Project work 1: NASA Turbofan Jet Engine A3

Alessia Tani, Iiro Vendelin, Sara Zambetti

September 14, 2025

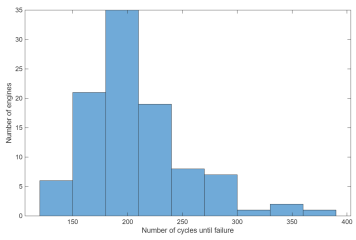
1 Communication and Code Sharing

For our group, we established a communication channel via Teams and share code using GitHub. All MATLAB scripts and figures are stored in a shared repository accessible to all group members.

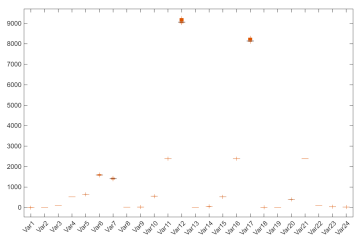
2 Data analysis

The analysis is based on the NASA C-MAPSS turbofan engine degradation datasets (FD001–FD004), which contain training and test records of engines monitored over multiple operational cycles. Each row represents one cycle of a specific engine, including operational settings and sensor measurements, while the labels provide the Remaining Useful Life (RUL).

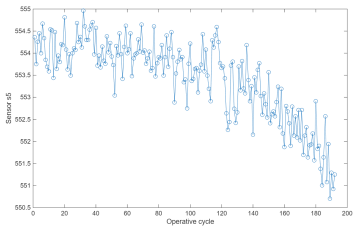
From the training sets, predictor matrices were extracted to structure the data with observations as rows and variables as columns. An exploratory analysis was then performed: the engine lifetime distributions (Figure 1.a) showed unsynchronized failure cycles between units, and the average engine life was calculated for each data set. Data quality checks confirmed the absence of missing values, while constant variables were identified in the FD001 and FD003 data set.
To better understand the data, boxplots (Figure 1b) were plotted, highlighting outliers and variability between measurements. The time series plots for the selected sensors (Figure 1c) in individual engines confirm that the measurements are organized as time series per engine.
In general, the data sets contain a large number of observations and 24 predictor variables. The exploratory analysis confirms the necessity of preprocessing, including normalization, outlier handling, and possibly dimensionality reduction, to prepare the data for modeling RUL.



(a) Distribution of the engine's life



(b) Boxplot

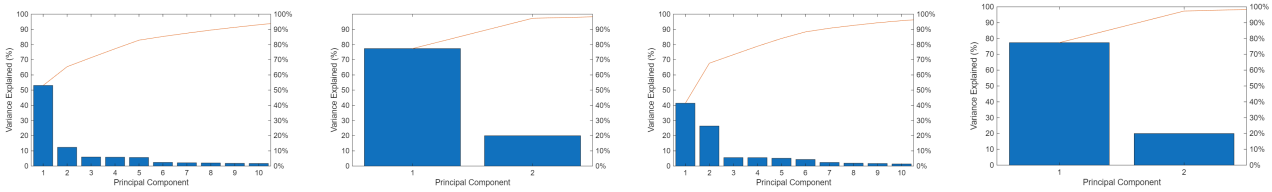


(c) Time series plot for selected sensor

Figure 1: FD001

3 PCA Analysis

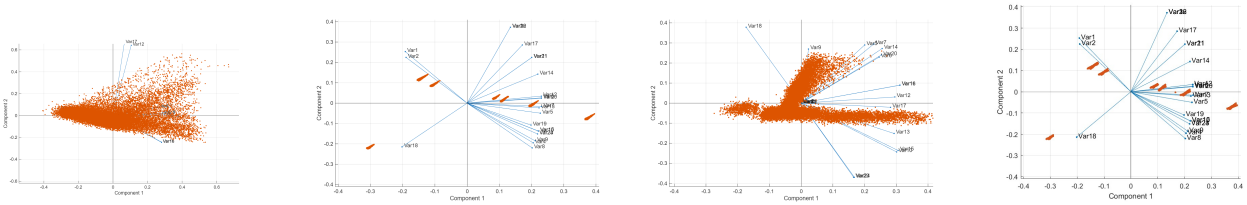
Principal Component Analysis (PCA) reduces data dimensionality by transforming correlated variables into uncorrelated principal components. These are defined by eigenvectors, which capture directions of maximum variance and with eigenvalues indicating the variance explained. Plots were used to visualize how this variance is distributed between components.
The FD001 scree plot (Figure 2.a) shows a gradual decline in explained variance, with the first component capturing just over half of the total variance, indicating complex, multi-dimensional sensor interactions requiring many components to preserve most information.
In contrast, FD002 and FD004 (Figures 2.b and 2.d) show very similar patterns: the first two principal components alone capture nearly all of the variance. This indicates more constrained operational conditions where sensor readings behave in predictable ways.
FD003 (Figure 2.c) falls in between. Its variance is more evenly distributed across the first few components, but still more concentrated than in FD001. The first 10 components capture almost all of the variance.



(a) FD001 (b) FD002 (c) FD003 (d) FD004

Figure 2: Scree plot of the four datasets.

The biplots reveal distinct clustering behaviors across the data sets. In FD002 and FD004, the variable vectors cluster tightly along the first principal component, indicating strong correlations and redundancy between sensors. In contrast, FD001 and FD003 show more dispersed patterns, with several variables pointing in different directions, suggesting more complex interactions and the need for additional components. In all cases, some sensors consistently align, confirming subgroup correlations. PCA effectively compresses this information, reducing dimensionality while retaining the main variance structure.



(a) FD001 (b) FD002 (c) FD003 (d) FD004

Figure 3: PCA Biplots - First two PCs.

4 Pretreatment plan

Based on the PCA results, retaining the first 5–10 principal components, depending on the dataset, appears sufficient to reduce redundancy while preserving the essential information. Strongly correlated sensors can be effectively compressed through PCA, further reducing dimensionality. Variables with constant values should be removed, as they provide no useful information for modeling and may introduce numerical issues. In addition, the removal of extreme outliers is recommended, since they can distort the principal component directions.