1 Introduction

We studied the Wind Turbine SCADA dataset for fault detection, ensuring correct import, inspection, and visualization. A GitHub repository (`Wind_Turbines_A4`) was used for version control and code sharing.

2 Data Import (0.25p)

Data was imported from Excel with observations as rows and variables as columns. WT2 (healthy) had 28 variables, WT39 and WT14 (faulty) had 27, while WT3 had 31 and was excluded due to inconsistency. We used WT2, WT39, and WT14 with 27 aligned variables.

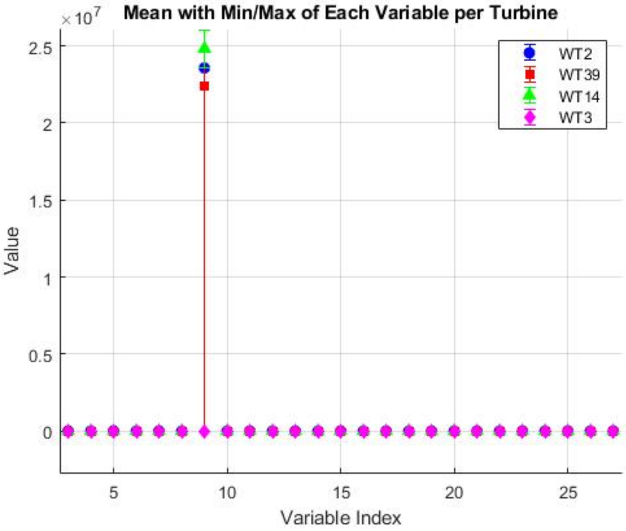


Figure 1: Comparison of variable ranges; WT3 differs strongly and was excluded.

The final dataset had 27 variables and 4359 records. Values ranged from -21.8 to 10^7, means were mostly $< 10^3$, and standard deviations varied widely, motivating standardization.

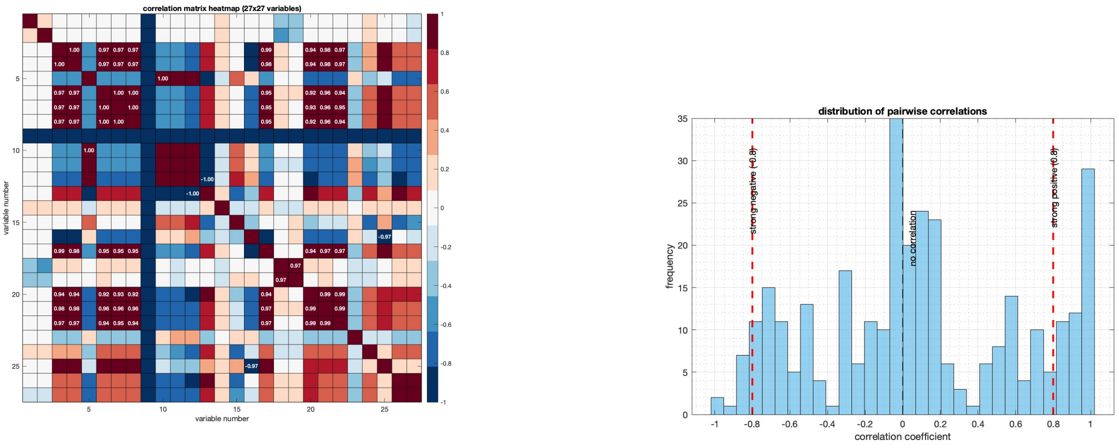
3 Challenges (0.5p)

When processing our wind turbine SCADA data, we were able to identify a number of critical data issues. The biggest problem was the extreme differences in scales of variables some of which were in the range of -0.565 to 23,325301, which gave a scale ratio of 31,800,502 between the highest and smallest standard deviation. Such excessive variability would predominate PCA outputs unless it is well standardized. Another issue was differences in the number of variables in the turbines, with the healthy turbine (No.2WT) having 28 variables and the faulty turbines having 27 variables, so we eliminated the extra column to ensure consistency. We also had 1 missing value in our 3,664 observations (0.03 missing rate) dataset which we had to remove. Another significant problem was the removal of redundant measurements by using perfect correlations (r = 1.000) of variables 6, 7, 8 and variables 5, 10 indicating the same physical phenomenon is being measured by these sensors and they may possibly be combined to eliminate redundancy.

