Capstone Project
Turning and tension process big data analysis from high-frequency sensors

Project definition

Project Overview

In the last 20 years, the industry has improved predictive maintenance processes and the tools used to perform it. In a fully developed factory, the high frequency sensors are applied to the manufacturing process. In particular, vibration analysis technology has evolved to unimaginable levels. To ensure the turning process through the best digital solution is my job. Turning process is an essential process in the manufacturing. The goal is to monitor the whole process to ensure the product quality by the sensors and sampling. Another goal is to model the tool life. The prediction of tool life can effectively avoid part quality problems caused by abnormal tool status and improve tool utilization rate. In this project, the data from the vibration sensors, tension sensors database and MES system are collected. The three data sources are combined through the time. The vibration sensors' dataset are x, y, z acceleration speed at 2000 HZ with its own product unique ID and time. The tension collects voltage and dynamic tension at 300 HZ with its own product unique ID and time. The MES system collects the time and product unique ID and the time we switched the knife of turning.

Project Statement

The tool wear process is complicated and the residual life of tool is difficult to be predicted accurately by the influence of working conditions. There are two important objectives for the turning process: (1) To model the tool life. (2) To model the roughness of the turning product. It is relying on the feature selection. For the tension process, we are focused on (1) the relationship between tension and voltage (2) the balance between the left and right tension.

We don't have any labels in the project. We need to define the goals of the project. I would go for the regression model with lots of features constructed through the time-series. I assume we can model the relationship between the tension and voltage through the data. We can also capture the abnormality (separation) through the analysis of big data by time series. I would put two months of data in the training dataset and the other dataset as a test dataset.

Metrics

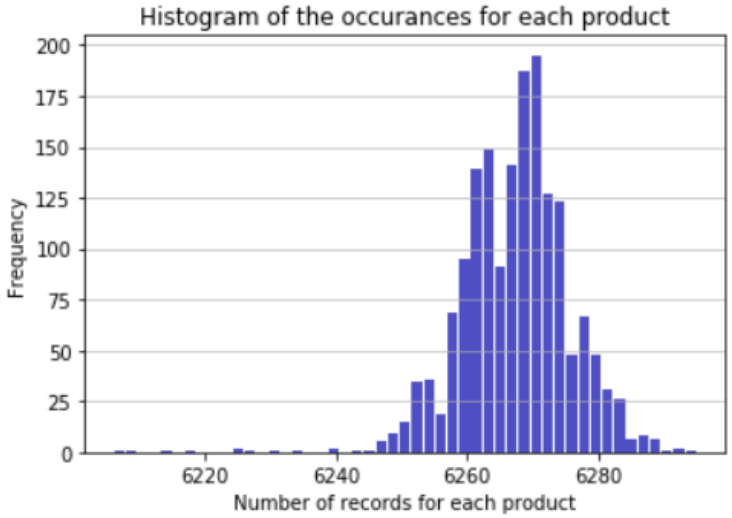
Both of the two problems are regression problems. We are going to use the R squared as the basic metrics to measure the performance of the model.

