

Relationship Between Track Geometry and Track Components' Health Conditions

This manuscript ([permalink](#)) was automatically generated from [uiced/cee-492-term-project-fall-2022-shans@7252118](#) on December 2, 2022.

Authors

- **Arthur Bilheri**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [artbil94](#) ·  [I dont like birds](#)

Hogwarts school of witchcraft

- **Negin Shafie**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [negins2](#) ·  [NA](#)

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign · Funded by none

- **Shirin Qiam**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [Shirin-qiam](#) ·  [NA](#)

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign · Funded by none

- **Sadaf Shafie**

 [XXXX-XXXX-XXXX-XXXX](#) ·  [sadafs2](#) ·  [NA](#)

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign · Funded by none

Introduction

Railroad tracks have four main components: rail, ties, fastening systems, and ballast. The rail is the main component that supports the loads transmitted by the trains. The fastening system is responsible for fastening the rail to the ties and keeping the rail in a proper position. Fastening systems are made of spikes, tie plates, anchors, and sometimes clips. The ties are the interface between the rails and ballast. The two main materials used for ties in the United States are timber and concrete. The last layer on the railroad superstructure is the ballast. Its main functions are to spread the loads on the ground and to provide proper drainage for the track system.

These components are inspected using the LRAIL technology that combines 2D imagery and laser triangulation to assess the health and condition of each component. The collected data is processed through a DNN (Deep Convolutional Neural Network) model that identifies and classifies the components. This technology has been used under the scope of research led by the RailTEC group at UIUC, funded by the FRA-DOT with two Class I railroads in the US. The output of this technology, which will be used in this project, is described in an Excel file (filename: Datasets) attached to this proposal and Table 1.

Track geometry can be defined as the relative position of the rails. The common measurements are related to horizontal and vertical irregularities, gage, and superelevation. CFR 213 establishes safety limits that must be kept in order to provide the trains with a safe ride. Railroads use laser-based contactless systems to measure track geometry. Recently these systems have been installed in boxcars and locomotives to provide autonomous measurements, reduce inspection costs, and collect more data. Geometry cars, as they are called, collect measurements of each foot of the track to calculate the deviations.

The research question behind this project is to look for relationships between components' health and geometry data collected over the same tracks in a short period and to identify possible correlations. This is important since track inspectors (railroad personnel that inspects the tracks in a weekly basis) can utilize this correlation to inspect the track when no geometry data is available or when the inspections are delayed. There is a paper that shows that one bad tie can affect the life of the adjacent ties. (Soufiane, Kenza & Zarembski, Allan & Palese, Joseph. (2022). Effect of Adjacent Support Condition on Premature Wood Crosstie Failure. Transportation Infrastructure Geotechnology. 9. 10.1007/s40515-021-00168-5). This research was conducted using only geometry data, not components' data. Since the technology used to collect the component data is still under development, there are no previous works done.

The first part of this project is EDA on the dataset, to better understand how the data is distributed, identify possible outliers and to identify possible relationships and linear dependencies between the measurements.

After the EDA part, we will perform a predictive modelling using the components' data to predict gauge (data from geometry cars). This part will fulfill the research question of the project with a model that from the components data is able to predict the gauge.

With the insights from EDA part we are able to perform some data cleaning before moving on with the next phase. This is important to keep the dataset out of outliers and measurement errors.

It can be possible, at the end, that there is no directly correlation between components' data and geometry data. In this case, new methods, data will be suggested to perform that task.

The data can be found in [Project](#) , the dataset used is already a combination of component and geometry data.

Table 1: Description of data

Column name	Type	Range	Unit	Observation
SectionID	int	0-inf		The ID number of section containing the tie. Each section is 2m wide.
Distance_m	float	0-inf	meters	The position of the tie. The distance is computed from the beginning of the survey. The unit is meter
Material	string	Wooden, Concrete		The material of the tie
Rating	int	0-3 (Wooden) 0-2 (Concrete)		The rating of tie based on condition of crack on surface of tie. 0: good. 3 or 2 is bad.
Askew_Angle	float	-inf to inf	degrees	Askew angle of the tie. This is angle between the horizontal line and the center line of the tie (passing through middle of tie) positive if we need to rotate the tie clockwise to make the center line fit the horizontal line and it is negative if we need to rotate the tie counter-clockwise. The unit is degrees.
FastenerConditionROI_1	int	1, 2, 3, 20		The condition of fastener in ROI 1 to 4. 1: Good, 2: Covered, 3: Missing and 20: Defective
SpikeTotalROI_1 to 4	int	0-inf		Number of spike in each ROI on tie
SpikeMean_Height_ROI1_mm	float	0-inf	milimeters	The mean height of all tie in each ROI on tie. The unit is millimeter
SpikeNearRailROI_1	int	0-inf		Number of spike is near the rail boundary in each ROI on tie
AnchorROI_1	int	0-inf		Number of anchor in each ROI (relating to the tie)

Column name	Type	Range	Unit	Observation
Tieplate_Right	int	0-inf		Number of tieplate on left/right side of tie
TieplateCondition_Right	int	1-5		Condition of tieplate. 1: Good, 2: Sunken, 4: Covered, 5: Twisted
Latitude	float	-90 to +90	degrees	GPS coordinates of the center of the tie
Longitude	float	-90 to +90	degrees	GPS coordinates of the center of the tie
GAUGE	float	-inf to +inf	inch	distance between rails
CURVE	float	-inf to +inf	degrees	degree of curvature

Exploratory Data Analysis

Statistical Analysis on Track Elements-Tie

In this project, we are studying a rail track with multiple elements. In this section, we study some of the most important elements that help us build our understanding of the track system. We started with ties. There are two types of ties in the track under study: concrete ties, and wooden ties. [Figure 1](#) shows the total number of each tie material. The total number of ties is 89985. 62.5% of the ties are made of concrete (56295 ties) and the remaining 37.5% are wooden ties (33690 ties).

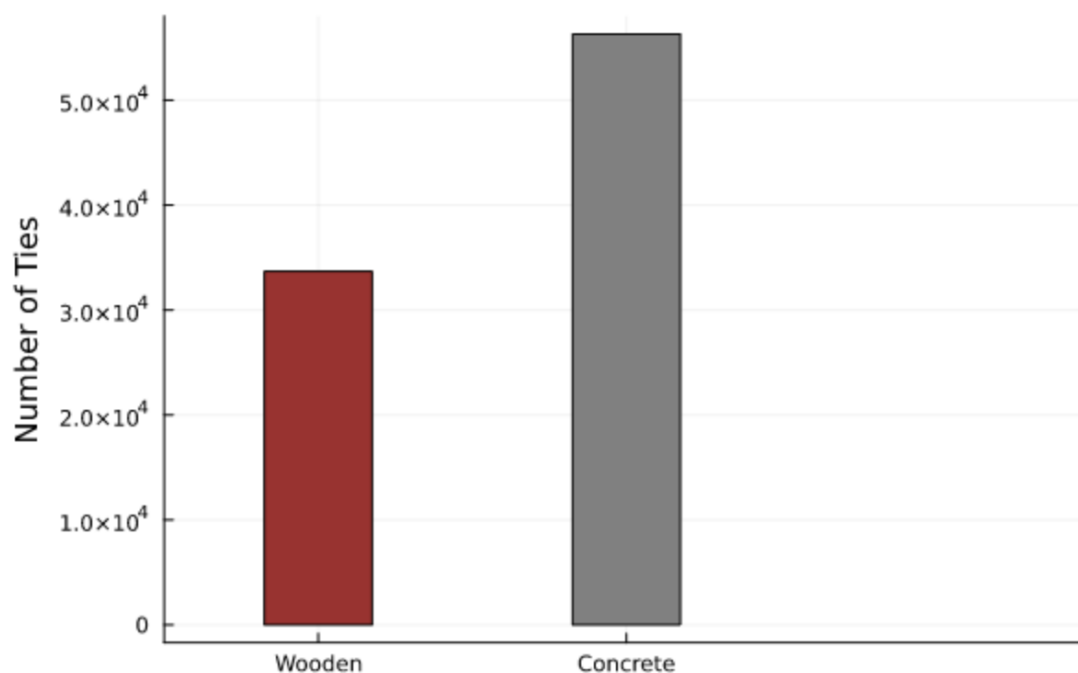


Figure 1: Tie Material

The data set provides condition ratings for each type of tie. [Figure 2](#) illustrates the distribution of different tie conditions for concrete ties. the rating goes from 0 to 2. Zero represents the good tie condition, and 2 represents the worst tie condition. To create this figure, first, we filtered the data frame to eliminate the wooden ties. Next, we plotted the histogram of the concrete tie ratings to get the number of ties with different conditions. In the end, we calculated each condition rate's percentage to better understand our tie health rate. The percentages are as follows: 79.5% of the ties are labeled as 0 (good), 13.5% have a rating of 1, and the remaining 7% are in poor condition.

