ADVANCED DATA ANALYSIS AND MACHINE LEARNING

NASA TURBOFAN LEVEL A2: PART 1

LUT University

2025

Group: Amin Hassanzadehmoghaddam, Antonio Oliva, Nico Niemelä

https://github.com/aminhm/NASA-Turbofan-A2.git

Content

1 Communication channel and code sharing strategy

The group will use Teams as a primary communication channel and GitHub will be used as a coding collaboration platform. Python was decided to be used as a tool to research the dataset.

2 Dataset description and identification of challenges of the data

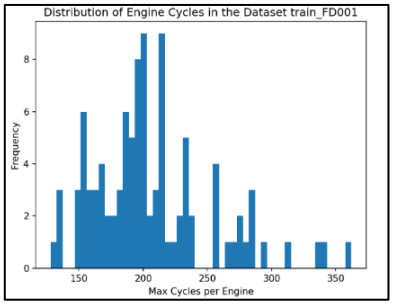

Figure 1. Distribution of engine cycles in the dataset train_FD001.

The dataset used in this study is the NASA Turbofan Jet Engine Degradation Dataset. The dataset consists of four different datasets: FD001, FD002, FD003 and FD004. Datasets contain multiple multivariate time series, each representing the operational history of a different engine.

Each engine has unequal amount of data. For example, distribution of maximum engine cycles in the dataset train_FD001 is shown in figure 1.

There seems to be no missing or extra values in the data, however datasets FD001 and FD003 have some columns with constant values. Physical meaning for the operational settings and sensor measurements are not specified.

3 Visualization and comment on the dataset

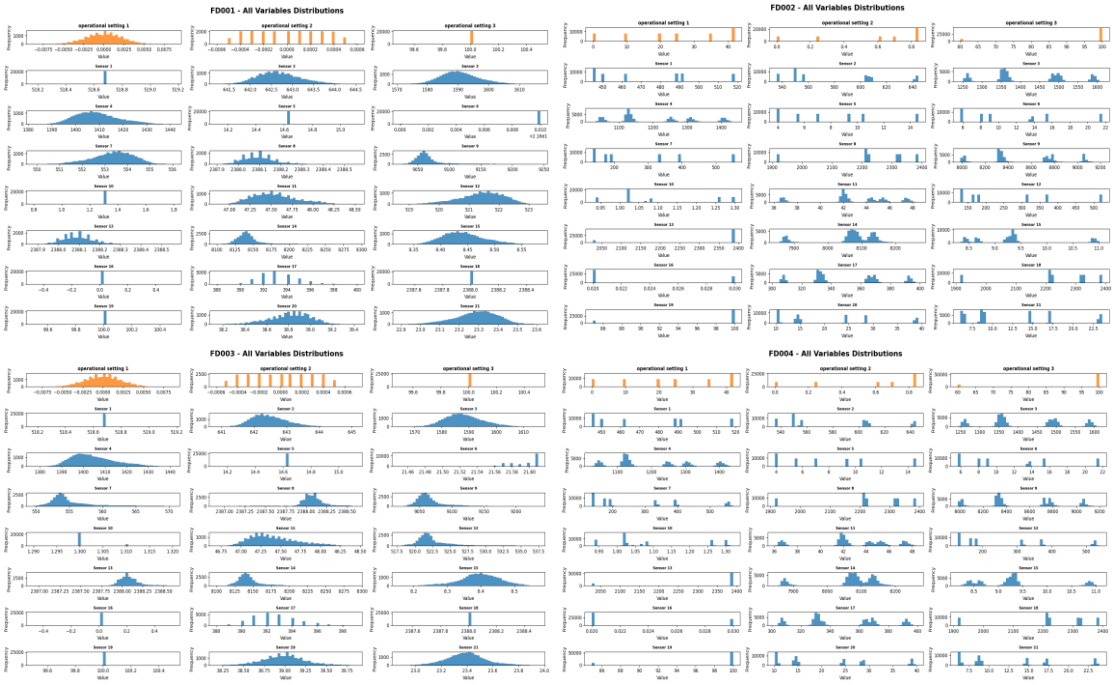

Figure 2. Distribution of variables in all datasets

The training datasets consist of 26 variables including unit number, time in cycles, three operational settings, and 21 sensor measurements. The number of observations is 20,631 for the FD001, 53759 for the FD002, 24720 for the FD003, and 61249 for the FD004. According to Figure 2, operational settings show distinct patterns. Setting 3 is nearly constant, setting 1 is normal in FD001 and FD003 but multimodal in FD002 and FD004, and setting 2 is multimodal depending on the system. Sensor measurements range from constant to normal and multimodal. They also reflect diverse engine states. As multivariate time series, the data capture each engine's evolution over time, and the 'time, in cycles' variable increases monotonic over the data, so it is a time-series problem.