- Inconsistent number of variables across turbines (WT3 excluded).
- One missing value in WT14 (row 358, col 9); removed, mean imputation also possible.
- Extra last column dropped from WT2.
- Variables unlabeled (Var1, Var2, . . .).
- No proper timestamp column, turbines not synchronized (10-second interval only stated in description).
- Strongly different scales across variables, requiring standardization.
- High correlation among variables, motivating dimensionality reduction.

4 Dataset Visualization (0.5p)

The dataset contained 3663 observations and 27 variables: 1571 healthy (42.9%), 687 and 1405 faulty (18.8%). Correlation analysis showed average $r = 0.44$ and 66 pairs above 0.8, including perfect correlations. Distributions varied across sensors, some skewed, confirming the need for standardization.



(a) Correlation matrix heatmap.

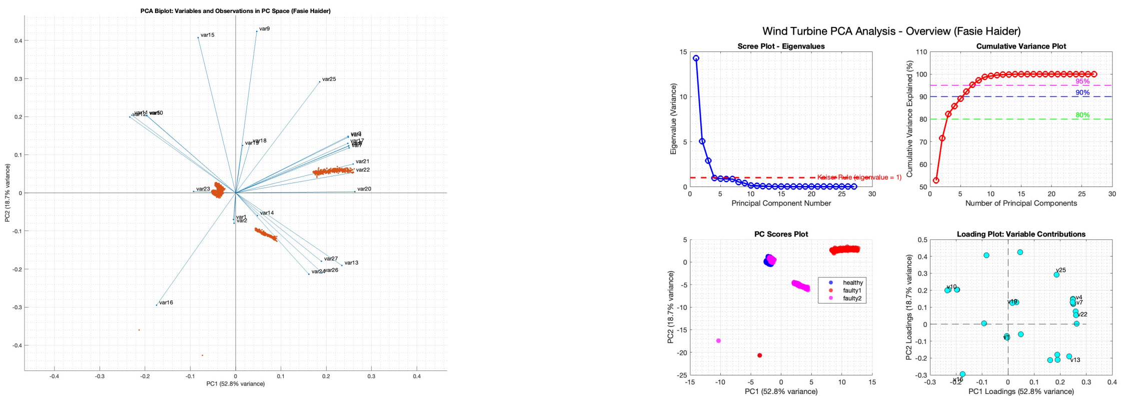
(b) Pairwise correlation distribution.

Figure 2: Correlation analysis of the 27 sensor variables.

5 PCA Exploratory Analysis (3p)

PCA reduced the 27D space to 3 main components explaining 82.2% of variance (PC1=52.8%, PC2=18.7%, PC3=10.7%). PC1 was driven by variables 20, 22, 21, 7, 8; PC2 by 9, 15, 16, 25, 24; and PC3 by 18, 19, 2. Healthy data showed low variability (PC1 mean -1.91 ± 0.20), while faulty turbine 1 was highly variable (PC1=4.52, SD=6.02), and turbine 2 was intermediate.

Group mean distances (6.42 and 1.83) provide strong discrimination. Scree and loading plots confirmed three key components.



(a) PCA biplot.

(b) PCA overview plots.

Figure 3: Principal Component Analysis results.

6 Pretreatment Plan (0.5p)

Preprocessing ensured reliable PCA: (1) standardization to zero mean/unit variance, (2) removal of one missing value (0.03%), (3) alignment of variables by dropping the extra column in WT2. The final matrix had rank 27, condition number 1.19×10^3, no severe multicollinearity. Although redundant variables exist, all were kept; future work may reduce them.