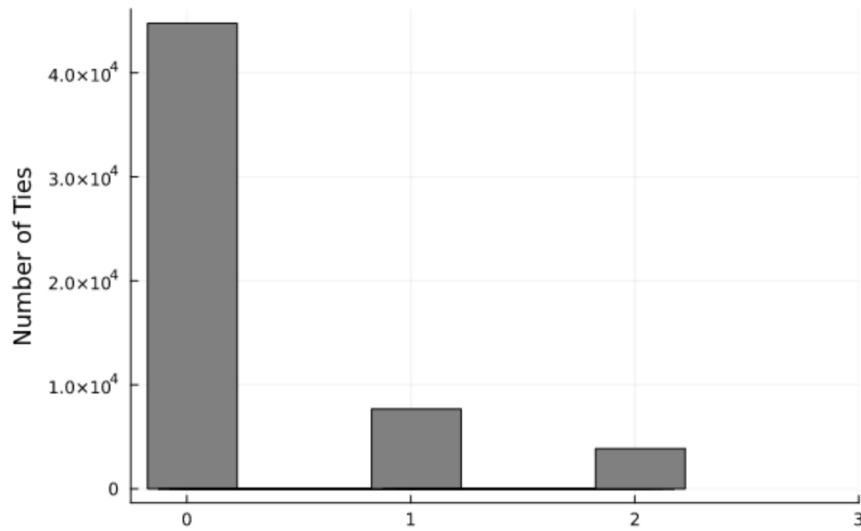


Figure 2: Concrete Tie Condition

The same process was done to get the health condition of wooden ties. Figure 3 shows the distribution of wooden tie ratings. The only difference between concrete and wooden tie ratings is we have a 3rd rate for wooden ties which represent the most damaged ties. The percentages of different condition rates are as follows: 98% of the ties are labeled as 0 (good), 1.5 % are 1, and the remaining 0.5% are in poor condition (2 and 3). These figures suggest that the overall condition of wooden ties is better than concrete ties.

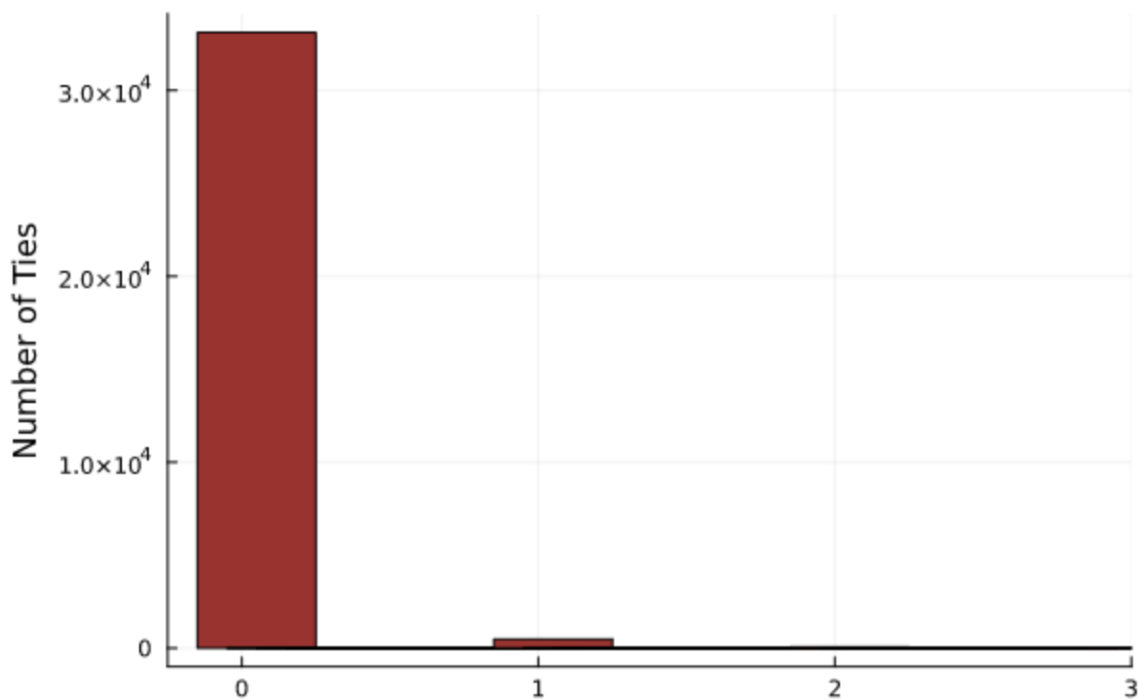


Figure 3: Wooden (Timber) Tie Condition

Moving forward to the next elements of the track: tie plates, anchors, and spikes. It is worth noting that concrete ties don't require any of the mentioned components, which means this section of the project only focuses on wooden ties. We will get back to the concrete ties later in this report.

Statistical Analysis on Track Elements-Tieplate

Our data set describes tie plate existence, along with their conditions. The condition of each tie is represented by a single rating value in the range of 1 to 5, 1 being the good tie plate condition, 2 meaning the plate is sunken, 3 meaning the plate is mildly damaged, 4 meaning the plate is covered, and lastly, 5 means the plate is twisted. To analyze the overall condition of tie plates, we counted the number of existing tie plates in right and left sides of the ties. To do so, first, we filtered out the rows of our data frame where the tie plate didn't exist. Our results show that a total number of 33517 tie plates exist on the right side of the ties, and 33397 tie plates exist on the left side of the ties. As mentioned earlier, the total number of wooden ties is 33690. This means about 0.7% of the ties miss at least one tie plate. A total of 214 wooden ties, don't have any tie plates, which is about 0.6% of wooden ties. Following the same methodology explained for ties, we plotted the histogram of different tie plate conditions for each side of the ties. On the right side of figure 4, the distribution of right tie plate conditions is shown, and the left side illustrates the left tie plate conditions.

