。

2 相关工作

风力发电预测(WPF)在过去几十年中得到了广泛研究[4, 5, 13, 15]。根据风电的时空尺度, 该问题可以分为单个风力涡轮机、一个风电场和一组风电场[9]。本次挑战的数据集属于风电场规模。一些精细模型基于统计模型[12, 13]、机器学习方法[8, 17]和深度学习方法[6, 14]专门设计用于处理具有不同时空尺度的WPF问题。许多先进的时序预测方法如[7, 10, 11, 16, 18, 19]也具备解决此问题的巨大潜力。

尽管有一些公开的WPF数据集, 但它们通常只包含有限的数量的风涡轮机并且不提供空间信息每个风力涡轮机。例如, Penmanshiel数据集只有14台风力涡轮机[1], 而Kaggle数据集只有1台风力涡轮机[2]。我们把关于WPF方法和WPF数据集的全面讨论留作未来的工作。

¹https://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>

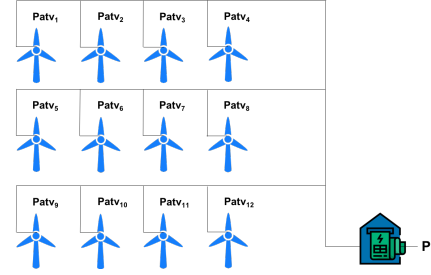


图 1: 一个风电场的示例。

3 SDWPF 数据集

在本节中, 我们提供了SDWPF数据集的简要介绍, 包括数据来源、总体统计信息、模式和空间分布。SDWPF数据集是从该监控和数据采集(SCADA)系统 của 风电场收集的。SCADA数据是从风电场中的每台风力涡轮机中采样每 10 分钟得到的, 该风电场由 134 台风力涡轮机组成。SDWPF数据集的重要信息统计如表 1 所示。

Days	Interval	# of columns	# of turbines	# of records
245	10 minutes	13	134	4,727,520

表 1: SDWPF 数据统计。

该数据集包括影响风力发电的关键外部特征, 如风速、风向和外部温度; 以及指示每个风力涡轮机操作状态的基本内部特征, 如内部温度、机舱方向和叶片的桨距角。

每个风力涡轮机可以单独产生 $Patv^i$ 的风能, 而风电场的总输出功率是所有风力涡轮机功率之和。换句话说, 在时间 t 时, 风电场的输出功率为 $P = \sum_i Patv^i$ 。一个风电场的示例如图 1 所示。我们还在表 2 中提供了关于数据主要属性的详细介绍。有关风力涡轮机组件⁴的更多细节, 请参阅维基百科。

风电场中所有风力涡轮机的相对位置也被发布, 以表征风力涡轮机之间的空间相关性。总共 134 台风力涡轮机的空间分布示例如图 2 所示。x 和 y 的单位是米。

⁴我们建议参考 https://en.wikipedia.org/wiki/Wind_turbine#Horizontal_axis 和 https://en.wikipedia.org/wiki/Wind_turbine#Components

Column	Column Name	Specification
1	TurbID	Wind turbine ID
2	Day	记录日期
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 surrounding 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)

表 2: SDWPF 数据的列名及其规格。

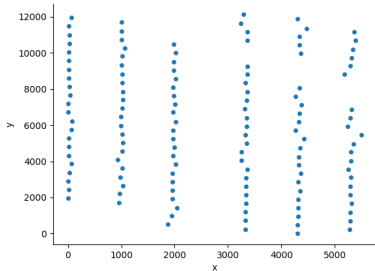


图 2: 所有风力涡轮机的空间分布 (x 和 y 的单位为米)。

4 评估

百度 KDD 杯 2022 要求解决 48 小时前的时空动态风力发电预测问题。例如，今天早上 6:00 给出的数据需要有效预测从当天早上 6:00 到 5:50 后天的风电场和相关的台风力发电机的历史记录序列。要求每 10 分钟输出一次预测值。具体来说，在一个时间点，需要预测长度为 288 的未来风电供应时间序列。均方根误差 (RMSE) 和平均绝对误差 (MAE) 的平均值被用作主要评估分数。

请注意，在我们的设置中，我们旨在通过 SCADA 数据和空间数据在时空建模范式上预测风电场产生的电力，而不了解未来的气象数据（风速、温度等）。在 2022 年百度 KDD 杯挑战赛期间，除了发布的 245 天数据外，我们仍保留几个月的数据以评估参赛者提交的模型。在给出指标的正式定义之前，我们先介绍一些关于数据的注意事项。

4.1 关于数据的注意事项

这里我们介绍了一些在使用这些数据训练和评估模型时需要注意的事项。**零值**。有一些有功功率和无功率小于零。我们将所有小于 0 的值简单地视为 0，即。如果 $Patv < 0$ ，则 $Patv = 0$ 。**缺失值**。请注意，由于某些原因，有些时间点的数据没有从 SCADA 系统中采集到。这些缺失值将不用于评估模型。换句话说，如果 p_{t_0+j} 是一个缺失值，我们会设置 $|Patv_{t_0+j} - \overline{Patv}_{t_0+j}| = 0$ ，而不考虑实际预测值 $\overline{Patv}_{t_0+j}$ 。**未知值**。一段时间后，由于风力涡轮机翻新和/或主动调度供电以避免电网过载等外部原因，风力涡轮机会停止发电。在这些情况下，风力涡轮机的实际发电量是未知的。这些未知值也不