Analysis

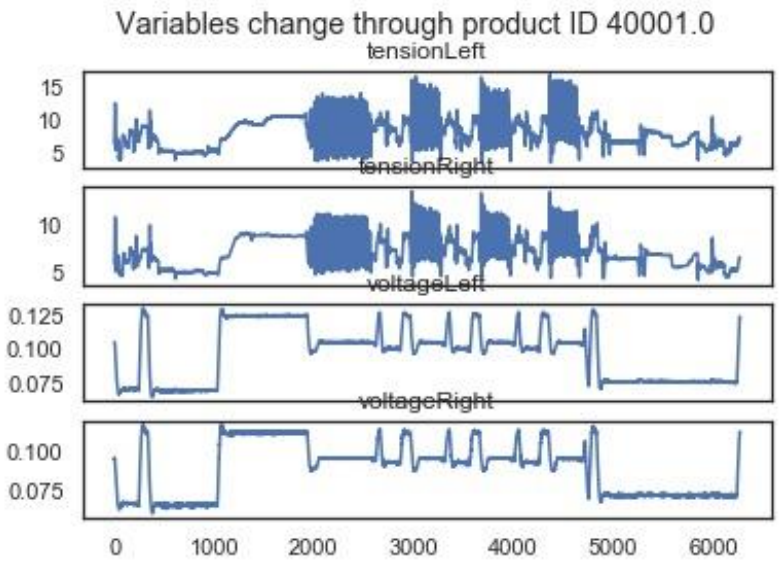
Data Exploration

For the tension problem, it consists of 11962005 rows and 8 columns. The data records the tension/voltage change from 6/28 0:00 to 6/28 18:00. There are 1853 products in the dataset. There are still some mislabeled products. For better separations of the time-series data, we only use the data corresponding to 6200-6300 rows. This would be 1692 products (~91% of the original dataset). There is no missing value in the case.
Tension dataset: one column is string (time), id, cur_right, cur_left, tension_right, tension_left are numeric variables.
Turning dataset: one column is string (time), id, x, y ,z are numeric variables.
MES dataset: one column is string (time), knifetime, id are numeric variables.

Data visualization



[Fig] The number of records for each product in the tension problem after removing outliers in the tension dataset



[Fig] The relationship between tension and voltage across one product in the tension dataset

Methodology

Data preprocessing

For the tension sensor's dataset, there are more or less rows corresponding to one product. We have to remove the outliers. The sensor for the tension is around 300 HZ. The sensor for the turning is around 2000 HZ.

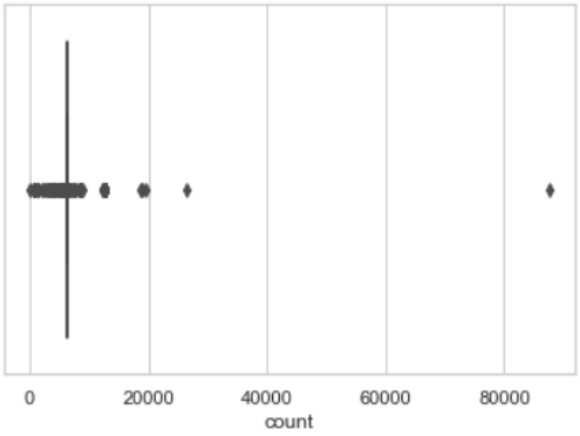
For each product, we calculate the mean, median, peak, root mean square, peak factor, form factor and so on. We also separate the data by the process itself.

For the turning project, original May data consists of 11962005 rows. Original June data consists of 116119892 rows. Original July data consists of 134806335 rows. For each row, we have the time stamp and the vibration (accelerations) of x, y, z-axis. It is 2000 HZ. For each product, we stored data around 6800 rows. We also remove the outliers corresponding to much more than or less than 6800 rows.

Step 1. Remove the outliers: if a certain product corresponds to too many (small) rows, we remove it.

```
]: freqtable=my_tab[(my_tab['count']<6300)&(my_tab['count']>6200)]
```

```
import seaborn as sns
sns.set(style="whitegrid")
ax = sns.boxplot(x=my_tab['count'])
```



Step2. Get the features for each product. Se calculate the mean, median, peak, root mean square, peak factor, form factor and so on.

Step 3: Generate a new column as the use times of the knife in the vibration dataset. To match the time of MES and the sensors'data.

Turning data mining project
Skewness (Sk, 偏度)

❑Interpretation of *Sk*

- ▶ It is a measure of symmetrical attributes.(对称程度)
- ▶ *Sk* > 0, positively skewed (正偏态/右偏，长尾在右)
- ▶ *Sk* = 0, symmetrical(正态，对称)
- ▶ *Sk* < 0, negatively skewed(负偏态/左偏，长尾在左)

❑Formula of *Sk*

$$Sk = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{\frac{n}{S^3}}$$

Negatively Skewed | Symmetric (Not Skewed) | Positively Skewed

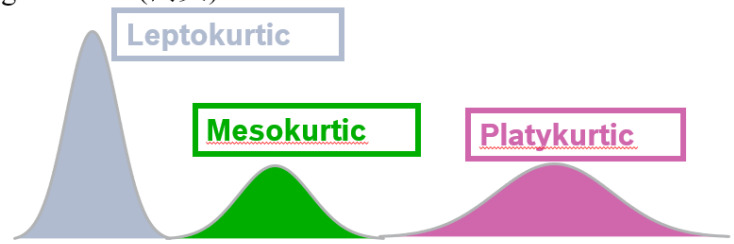
Turning data mining project
Kurtosis(Kur, 峰度)

