Wind Turbine Project

Olivier Kanamugire, Durbar Hore Partha, Bright Wiredu Nuakoh

September 2024

---

1 Project Overview and Objectives

This project focuses on fault detection in wind turbines using control charts and sensor diagnostics. By developing a model based on data from healthy turbines, we aim to capture faults in malfunctioning turbines. The approach involves optimizing the number of principal components in the healthy turbine model and creating multivariate control charts, specifically T^2 and SPEx charts, to monitor performance. By analyzing the contributions of individual out-of-control observations, the goal is to identify which sensors are most effective at detecting faults.

The specific objectives of the project are to:

- Optimize the number of principal components in the healthy turbine model.
- Develop multivariate control charts (T^2 and SPEx) to monitor turbine performance.
- Detect faults in faulty turbines using control charts developed from healthy turbine data.
- Analyze out-of-control observations to identify sensors that can accurately capture faults.

2 Exploratory Data Analysis of the Wind Turbine Failure Dataset

2.1 Dataset description

The wind turbine failure dataset contains sensor measurements recorded at 10-second intervals from multiple wind turbines. The dataset is divided into four subsets: WT2, WT3, WT14, and WT39, of which three (WT3, WT14, and WT39) represent faulty turbines, while WT2 is functioning well. Although the exact names of the variables are not provided, they correspond to sensor readings that reflect different physical processes and operational properties of the turbines like spinner temperature, Nacelle temperature, etc.

For this task, three turbines—WT2, WT14, and WT39—were selected based on inspection of their means from Figure?? to achieve our objectives. It is obvious from the table above that the center of Variables on W3 differ from other tables hence resulting in our decision to exclude it. WT2 contains 1,571 observations with 28 features. Some variables have very small values, while others have significantly large values, indicating the need to standardize the data. For instance, the first variable ranges from -0.16052 to 1, further reinforcing the need for standardization.

Identification of missing values: Among the selected subsets of our dataset, only one has a missing value and we have decided to remove that observation since it has no much effect on the data distribution!

2.2 Identification of outliers using box plots

The figure below contains four box plots for the four turbines. Since the dataset includes 28 attributes, we have randomly selected 10 variables for easy data visualization.

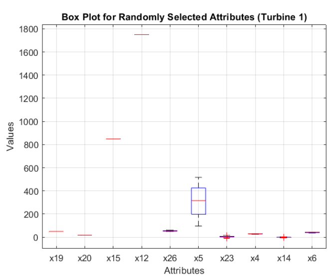

Figure 1: Box plot for Turbine 4

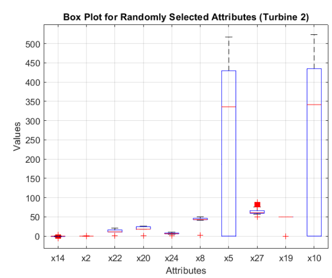

Figure 2: Box plot for Turbine 3

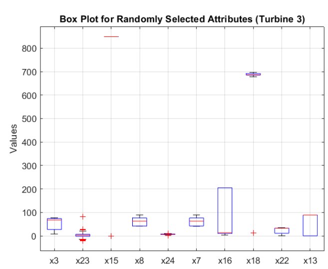

Figure 3: Box plot for Turbine 1

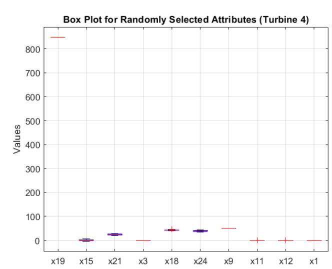

Figure 4: Box plot for turbine 2

In Figure 4, we observe that the range of values for the 15th variable is wider compared to that of the 20th variable. This pattern is consistent across

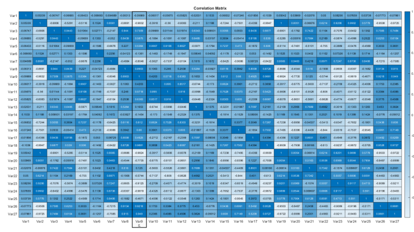

Figure 5: Correlation matrix for Turbine 1.

all turbines.

2.3 Is it necessary to do dimensional reduction? Are our variables correlated?

It is important to perform dimensionality reduction when the variables are correlated. Here we have used a correlation matrix to see if we can get insights about that.

Considering the figure 5, we see that some of the variables are highly correlated up to 90%. Therefore, dimensional reduction technique is required!
[1]

3 Data Pretreatment

3.1 Which variables should be removed and which turbines should we choose for our model?

Our goal is to ensure that all selected turbine datasets have an identical number of variables. We noted that the datasets contain 27 variables for WT14, 27 for WT39, 28 for WT2, and 31 for WT3. To achieve consistency, we identified the intersection of variable names across these datasets, resulting in a total of 27 common variables.

Additionally, in preparation for future analysis, such as standardization, we removed variables with constant values that were identical across all turbines. Specifically, the variables 'x12' and 'x15' were excluded from the datasets. But why do we exclude WT3? Considering figure 6, we see the means for every variable on each turbine. So if we compare the means, we see that WT3 is far from the rest in almost all variables.

[1] The team has chosen Microsoft Teams for communication and idea sharing, and GitHub will be used for code management and collaboration.

	WT2_Mean	WT3_Mean	WT14_Mean	WT39_Mean
x1	-0.0074363	0	-0.0093686	-0.0087091
x2	-0.011617	49.986	-0.0082247	-0.010234
x3	26.312	0.044193	50.91	28.644
x4	29.28	13.457	51.757	31.592
x5	315.31	12.315	199.95	266.6
x6	41.044	67.309	59.391	44.256
x7	40.906	88.993	59.234	44.217
x8	41.08	0	59.306	44.25
x9	2.3575e+07	50	2.4832e+07	2.2362e+07
x10	320.62	89	202.31	270.12
x11	1320.3	-0.0098357	698.93	962.62
x12	1750	-0.0098731	864.01	1181.3
x13	0.011046	0.018486	46.27	29.711
x14	-0.75414	0.015096	-0.28321	-0.52242
x15	849	0.42006	849	568.31
x16	223.68	4.0614	101.66	240.06
x17	9.0433	2.8136	24.553	11.059
x18	685.09	42.452	687.91	685.22
x19	50.002	849	50.017	49.962
x20	17.623	41.722	27.291	20.238
x21	22.912	24.338	32.181	24.5
x22	10.736	27.727	22.609	12.861
x23	5.1643	38.24	1.9584	3.9436
x24	5.4798	38.849	6.9757	6.8836
x25	16.923	38.913	35.43	14.099
x26	53.211	-0.15317	63.912	63.932
x27	59.822	-0.32617	65.019	63.845

Figure 6: Comparing the means of the Variables for all wind turbines

3.2 Standardization Of Data

We implemented the Z-score standardization method on the healthy turbine W2 to center each variable to zero mean and one standard deviation. Later the mean and standard deviation for the Healthy turbine was used to center and scale the chosen faulty turbines.

There was not much data synchronization done since all the observations were captured within 10 seconds implying that all variables have the same scale.

4 Visualization of Pretreated Data

Figure 7: Boxplot for random chosen Variables

Considering boxplots for variable X23 and X2 form Figure7, after applying the mean and standard deviation of healthy turbine H on faulty turbines F1 and F2, we observe that the observations have similar distribution and closer means. However, there are identified outliers in both faulty turbines and the healthy one. Additionally, it also observed that in variable x10 and X11, the mean of faulty turbine F1 has deviated from zero and the faulty turbines have very large variations.