Preliminary analysis of selected wind turbine SCADA datasets

Kelmarsh

The first analysed dataset comes from Kelmarsh wind farm located in the UK. Data was collected using the SCADA system (Supervisory Control And Data Acquisition) from January 1st to June 30th 2021 from six wind turbines. The measurements were averaged over 10-minutes time blocks assuming that the variables were nearly constant during this time.

Results

The following results were obtained during the first analysis of the data:

	Turbine 01	**Turbine 02**	**Turbine 03**	**Turbine 04**	**Turbine 05**	**Turbine 06**
Columns (variables)	299	299	299	299	299	299
Datapoints	26064	26064	26064	26064	26064	26064
First timestamp	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00
Last timestamp	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00
Missing values	564309	563890	561447	567260	568249	583129

This analysis shows a large number of missing values. After taking a closer look at the data, we noticed that some of the columns contain only missing data for all turbines. To be precise, the following columns turned out to be empty:

1. 'Lost Production (Contractual Global) (kWh)',
2. 'Lost Production (Contractual Custom) (kWh)',
3. 'Potential power met mast anemometer (kW)',
4. 'Potential power met mast anemometer MPC (kW)',
5. 'Time-based Contractual Avail. (Global)',
6. 'Time-based Contractual Avail. (Custom)',
7. 'Production-based Contractual Avail. (Global)',
8. 'Production-based Contractual Avail. (Custom)',
9. 'Equivalent Full Load Hours counter (s)'

For this reason, we decided to remove those variables from the dataset, which led to results presented below:

	Turbine 01	Turbine 02	Turbine 03	Turbine 04	Turbine 05	Turbine 06
Columns (variables)	290	290	290	290	290	290
Missing values	329733	329314	326871	332684	333673	348553

Penmanshiel

The second analysed dataset comes from a British wind farm located in Penmanshiel. As above, the SCADA system with 10-minutes time blocks was used. Data was collected from January 1st to June 30th 2021 from five wind turbines.

Results

The preliminary analysis led to following results:

	Turbine 01	**Turbine 02**	**Turbine 03**	**Turbine 04**	**Turbine 05**
Columns (variables)	300	300	300	300	300
Datapoints	26064	26064	26064	26064	26064
First timestamp	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00	2021-01-01 00:00:00
Last timestamp	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00	2021-06-30 23:50:00
Missing values	459162	361192	367388	359515	363038

For this dataset, the nine columns mentioned above also turned out to be empty. After removing them, we obtained the following results:

	Turbine 01	**Turbine 02**	**Turbine 03**	**Turbine 04**	**Turbine 05**
Columns (variables)	291	291	291	291	291
Missing values	224586	126616	132812	124939	128462