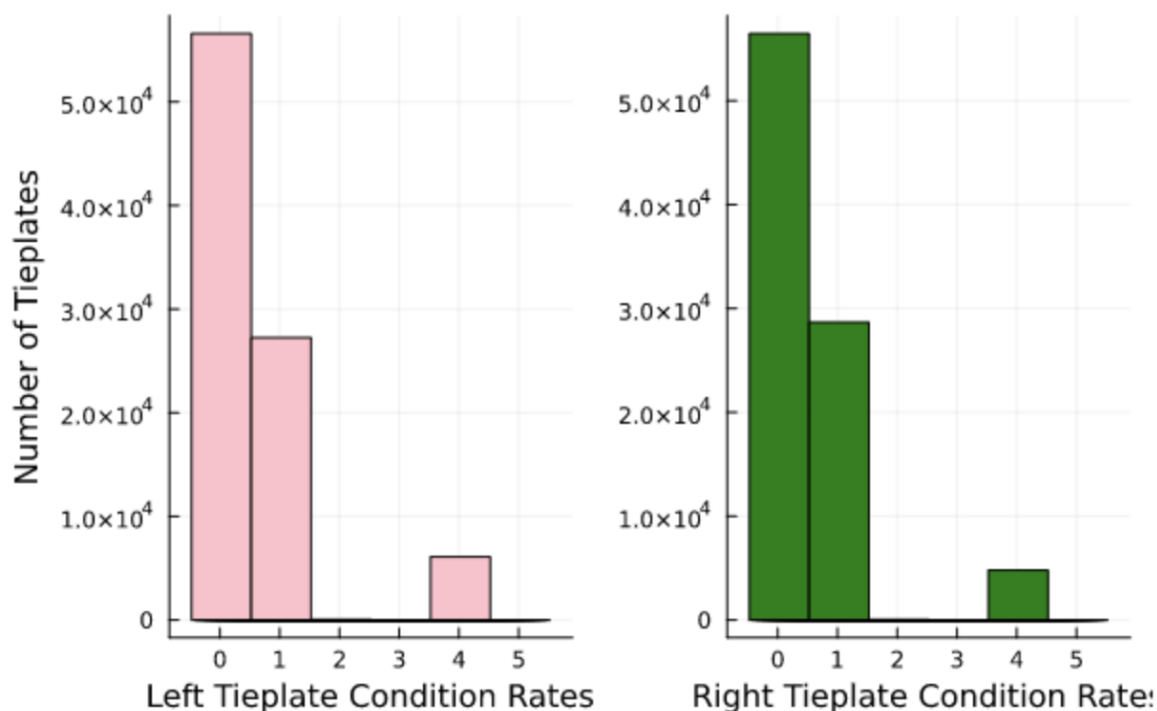


Figure 4: Tieplates Condition

Statistical Analysis on Track Elements-Anchor

We have four regions of interest on track, two on the left-hand side and two on the right side of each tie. This will be discussed in more detail further on the project. There are 4 columns in our data frame describing the anchors in each of the regions. In this section, we counted the number of existing anchors in each region, along with plotting their distribution. To do so, we selected the 4 channels from the data frame, then we filtered each region of interest to use in our calculations. And lastly, we plotted the bar chart of our selected channels. Given the fact that wooden ties require anchors on every other tie, we expect to have around 16850 anchors on the track. Based on our calculations there are 14470 anchors in the first region of interest, 14159 in the second one, 14444 in the third

one, and 14355 in the last region. On average, 42% of the wooden ties have anchors on them (8% less than our expectations). Figure 5 illustrates this.

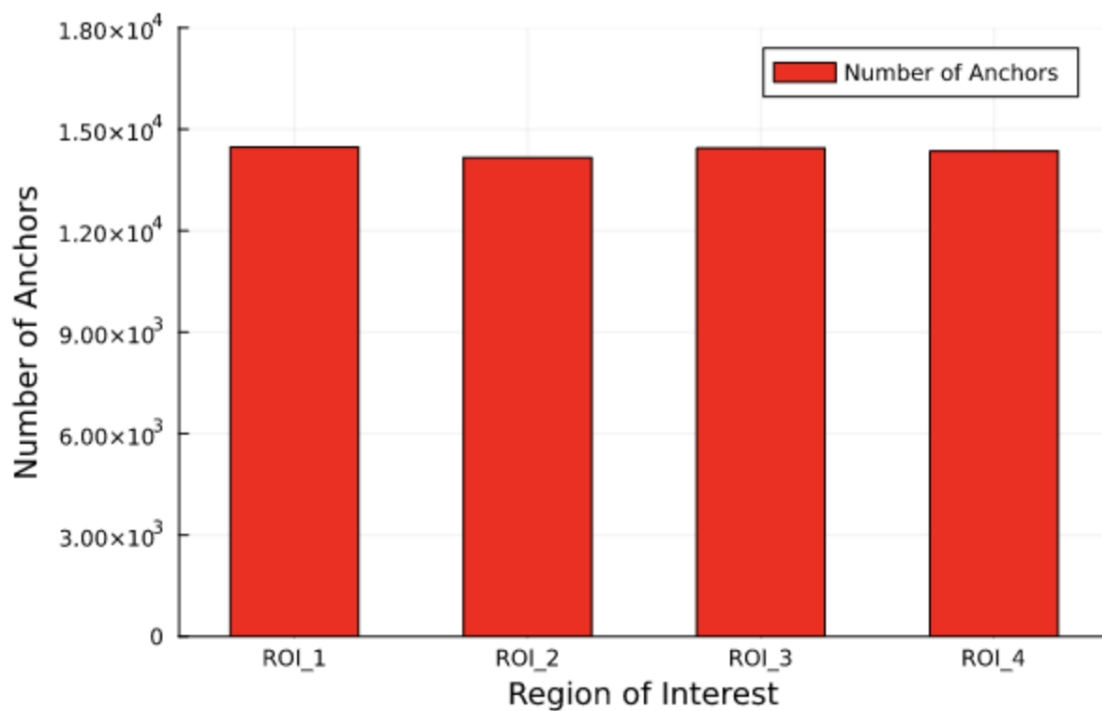


Figure 5: Anchors Condition

Statistical Analysis on Track Elements-Fastener

Our data frame describes the condition of the fasteners in each region of interest in 4 different channels, each belonging to one region. The condition rating is 1,2,3, or 20. 1 means the fastener is in good condition, 2 means the fastener is covered, 3 means the fastener is missing, and 20 means it is defective. To analyze our data, first, we selected 4 channels corresponding to fasteners from our data frame, then we filtered each region of interest to analyze them separately. We plotted the histogram of the fastener condition of each of the regions, then plotted all 4 plots together, which can be seen in Figure 6.

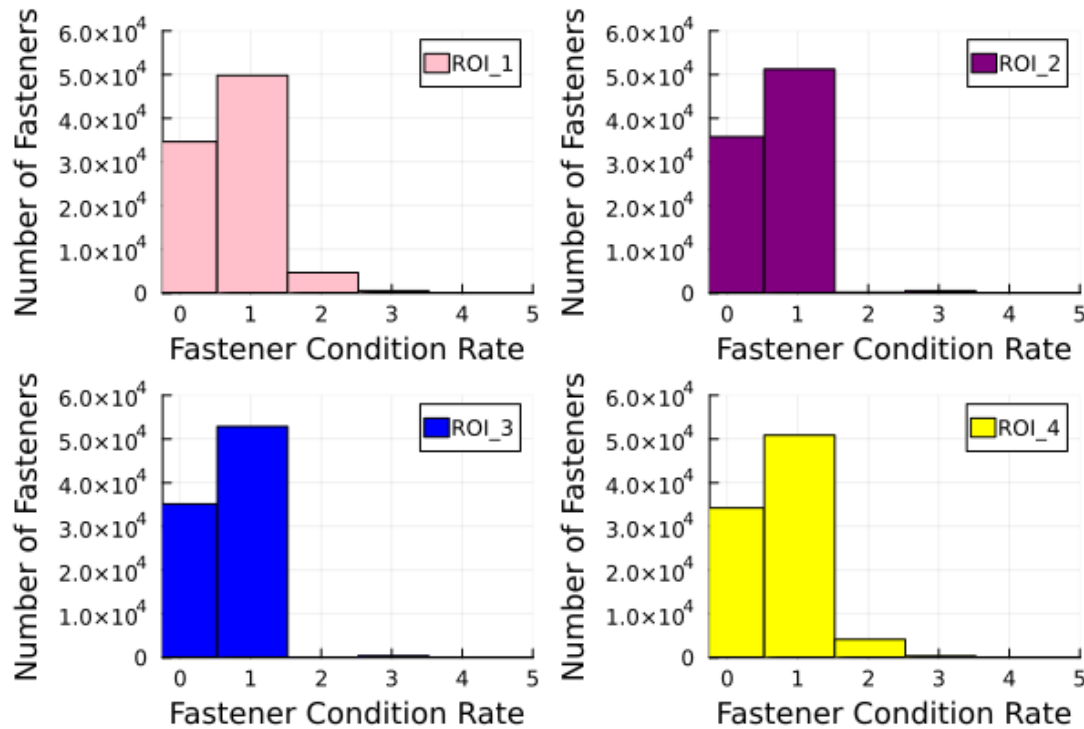


Figure 6: Fastener Condition

Does Askew Angle relate to the number of anchors?

