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 normally. 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 to achieve our objective. 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 variable in the 15th column has a minimum value of 15 and a maximum of 849, which is a considerable difference. Additionally, the first variable ranges from -0.16052 to 1, further reinforcing the need for standardization.

Identification of missing values: Among four subsets of our dataset, only one has a missing value and we have decided to replace it with a mean value!

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 selected variables 10, 15, and 20 for visualization.

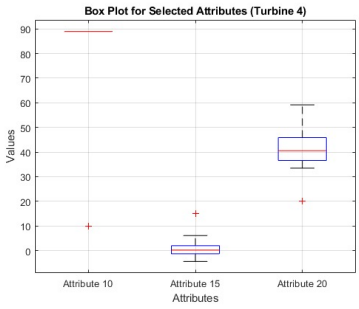

Figure 1: Box plot for Turbine 4

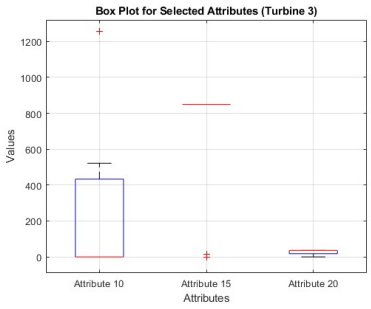

Figure 2: Box plot for Turbine 3

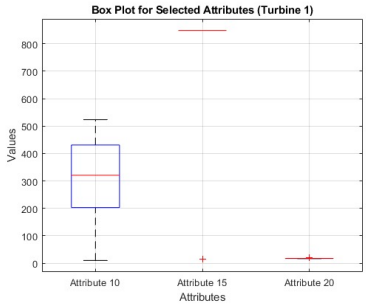

Figure 3: Box plot for Turbine 1

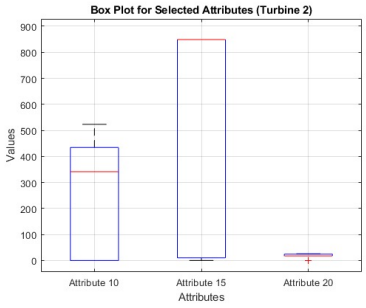

Figure 4: Box plot for turbine 2

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 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.

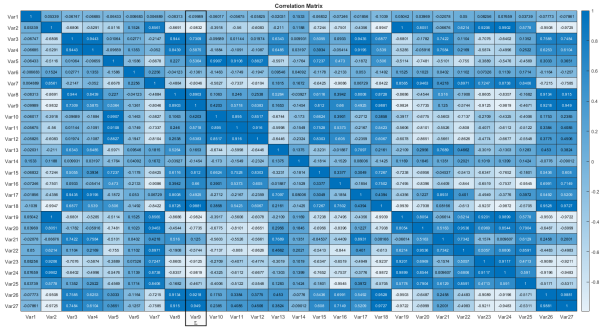

Figure 5: Correlation matrix for Turbine 1.

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

[1]

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