Predictive Fatigue Analysis on Wagon Drawgear by Position in Train

Congyu Liu

R. Nazim Khan
Mathematics and Statistics
The University of Western Australia

John McArthur
CEED Client: Rio Tinto Iron Ore

Abstract

In the Pilbara region of Western Australia, RTIO runs the world's largest iron ore operations, including a world-class, integrated network of 16 iron ore mines, four port facilities, and a 1,700km rail network. The company has a fleet of 4,500 ore car pairs which transport the iron ore through rail network from Pilbara to ports in Dampier and Cape Lambert. Coupling components (including couplers, knuckles, yokes, drawbars and draft gear) connect ore cars to each other. This study is conducted on behalf of Rio Tinto Iron Ore (RTIO) and apply machine learning technology on data of train operation to predict failure of coupling component that causes train separation. The principal objective of the project is to train statistical models to coupling component failure data based on the input variables related to the wagon position. The data from several different departments need to be merged for this project. Results obtained thus far show that generous fit of the model is achieved. We recommend other variables for which data should also be obtained, which will further improve the model fit.

1. Introduction

Rio Tinto Iron Ore (RTIO) runs one of the largest iron ore mining operations worldwide. The mine sites are located in the Pilbara region of Western Australia. Ore from the mines is transported to port facilities by train. The ore cars are connected to each other using coupling components. These components experience coupling force during loading and unloading, leading to fatigue failure. For safety and continuity of operations, it is essential to avoid failure.

Figure 1 shows the ore cars that RTIO currently uses in their operation. The coupling components can be seen between the cars and connects two ore cars working as a pair. The coupling components mostly used by RTIO is the F type coupler system shown in Figure 2. Failures in knuckles and yokes are our priority in the project because fatigue forces most likely cause them. The current maintenance strategy for coupling components is a combination of non-destructive testing result and replace components on age. Based on a report from rolling stock management division, most failures of knuckles occur in the second year, resulting in train separation.

Fatigue force is the leading cause of coupling component failures. The force experienced by coupling components depends on their position in the train of the ore cars they connect. A train may be up to 2.4 km long with 240 ore cars attached, each with a load capacity of 120 tonnes.

The forces in the coupling components towards the end of the train are very different from those at the beginning. For example, an ore car connected to the locomotive would be in position one and the force suffered from this ore car is different from an ore car in the middle part of the train. This is because tensile and compressive forces experienced in the front part of the train are different from the forces in the middle section. This results in components having different lifetimes. Consequently, the replacement times can be optimised to provide maximum useful life for the components.