Anchors are spring steel clips that attach to the underside of the rail baseplate and bear against the sides of the sleepers to prevent longitudinal movement of the rail, either from changes in temperature or through vibration. Based on the definition of askew angle in table 1, the anchors may have a relationship with the askew angle. Figure 7 shows regions of interest 1 to 4 for anchors. The following parts describe the steps we do to investigate any relationship between the askew angle and the number of anchors in a tie.

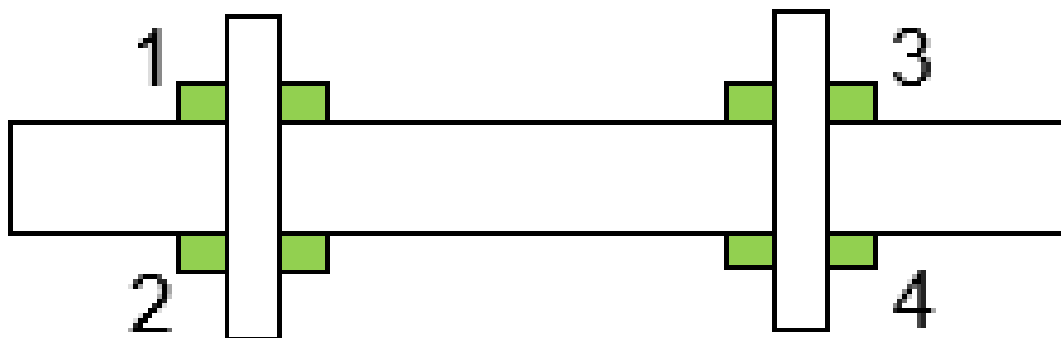


Figure 7: Anchor Regions

Selecting related columns:

To create the data frame, using the "select" function in Julia, we select columns about askew angle and anchors, which are "Askew_Angle", "Material", "AnchorROI_1", "AnchorROI_2", "AnchorROI_3", "AnchorROI_4". The material is important because we only have anchors for wooden ties. Therefore, the data frame has 89982 rows and six columns.

Filtering wooden ties:

Using the “filter” function, we choose only ties with wooden material. The new data frame for wooden ties has 33687 columns which shows more than half of the ties are concrete.

Adding total number of anchors:

Since we have only 0 or 1 for having or not having an anchor in the raw data, we need to know the total number of anchors for each tie. Therefore, we add a column to our data frame representing the total number of anchors for each tie by adding columns for anchors in each ROI.

Describing statistical features of data:

The “describe” function gives a good overview of the statistical features of the data frame. Using this function, we find the statistics for our data. Figure 8 represents the result.

	variable	mean	min	median	max	nmissing	eltype
	Symbol	Float64	Real	Float64	Real	Int64	DataType
1	:Askew_Angle	-0.741635	-14.19	-0.73	6.13	0	Float64
2	:AnchorROI_1	0.429483	0	0.0	1	0	Int64
3	:AnchorROI_2	0.420221	0	0.0	1	0	Int64
4	:AnchorROI_3	0.428682	0	0.0	1	0	Int64
5	:AnchorROI_4	0.42604	0	0.0	1	0	Int64
6	:Total_Anchor_Number	1.70443	0	0.0	4	0	Int64

Figure 8: Data Description

Plotting corrplot for data:

The first thing we can plot to check whether some variables are correlated or not is a corrplot. Figure 9 shows the corrplot for the data.

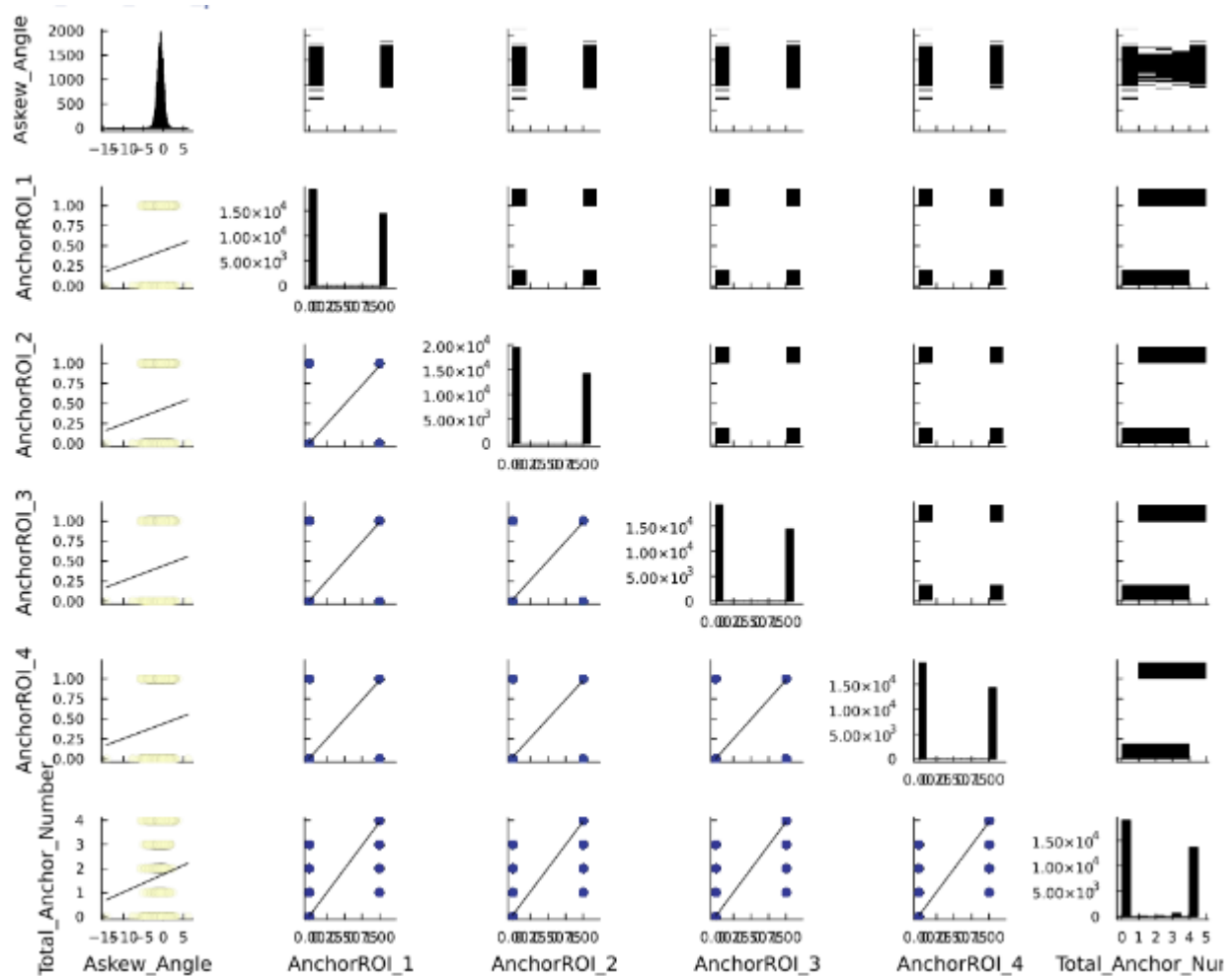


Figure 9: Askew Angle and Number of Anchors Corrplot

The corrplot indicates that there is no correlation between askew angle and anchors. However, this idea comes to mind that comparing the statistics for situations with a different total number of anchors assures us that we don't have any relationship between askew angle and anchors. Therefore, we filter the data based on each value we can have as the total number of anchors (0 to 4) and find their distribution of askew angle using the "histogram" function.

Plotting histogram for different values of total number of anchors:

Plotting all five histograms in a graph shows us that they are so close and different numbers anchors have not caused different askew angles. Figure 10 shows these histograms.