❑Interpretation of *Kur*

- It is a measure of Peakedness .(峰部的尖度)
- *Kur* > 0, platykurtic(厚尾), flat and spread out(矮胖)
- *Kur* = 0, mesokurtic(正态), normal in shape
- *Kur* < 0, leptokurtic(瘦尾), high and thin(高尖)

❑Formula of *Kur*

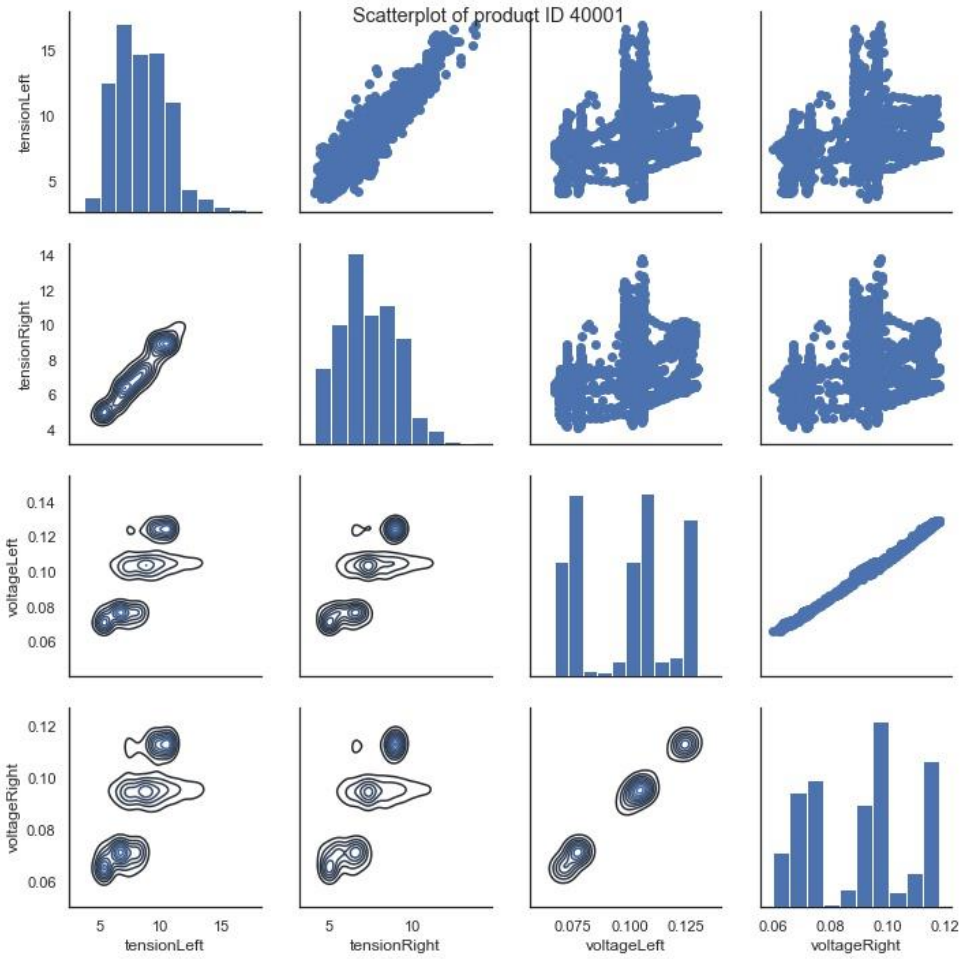
$$Sk = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{\frac{n}{S^4}} - 3$$



Turning data mining project
RMS and standard deviation

- RMS: The RMS (root mean square) value is generally the most useful because it is directly related to the energy content of the vibration profile and thus the destructive capability of the vibration. RMS also takes into account the time history of the wave form.
- In the case of a set of *n* values $\{x_1, x_2, \dots, x_n\}$, the RMS is
- $x_{RMS} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \cdots + x_n^2)}$
- Standard deviation: the standard deviation is a measure of the amount of variation or dispersion of a set of values.
- $s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$