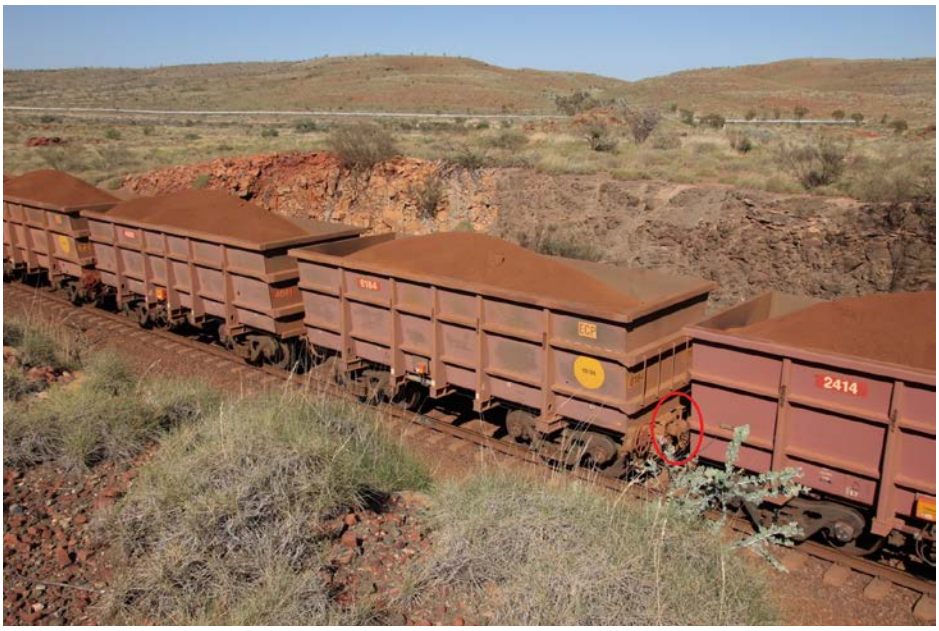


Figure 1 Ore Car

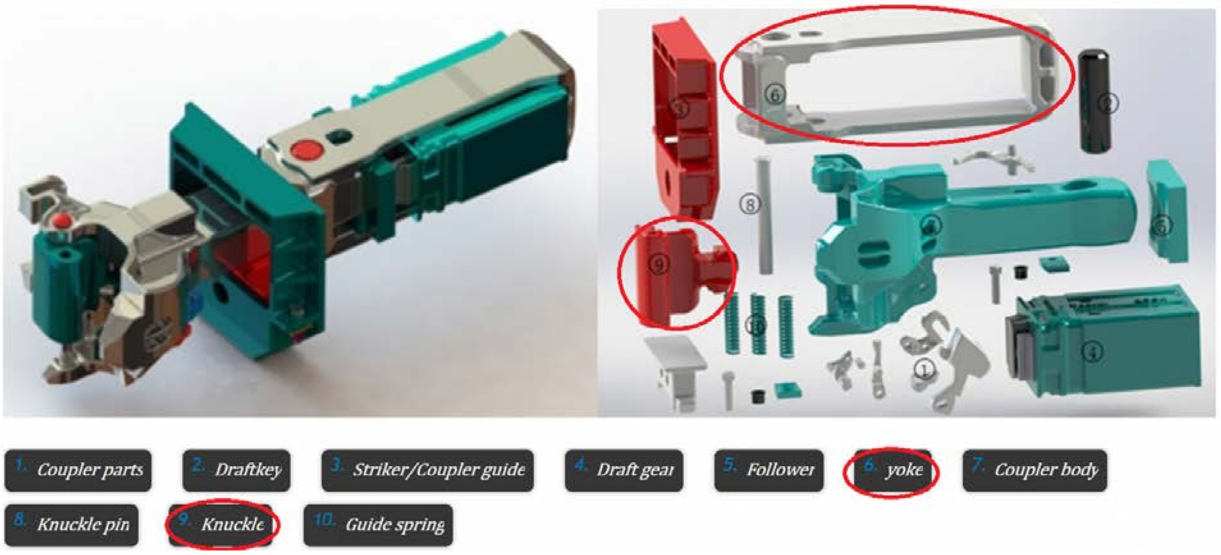


Figure 2 F Type Coupler

RTIO currently works with the Institute of Railway Technology (IRT) at Monash University to analyse fatigue damage of coupling components across their operation. The methodology used is rain-flow analysis to count the number of loading and unloading cycles. The power of three then weights these cycles on all ore cars based on the AS 4100-1998 for steel structures for loading up to five million cycles（AS, 1998). The power three weighting means that if the force is doubled then the damage index increases by a factor of 2^3 =8. Thus a 100-tonne force would contribute eight times more to the damage index than a 50-tonne force.

In a study from IRT（Bowey, 2018), power three weightings were applied on all cycles of ore cars. However, given that the force on ore cars depends on their positions, it is more realistic to use different power indices according to the position of ore cars on the train. Moreover, the prediction of life of coupling components is of paramount interest to the client; the current maintenance strategy leads to failure more often in the second year, and the current replacement plan can be optimised to reduce the costs. The objective of this project is to apply machine learning technology to optimise the maintenance strategy.

2. Methodology

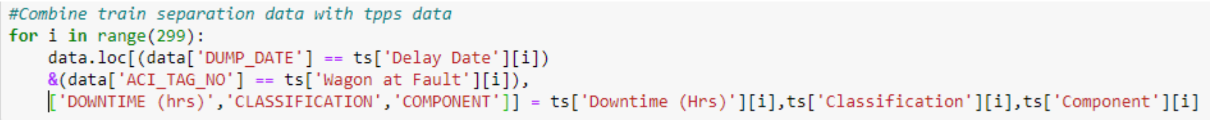
2.1 Data preparation

The data for the project come from three sources: train separation summary table: instrumented ore cars (IOCs) data from IRT; and real-time tracking data from Rio Tinto operation center. The first step is to combine the dataset from these departments and form a large dataset to be used to train the model. We found that the data from IOCs are not suitable to merge with the other two datasets, because there are only a few IOCs in a train of 240 ore cars, and it is impossible to match the data with real-time tracking data that have millions records available.

The real-time tracking data can be extracted by using SQL query from databases of the operation center. The data from January 2017 to March 2019 have been obtained, and the corresponding train separation record can be merged by matching dumping data and wagon ID. The combined dataset contains approximately 8 million transactions.