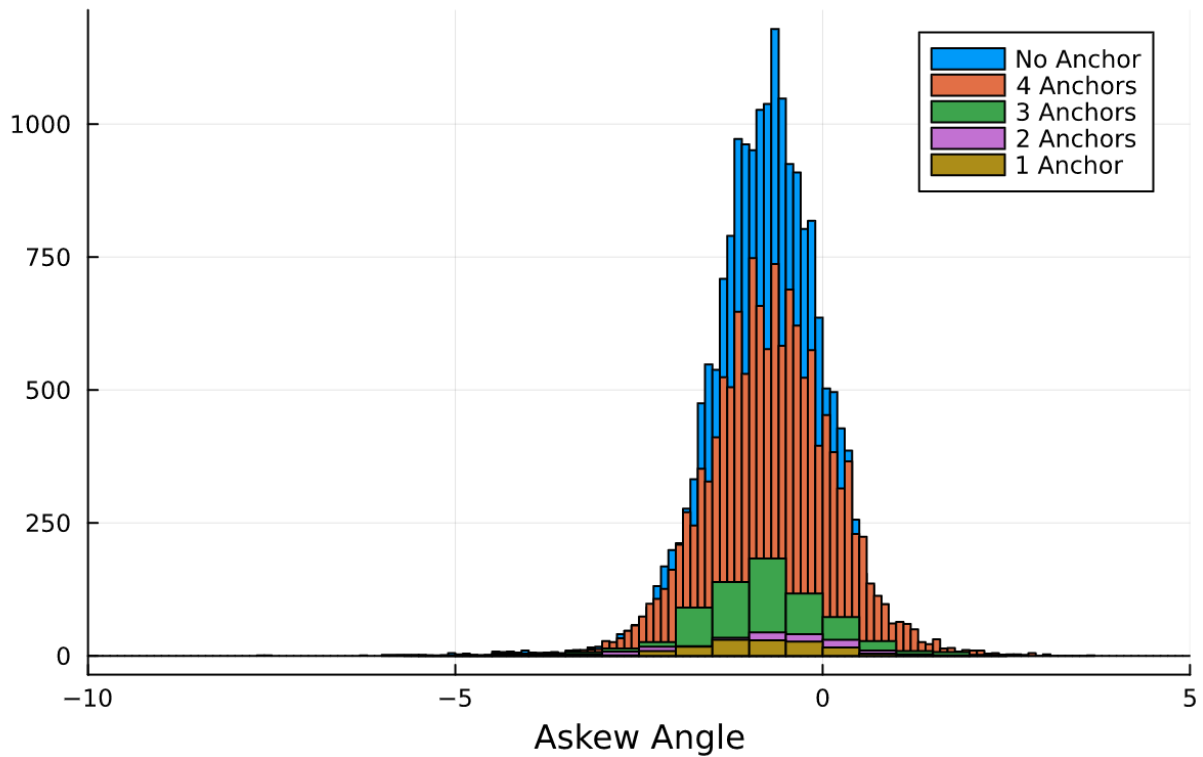


Figure 10: Number of Anchors per Tie Histogram

Different numbers of bins are because of the different numbers of data we have in each situation. To make our conclusion more precise, we normalize the data and compare situations with no anchor and 4 anchors. Figure 11 represents no meaningful difference between these two conditions.

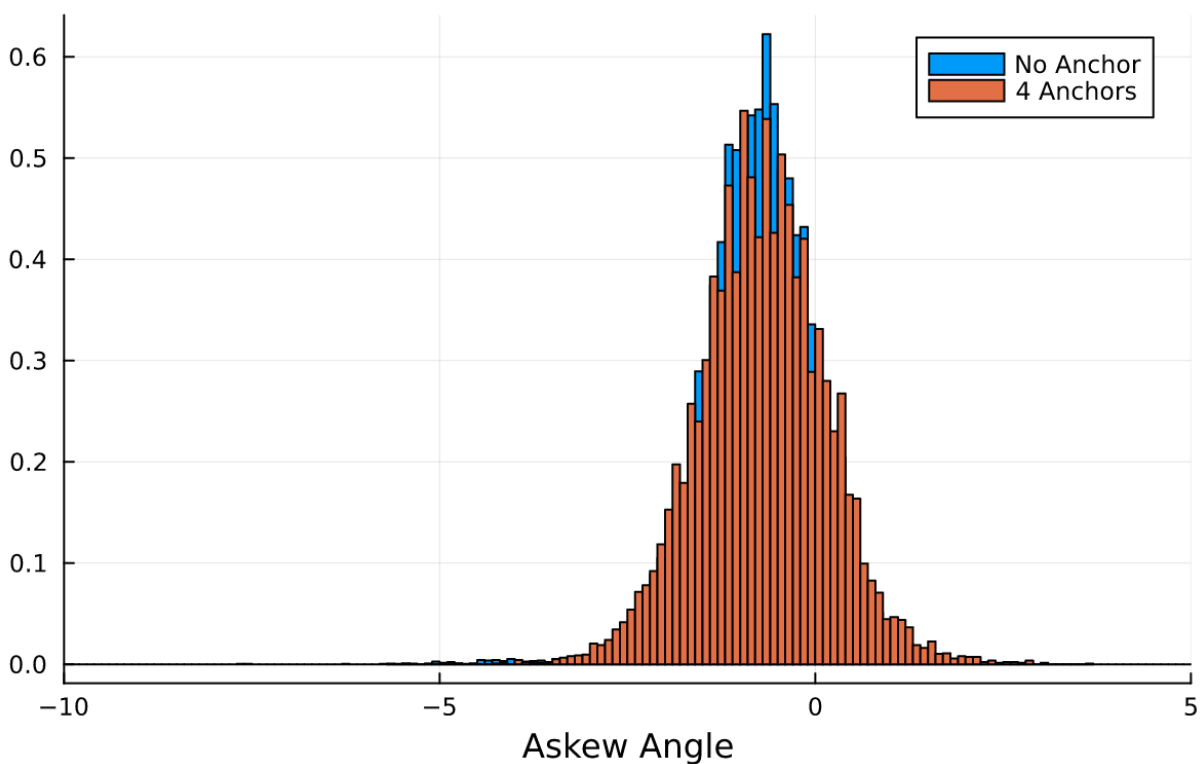


Figure 11: Normalized Histogram

Plotting boxplot:

The last graph we plot is a boxplot which provides a good sense of statistics to compare skew angle with different total numbers of anchors. Figure 12 shows these boxplots and confirms no such

difference between the mean and other statistics of askew angle.

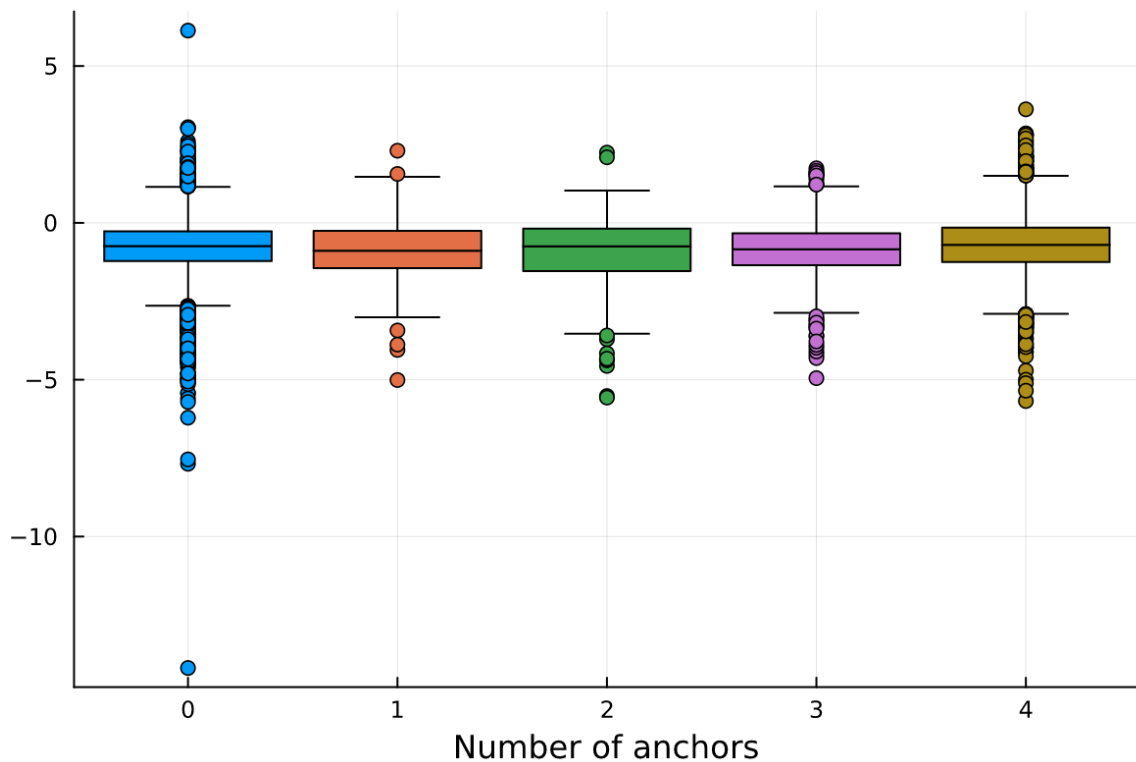


Figure 12: Box Plot

Conclusion:

Even though the anchors seem to play a significant role for the askew angle, the data surprisingly shows they do not have any relationship.