Implementation

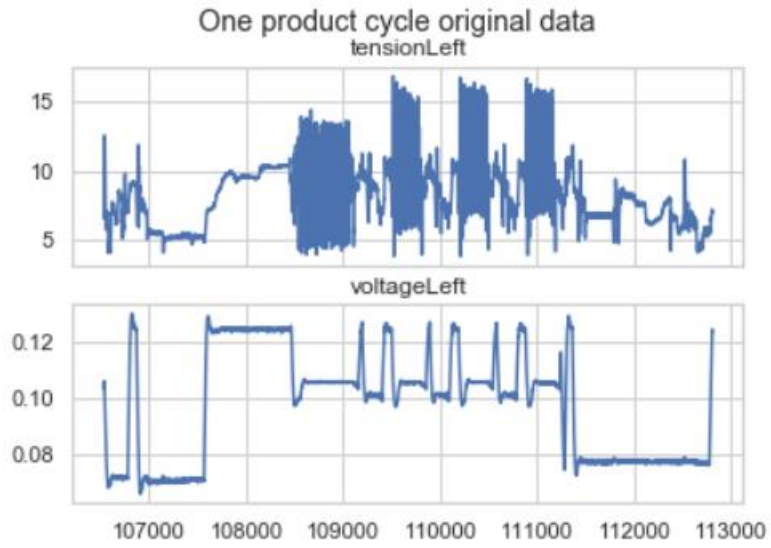


[Fig] The scatterplot, the density plot and histogram for four different variables in the tension sensor

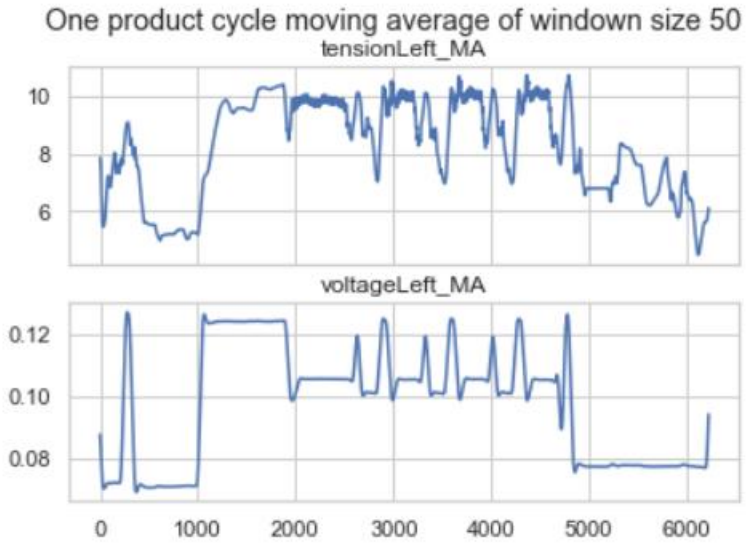
From one product, we can explore the tension and voltage relationship. We can see from this scatterplot more clearly: when the voltage is low, the corresponding voltage is relatively low. Voltage left and right are linear. Tension left and right are also very close to linear. Voltage are tri-modal while the tension is more normal distributed.

Refinement

In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full dataset. It will make the trend more smooth. It is a common way in the time-series analysis to get rid of the white noise.



[Fig] The original product cycle data for the tension problem.



[Fig] The moving average of product cycle data for the tension problem.

Original data: Tension = 1.57 + 69.86 * Voltage

Moving average data: Tension = 1.34 + 72.28 * Voltage

The model is significant better in the moving average data.The correlation increases from 0.64 to 0.82. We can also confirm this point by the OLS output. This is in the 20200628_tension_data_analysis Notebook in [67]