2.2 Data cleaning

Data cleaning is critical for any modelling. The data contains a large number of null values that correspond to missing values. These empty slots are filled based on their column. For example, in the column ‘Downtime caused by train separation’, the mean value of downtime in hours is calculated and then loaded into the dataset. For columns such as ‘Classification’ and ‘Breaking component’, null values are relabelled into the ‘Unknown’ category.
It is hard to deal with the unfixable data in the train separation table, and some wagon ID is ambiguous that are not able to be replaced. By using the code of data integration shown in Figure 3 (Jupyter Notebook, 2019), around 100 records of train separation loss.



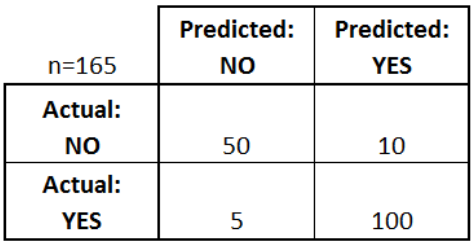
```
#Combine train separation data with tpps data
for i in range(299):
    data.loc[(data['DUMP_DATE'] == ts['Delay Date'][i])
    &(data['ACI_TAG_NO'] == ts['Wagon at Fault'][i]),
    ['DOWNTIME (hrs)','CLASSIFICATION','COMPONENT']] = ts['Downtime (Hrs)'][i],ts['Classification'][i],ts['Component'][i]
```

Figure 3 Code for gathering the data in Python

2.3 Model selection

Machine Learning techniques can be classified according to the type of supervision they get during training. Three categories are considered in this project: supervised learning, unsupervised learning and semisupervised learning. In supervised learning, a labelled training set is one that contains the desired solution (a.k.a. label) for each observation. The two most common supervised tasks are regression and classification. Based on the combined dataset, the task we tackle is to predict if there will be a train separation during the operation so that the classification is suitable to analyse the dataset.

Two types of classification are conducted in this project: the first model is to predict whether the failure will happen, and the second model is to predict the type of broken coupling components such as knuckles or yokes. The binary classification algorithm is applied to the first model and multi-class classification suits for the second. One way to evaluate the performance of a classifier is to look at the confusion matrix. Figure 4 shows a sample confusion matrix.



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Figure 4 Sample confusion matrix

3. Results and Discussion

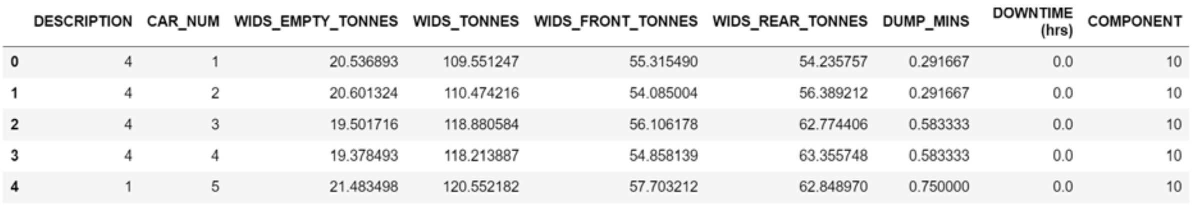
3.1 Software package used

Oracle SQL developer is applied to explore the database from Rio Tinto operation center (SQL Developer, 2019). Jupyter notebook is a web application that it is one of the most popular tools to do data analytics (Jupyter Notebook, 2019). It is used to do the rest part the project including data cleaning, data integration, model training and visualisation.

3.2 Combined datasets

The current results are derived from analysing train separation record and real-time tracking data from January 2017 to March 2019. Training more data to get better performance of the model is possible but time-consuming would be the stepping stone for the project. There are approximately 8 million transactions during this period, and every training attempt took many hours to come alive.

A sample of the data is provided in Figure 4. Some attributes such as dump data, wagon size and tag number have been dropped because they are not suitable to encode as an integer, which is the requirement for training a classifier. Attributes like description and broken components can be encoded. For example, Q series wagon can be set to class 1, and S series corresponds to class 4. Similarly, breaking knuckle is equivalent to class 10 and Retainer Plate Bolts can be seen as class 9.



	DESCRIPTION	CAR_NUM	WIDS_EMPTY_TONNES	WIDS_TONNES	WIDS_FRONT_TONNES	WIDS_REAR_TONNES	DUMP_MINS	DOWNTIME (hrs)	COMPONENT
0	4	1	20.536893	109.551247	55.315490	54.235757	0.291667	0.0	10
1	4	2	20.601324	110.474216	54.085004	56.389212	0.291667	0.0	10
2	4	3	19.501716	118.880584	56.106178	62.774406	0.583333	0.0	10
3	4	4	19.378493	118.213887	54.858139	63.355748	0.583333	0.0	10
4	1	5	21.483498	120.552182	57.703212	62.848970	0.750000	0.0	10

Figure 5 First 5 transactions of the dataset

3.3 Model performance

Three models used to predict whether train failure will happen is built. Stochastic Gradient Descent (SGD) classifier is a linear Support Vector Machine (SVM) classifier that is simple yet very efficient to learn large datasets. Although linear SVM classifier is fast to train and work surprisingly well in some cases, many datasets are not close to being linearly separable. Nonlinear SVM classifier with a polynomial kernel is applied to be the second model. Random Forest classifier is an ensemble of decision trees, and the algorithm introduces extra randomness when growing trees. It is added to be the third model to make a comparison.