Does Askew Angle relate to the number of spikes?

As what we had for anchors, we are interested in finding a relationship between spike conditions and askew angle. The steps to figure out this relationship is as follows.

Selecting related columns:

To create the data frame, we select columns about askew angle and spikes, which are "Askew_Angle", "SpikeTotalROI_1", "SpikeMean_Height_ROI1_mm", "SpikeNearRailROI_1", and these last three columns for ROI 2, ROI 3, and ROI 4. Therefore, the data frame has 89982 rows and 13 columns.

Plotting corrplot for data:

The first thing we can plot to check whether some variables are correlated or not is a corrplot. Figure 13 shows the corrplot for ROI 1. As the corrplot for other regions of interest are similar to ROI 1, we do not put it on the report. For the rest of this part, we just consider ROI 1 because the results are the same for all the regions.

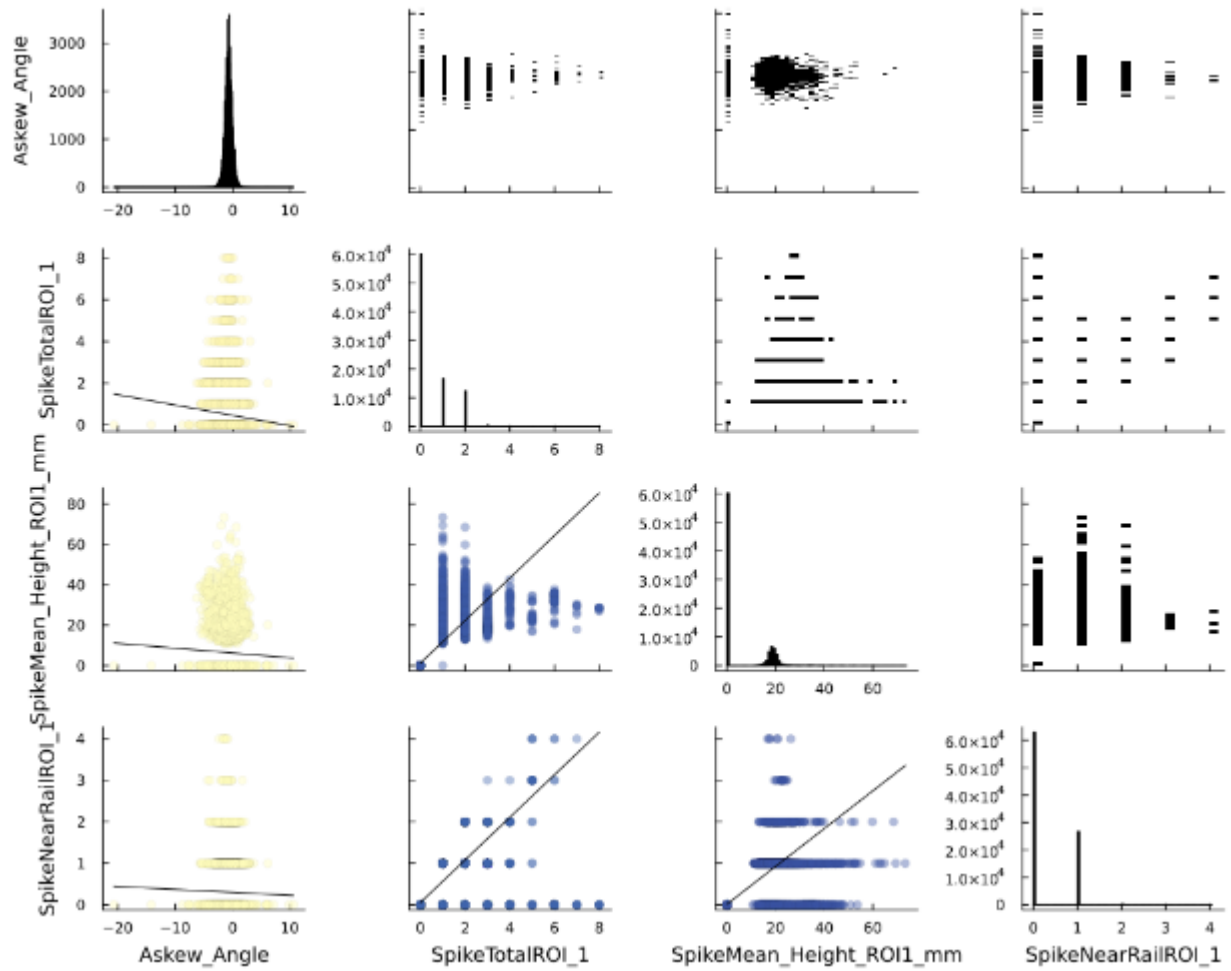


Figure 13: Askew Angle and Spikes Corrplot

Like anchors, the corrplot indicates no correlation between askew angle and spike conditions. However, an interesting pattern can be seen in the scatter plots we have in corrplot. To illustrate this pattern better, figure 14 shows a scatter plot for the number of spikes in ROI1 and askew angle, which is one of the plots in corrplot. We can see that the range of the askew angle for fewer spikes is larger. It is reasonable and having no correlation between askew angle and spike condition is because a large share of ties are healthy and do not have defective spikes. However, the unsatisfactory situation is crucial for us. Therefore, as the askew angle is approximately symmetric relative to zero, we consider the maximum askew angle for the different numbers of spikes and find the correlation between it and the number of spikes.

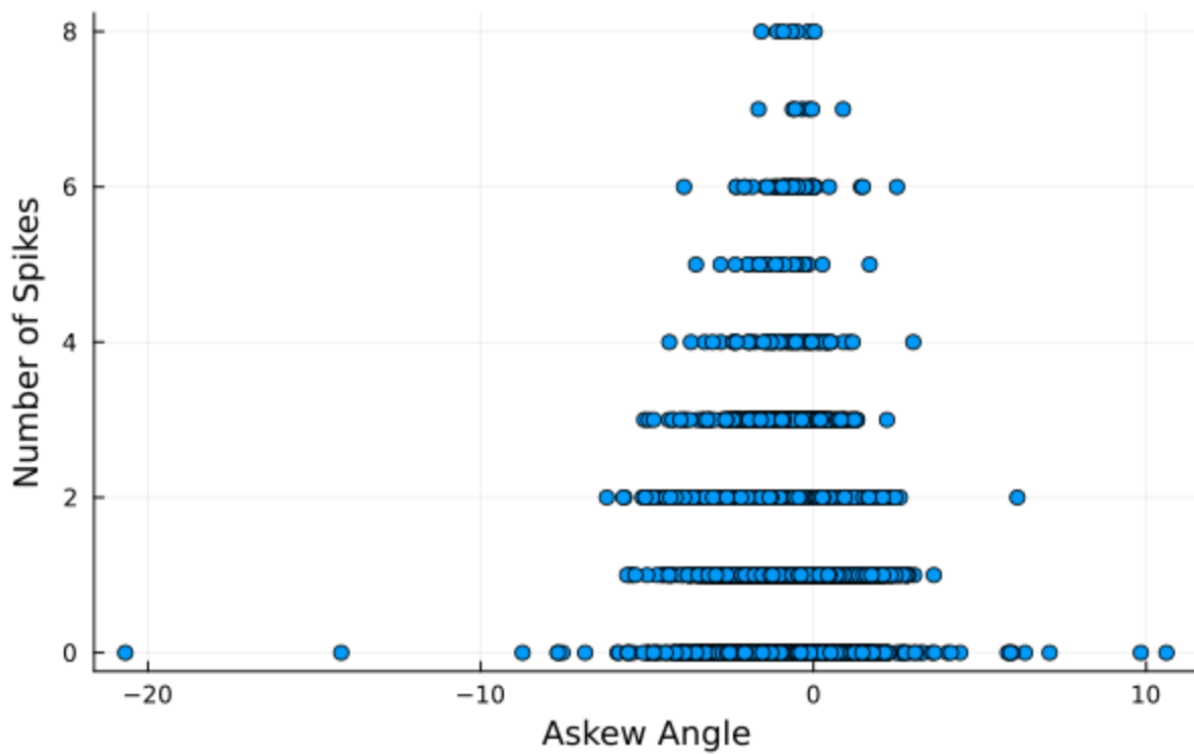


Figure 14: Askew Angle and Spike Scatter Plot

Correlation between maximum askew angle and the number of spikes:

Using “corr” function, we find the correlation value between the maximum askew angle and the number of spikes which is about -0.64. Therefore, we say they are correlated, and there is a relationship between the maximum askew angle and the number of spikes. The same results happen for other regions of interest and even the total number of spikes in all regions.

Relationship between Gauge Deviation and Concrete Tie Rating

During the EDA section we tried to look for relationships between the gauge of the track and the component's condition. The Gauge Deviation is the difference between the standard gauge (4'-8.5") and the measured gauge. So a value close to 0 indicates that the gauge is correct. Due to the train traffic and tie degradation the gauge deviation tends to increase over time. When it reaches certain limits, it is called a defect and maintenance must be made to restore the gauge. Negative values usually mean that the tie was installed with a tighter gauge or that there was a measurement error and should be considered outliers if it is too tight, around -1 in or less. To be reasonable in this section, the analysis are divided between concrete and Wooden ties, since they have different characteristics that make it harder compare both together. The most directly relationship is the gauge of the track and the condition of the ties. Based on the technology used the concrete ties are graded from 0 (Good) to 2 (Bad). 1 (Fair) is a intermediate condition. Figure 15 represents the distribution of gauge deviation by each concrete tie rating.

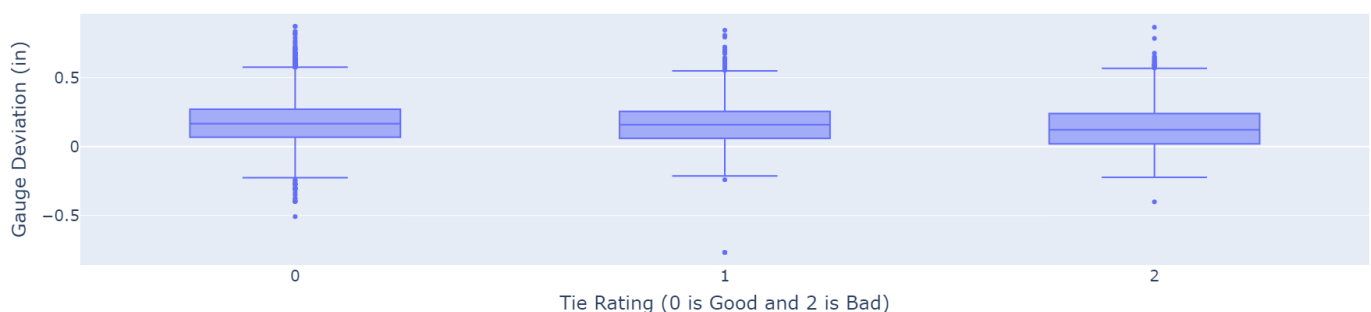


Figure 15: Gauge and Concrete Tie Condition

Intuitively we can think that bad ties would have wide gauge, but for this specific region it is not the case. Comparing the median values we see that the values are not too far from each other. Another interesting finding is that the max value does not have big changes when changing the tie rating, indicating that any tie rating can have a gauge defect.

Relationship between Gauge Deviation and Wooden Tie Rating

Now let's look to the same criteria but for Wooden ties. Wooden ties have one more bin for classification, still following the same idea as 0 being Good and 3 being Bad. Figure 16 shows the distribution of gauge deviation by each wooden tie rating.

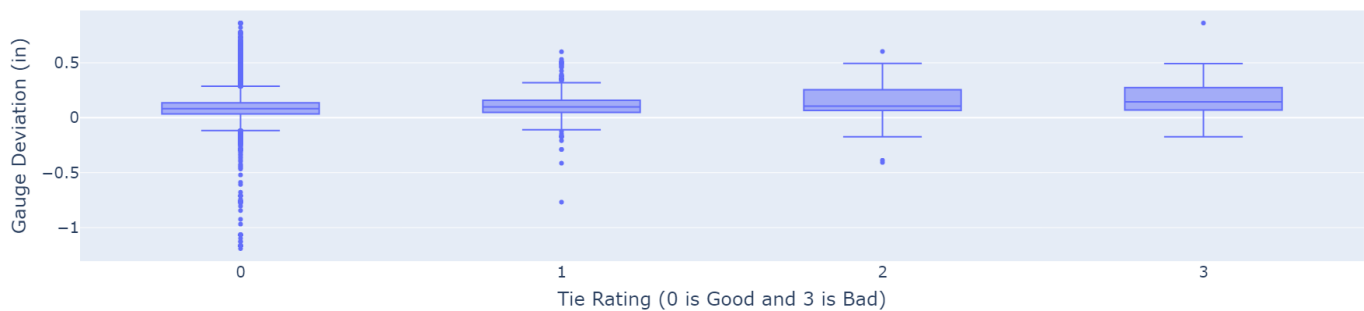


Figure 16: Gauge and Wooden Tie Condition

First thing to mention is that the Good ties (Condition 0) has more outliers than the other ones, this can be related to the fact that there are a lot more ties with this condition than the other ones. In contrast to the concrete ties, the median gets slightly higher when we have a bad condition tie, indicating that the condition of the tie can be related to the gauge of the track. Comparing the max values, only condition zero (Good) and 3 (Bad) have higher max values. It makes sense for condition 3 (Bad) ties but measurement errors should be account for condition 0 (Good) ties.

Relationship between Gauge Deviation and Tie Plate Condition

The next component to be investigated is the tie plate. Only wooden ties have tie plates. Their main function is to support the rail and distribute the load on the ties. Normal wooden ties have two tie plates – one per rail. The metric used to analyze the data is the number of defective tie plates per tie, 0 being a tie with 2 tie plates and 2 a tie plate without tie plates in good condition. Figure 17 shows the distribution of gauge deviation by the number of defectives tie plates per wooden tie.

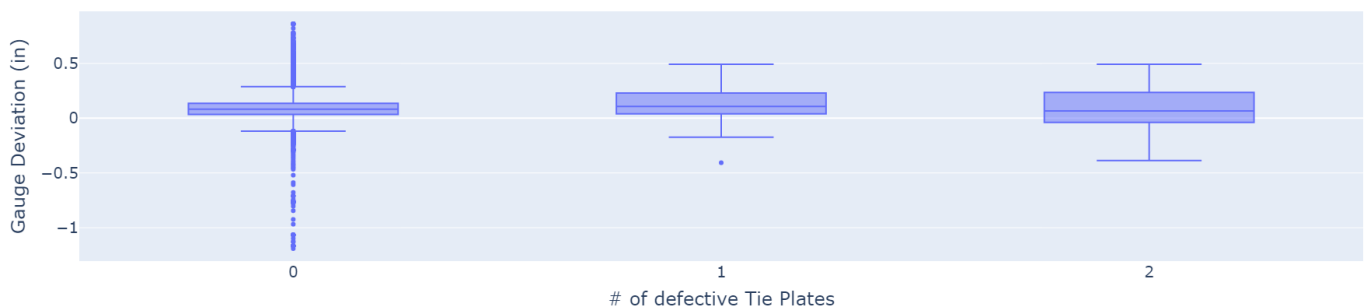


Figure 17: Gauge and Tie Plate Condition for Wooden Ties

Again, the number of outliers for ties with two tie plates must be taken into consideration since this is the highest population on the dataset. The main takeaway is that the number of defective tie plates are not directly related to the gauge.