会用于评估模型。与缺失值类似，如果 $Patv_{t_0+j}$ 是一个未知值，我们总是设置 $|Patv_{t_0+j} - \overline{Patv_{t_0+j}}| = 0$ 。这里我们引入两个条件来确定目标变量是否为未知：

- 如果在时间 t ， $Patv \leq 0$ 和 $Wspd > 2.5$ 时，那么这台风力涡轮机在时间 t 的实际有功功率 $Patv$ 是未知的；
- 如果在时间 t ， $Pab1 > 89^\circ$ 或 $Pab2 > 89^\circ$ 或 $Pab3 > 89^\circ$ ，则该风力涡轮机在时间 t 的实际有功功率 $Patv$ 是未知的。

Ndir 的合理范围是 $[-720^\circ, 720^\circ]$ ，因为涡轮系统允许机舱最多在一个方向上旋转两圈，否则会迫使机舱返回原始位置。因此，超出此范围的记录可以被视为由记录系统引起的异常值。因此，如果在时间 t 处有 $Nidir > 720^\circ$ 或 $Nidir < -720^\circ$ ，则该风力涡轮机在时间 t 的实际有功功率 $Patv$ 是异常的。

Wdir 的合理范围是 $[-180^\circ, 180^\circ]$ 。超出此范围的记录可以视为由记录系统引起的异常值。如果在时间点 t 处 $Widr > 180^\circ$ 或 $Widr < -180^\circ$ ，则该风力涡轮机在时间 t 的实际有功功率 $Patv$ 是异常的。

4.2 评估指标

形式上，在时间步长 t_0 ，需要预测风场 $P = \{p_{t_0+1}, p_{t_0+2}, \dots, p_{t_0+288}\}$ 的风电时间序列。然而，由于每个风力涡轮机的缺失和未知值，在这个挑战中，我们评估每个风力涡轮机的预测结果，然后将预测得分相加作为模型的最终得分。在时间步长 t_0 时，对于风力涡轮机 i 的评价分数 $s_{t_0}^i$ 定义为：

$$s_{t_0}^i = \frac{1}{2} \left(\sqrt{\frac{\sum_{j=1}^{288} (Patv_{t_0+j}^i - \overline{Patv_{t_0+j}^i})^2}{288}} + \frac{\sum_{j=1}^{288} |Patv_{t_0+j}^i - \overline{Patv_{t_0+j}^i}|}{288} \right) \quad (1)$$

其中， $Patv_{t_0+j}^i$ 是风力涡轮机 i 的实际功率，而 $\overline{Patv_{t_0+j}^i}$ 是风力涡轮机 i 在时间步长 t_0+j 处的预测功率。请注意，每个时间步长 j 为 10 分钟。预测模型 S_{t_0} 在时间 t_0

的整体分数是所有风力涡轮机预测分数的总和，即。

$$S_{t_0} = \sum_{i=1}^{134} s_{t_0}^i \quad (2)$$

一个长度为 L_x -长度-288 的预测窗口被采用，以步长 Δt 时间步骤滚动整个测试集（ Δt 的每个时间步骤是 10 分钟），并报告平均评估分数。请注意， L_x 表示输入时间序列的长度。

我们使用 K 数据实例来评估预测模型的性能。对于每个数据实例 k ，我们从范围 $[1, 10]$ 中随机采样一个步长时间步骤 Δt_k 。换句话说，步长时间步骤随机分布在 10 分钟到 100 分钟之间。形式上，模型的评估分数是：

$$score = \frac{1}{K} \sum_{k=0}^K S_{t_0 + \sum_{r=0}^k \Delta t_r} \quad (3)$$

计算得分的代码在我们的示例代码中可用。

如我们在示例代码中所示，由于我们对所有风力涡轮机的误差进行了求和，为了防止数值过大，我们使用兆瓦（而不是千瓦）作为单位来表示最终得分。

对于我们在提交系统中的评估（从 5 月 10 日开始），我们使用了 195 个随机抽样的步进时间步骤（即方程 3 中的 $\{\Delta t_0, \Delta t_1, \dots, \Delta t_{194}\}$ ）来评估提交的模型。此测试时间段和抽样步进时间步骤将在未来的阶段中更新（6 月 20 日和 7 月 15 日）。

5 基线代码

我们发布了一个带有门控循环单元 [3] 的简单基线代码在 PaddleSpatial⁵ 中。在我们的实验中，我们将所有风力涡轮机的有功功率相加形成一个风电时间序列。然后我们使用前 214 天的数据作为训练数据，剩下的 31 天作为验证数据。根据我们的统计，我们的基线在整个测试时间序列的 195 个预测（即 $K = 195$ ）上的得分是：RMSE: 47.081286，MAE: 37.558233，总体得分为 42.319760。在配备 Nvidia P40 GPU 的 Linux 机器上，对

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

这 195 个预测进行评估所需的时间是 1129.722821 秒。请注意，批量大小会影响性能评估。为了缓解这个问题，评估采用了不使用批处理策略的方法，并逐一测试每个实例的性能。

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 McKeeogh. 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 Hatziairgiou. 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.