The result shows that all models achieve 99% accuracy based on the test set. This result can not be fully trusted because there are over millions of transactions which are marked as no failure versus only a few hundred labelled as train separation. Under this circumstance, the confusion matrix is the better way to compare performance from all models.

The confusion matrix in Table 1 shows that the SVM classifier has the best performance over three classifiers with 15 correct predictions on train separation cases in the test set.

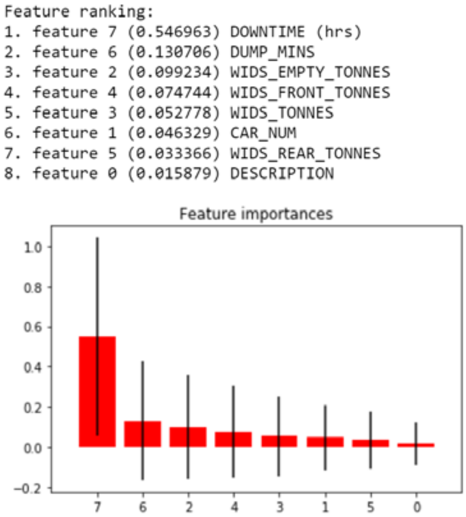
Model	True Positive	False Positive	False Negative	True Negative
SGD classifer	1539434	0	16	3
SVM classifer	1539428	6	4	15
RandomForest classifier	1539434	0	19	0

Table 1 Confusion matrix for all classifiers

It is essential to analyse feature importances. Figure 6 indicates that the attribute of downtime in hours is the most vital of all features in the Random Forest classifier. Based on Table 1, we can see that Random Forest classifier has the worst performance in three classifiers. Feature importance in the SVM classifier should be more reasonable. However, the complexity of calculating feature importance in SVM classifier is exceptionally time-consuming.

4. Conclusions and Future Work

At present, the results show that machine learning techniques can be an auxiliary tool in the rail industry. However, the accuracy of prediction still needs to be optimised. Ideally, car position in the train consist should be the essential feature affecting train separation, because the force experienced by coupling components depends on the position in the train of the ore cars they connect. Based on my result, the features of downtime and dumping minutes are more import than car number in train consist.



```
Feature ranking:
1. feature 7 (0.546963) DOWNTIME (hrs)
2. feature 6 (0.130706) DUMP_MINS
3. feature 2 (0.099234) WIDS_EMPTY_TONNES
4. feature 4 (0.074744) WIDS_FRONT_TONNES
5. feature 3 (0.052778) WIDS_TONNES
6. feature 1 (0.046329) CAR_NUM
7. feature 5 (0.033366) WIDS_REAR_TONNES
8. feature 0 (0.015879) DESCRIPTION
```

Figure 6 Feature importances in random forest classifier

Ongoing work focuses on using unsupervised learning and semisupervised learning techniques to achieve a better performance of the model. The unsupervised learning method, clustering, can be applied to the unlabelled data to find similarities in these transactions, which in turn segment car numbers into clusters. Semisupervised learning can deal with partially labelled training data, usually a large amount of unlabeled data and a small amount of labelled data. The concept of this technique perfectly fits our dataset.

Future work includes checking if non-destructive testing data can merge into the dataset to provide more attributes. More input features will lead to better model performance and eventually construct an application that outputs accurate prediction of train separation possibility.

5. Acknowledgements

I would like to thank John McArthur for the opportunity to work on such a stimulating project. To Hayden Huxtable, thank you for all the guidance on collecting the tpps data. To my supervisor, Dr Nazim Khan, thank you for all inspiring ideas and helps on editing all paperwork. Lastly, thanks for the staff at CEED office, Dr Jeremy Leggoe and Amanda Bolt for supporting the project.

6. References

Australian Standards (1998), AS 4100-1998 Steel Structure

Bowey R, Tan M, Chevin J (2018), Review of draw gear REPOS tables and fatigue damage across Rio Tinto Iron Ore’s operation

Jupyter Notebook (2019), Anaconda

Oracle SQL Developer (2019), Oracle Corporation