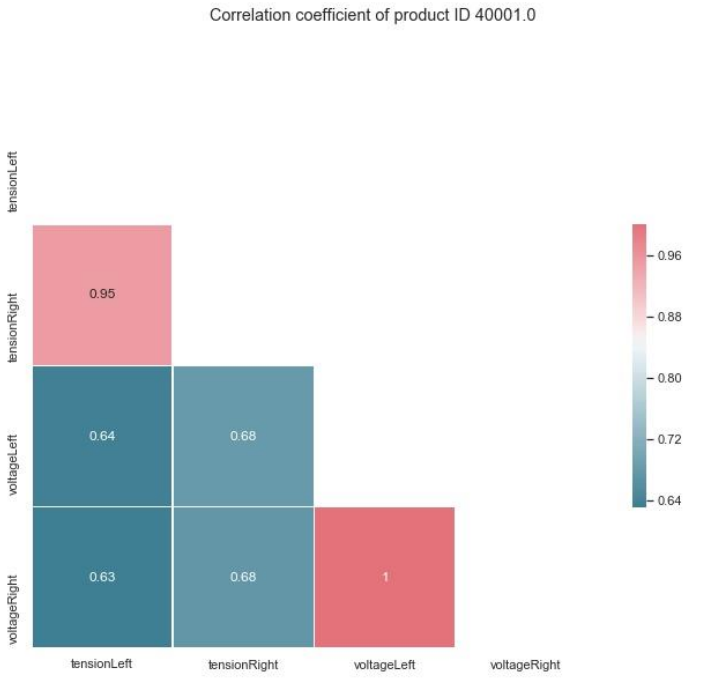
In [67]:
```
import statsmodels.api as sm
X = voltageleft_ma
y = tensionleft_ma
X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.677
Model:                            OLS   Adj. R-squared:                  0.677
Method:                 Least Squares   F-statistic:                 1.302e+04
Date:                Thu, 02 Jul 2020   Prob (F-statistic):               0.00
Time:                        13:34:32   Log-Likelihood:                -8479.2
No. Observations:                6217   AIC:                         1.696e+04
Df Residuals:                    6215   BIC:                         1.698e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3356      0.063     21.206      0.000       1.212       1.459
x1            72.2848      0.633    114.116      0.000      71.043      73.527
==============================================================================
Omnibus:                      230.778   Durbin-Watson:                   0.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              256.252
Skew:                          -0.493   Prob(JB):                     2.27e-56
Kurtosis:                       3.126   Cond. No.                         53.3
==============================================================================
```

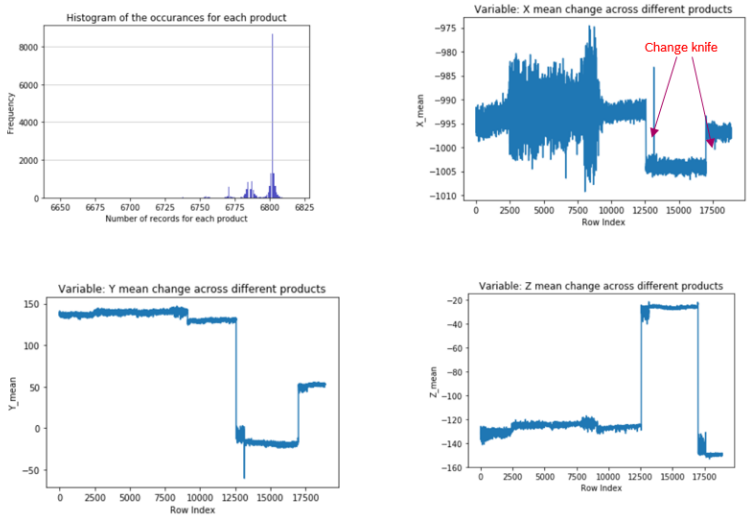
[Fig] The OLS regression output for the tension dataset

Results

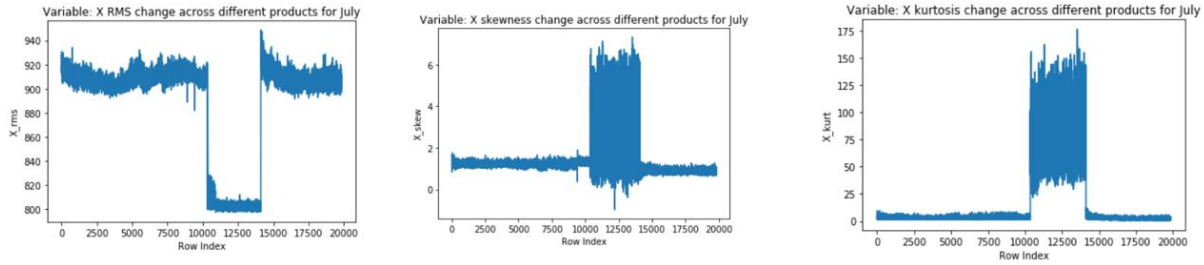
Model evaluation and validation



[Fig] The correlation between four different variables in the tension dataset



[Fig] The features' change in the turning dataset in May



[Fig] The features' change in the turning dataset in July

From the model, we can capture the trends through different variables.