4 Exploratory data analysis with PCA

Principal Component Analysis (PCA) was applied to the standardized matrixes of the datasets to analyse the correlations between sensors' data and retrieve the principal components that capture most of the variance in the dataset. In the scree plots in Figure 6, the first two components are the most impactful on the variance of the dataset, contributing to ~ 65% in FD001, ~ 67% in FD003 and ~ 97% in both FD002 and FD004.

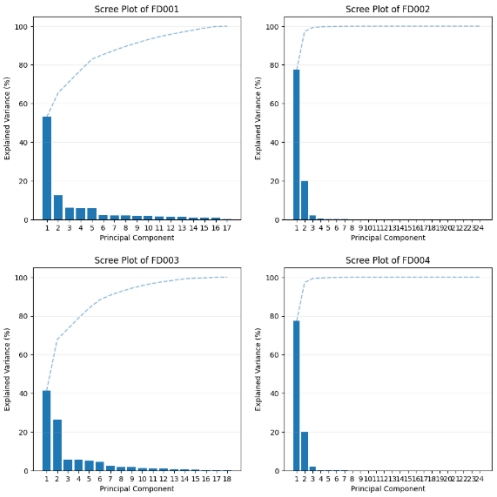

Figure 3. Scree plot of datasets with explained and cumulative variance per sensor.

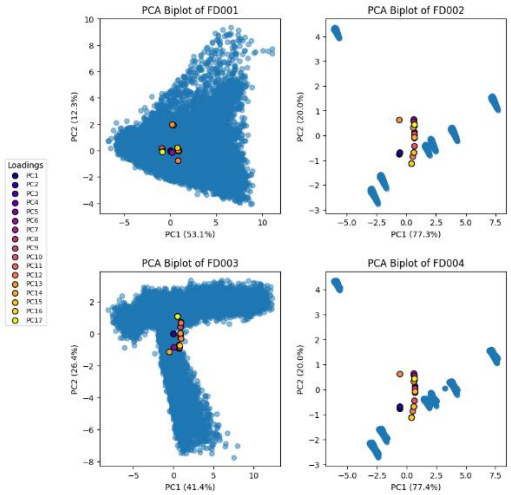

Figure 4. Biplots of data projected onto PC1 and PC2 and loadings for each principal component.

The loadings bars in Figure 8 show how each sensor relates to a principal component. Large positive or negative bars mean a strong link to that component. Bars close to zero indicate the sensor contributes little new information.

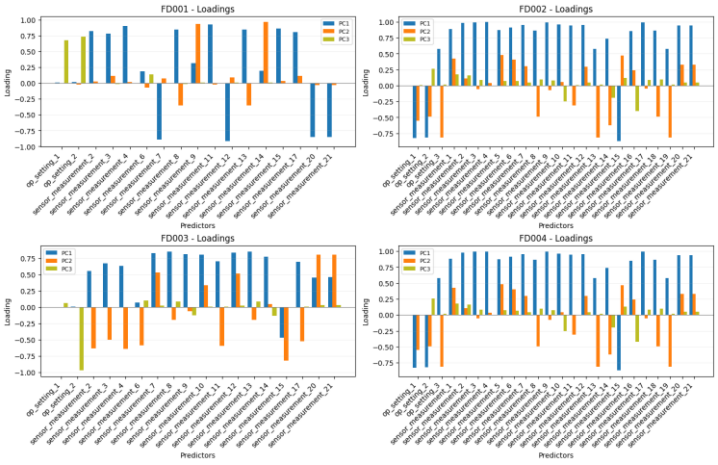

Figure 5. Loadings bar plots to analyse correlations between principal components and sensors

By looking at the plot, it is possible to determine that PC1 has a positive correlation with most sensors, while PC2 and PC3 focus attention on the differences between sensor groups. This shows that using the first three components is sufficient for further analysis.

5 Identification of the pretreatment steps

Pretreatment consists of removing constant or near-constant variables (e.g., operational setting 3, sensors 1, 5, 6, 10, 16, 18, 19 in FD001), handling missing values, and applying normalization to all variables (excluding removed ones). Outliers should be detected using methods like z-scores. Time-based features such as "time, in cycles" and "unit number" can be used to design lag and rollings.