Justification

The exploratory plots are similar when we analyze the data before/after change. The model is about the voltage left and voltage right. The differences between the left and rightVoltage is almost certain across different time.

Voltage left = 1.15*voltage right

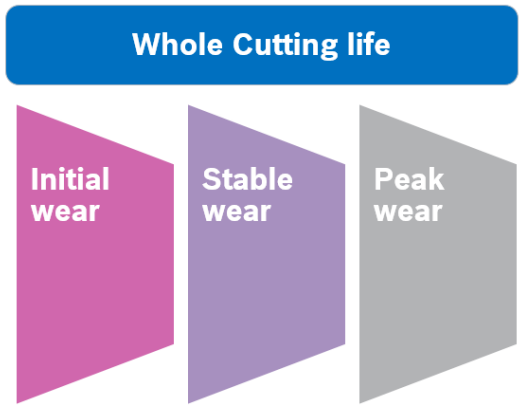
Voltage left = 1.13*voltage right

We slightly adjust the equipment to get less variance between left and right.

Conclusion

Reflection

Increase the manufacturing efficiency by monitoring machines, processes and environmental conditions. The turning sensor is a compact multi-sensor device for harsh environments. Machine condition tracking enables predictive and remote maintenance to save costs. With sensors, production yields can be optimized via live process monitoring. Due to its motion and environmental sensing abilities, the sensors' big data analysis is ideally suited for I4.0 applications.



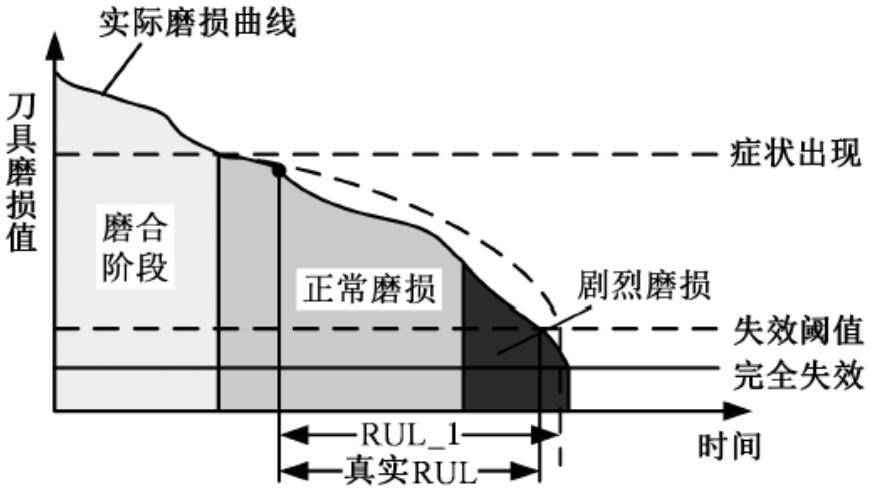
Improvement

We are able to draw a life curve for the turning tools. We also adjust the machine according to our big data analysis results.

The steps of the solution:

- 1. Correlation, PCA, clustering analysis to find the important variables and get rid of the outliers.
- 2. The exploratory data analysis to identify the trends of the important variables.
- 3. Collect the data from the newly recommended life time line. Verify the process by processing engineers.

The most challenging parts is (1) the feature selections part, we have to generate enough features (in time-domain and frequency-domain) to capture the separation; (2) The understanding of the process: we have to work with the process engineers.


实际磨损曲线
刀具磨损值
症状出现
磨合阶段
正常磨损
剧烈磨损
失效阈值
完全失效
RUL_1
真实RUL
时间