Relationship between Gauge Deviation and Number of Defective Fasteners on Wooden Ties

Fasteners are the components that have hold down force to hold the rails connected to the tie plates. They are used in very specific situations in the railroad, such as tight curves, switches and grade crossings. When used, each tie has four fasteners, two per rail, one on each side of the rail. Fasteners may or may not be used in wooden ties but must be used in concrete ties. Figure 18 shows the distribution of gauge deviation by the number of defectives fasteners per wooden tie.

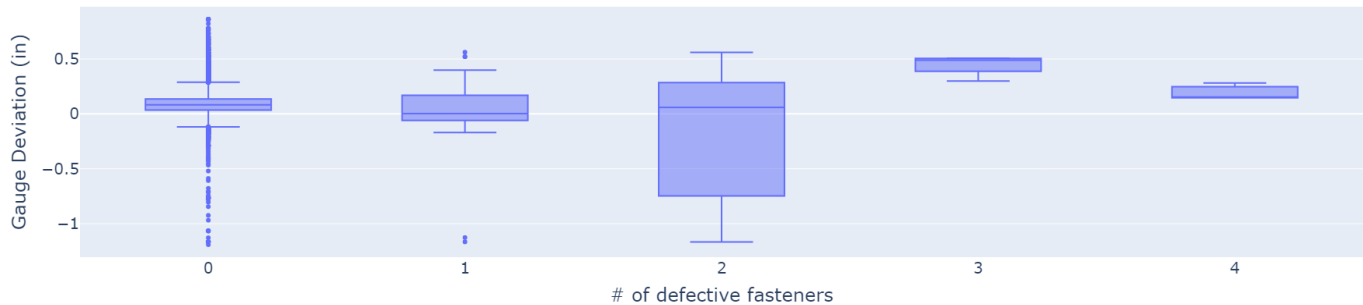


Figure 18: Gauge and Number of Defective Fasteners per Wooden Tie

Outside the outliers for ties without any defective fasteners, ties with only two defective fasteners have more variability in gauge. It can be explained by the fact that if the two missing fasteners are on the outside side of the rails, it is likely that the gauge will be widened. The ties with 3 and 4 bad fasteners are not that common, but when they happen they have wider gauge than the ties in better condition. Only one bad fastener does not have much influence in the gauge. [Relationship between Gauge Deviation and Number of Defective Fasteners on Concrete Ties](#)

For concrete ties there were no ties with zero defective fasteners (all four fasteners present) or with 4 defective fasteners (no fasteners at all). That can be related to the fact of having less concrete ties in the section analyzed. Figure 19 shows the distribution of gauge deviation by the number of defectives fasteners per concrete tie.

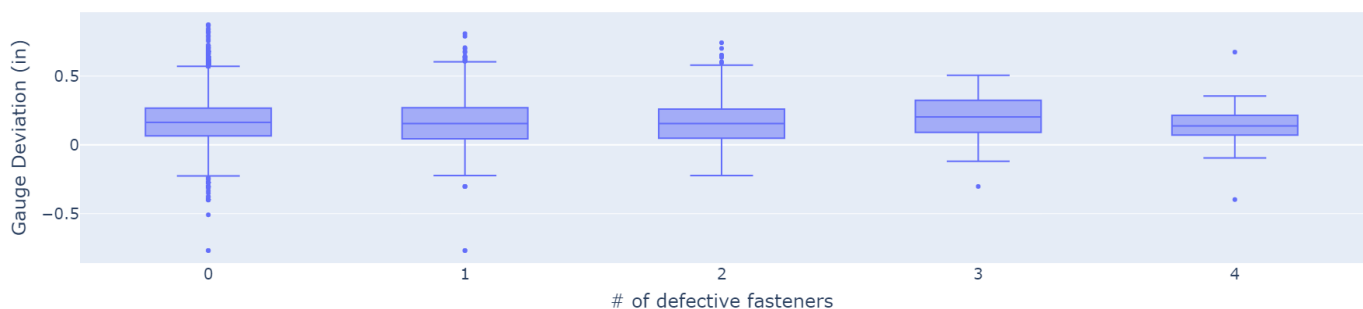


Figure 19: Gauge and Number of Defective Fasteners per Concrete Tie

Concrete ties are less prone to have gauge problems related to defective fasteners, as we can see by comparing the median of the distributions. It can be explained by the fact that they have metal shoulders that holds the rails in place, regardless of having a fastener or not. Concrete ties with 3 defective fasteners are more prone to have wide gauge, as shown by the higher median than the other conditions.

Is there a meaningful relation between the Degree of Curvature and Gauge Deviation?

Degree of curvature is a measure to determine the sharpness of a curve. The definition is found by connecting two points on an arc with a 100-foot chord, drawing radii from the center of the arc to the chord end-points, and then measuring the angle between these radii lines. The larger the degree of curvature, the sharper the curve is. A positive degree of curvature means the track route is towards right, and a negative degree of curvature represents a left turn.

First, we looked at the correlation plot shown in figure 20 to investigate a possible relationship between the Degree of Curvature and Gauge. We used the absolute value of the curvature here to make the resultant plot more accurate.

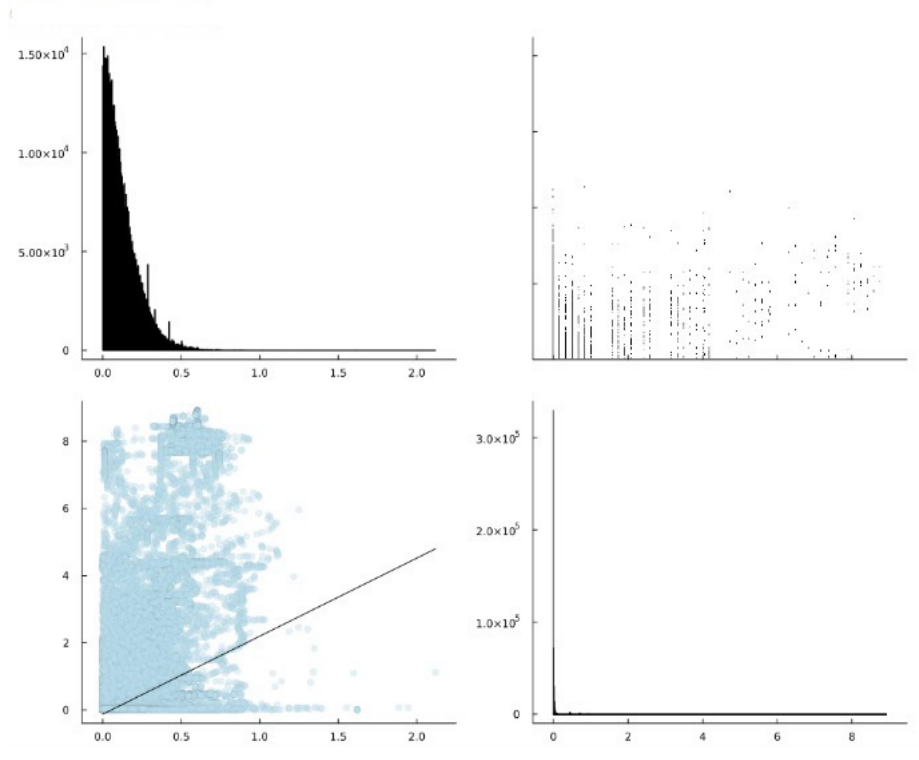


Figure 20: Correlation Between Gauge and Degree of Curvature

The correlation plot shows a slight positive correlation, meaning the greater the degree of curvature or in other words, for sharper curves, the changes in gauge are greater. By looking at the Curve histogram we realized that the majority of this specific rail track is on a straight route with 0 degree of curvature. However, for better visualization we filtered the curve column for values greater than 0.1 and less than -0.1. Also, the Gauge histogram in 21 suggests that a significant percentage of the measured gauges, are actually in good condition.

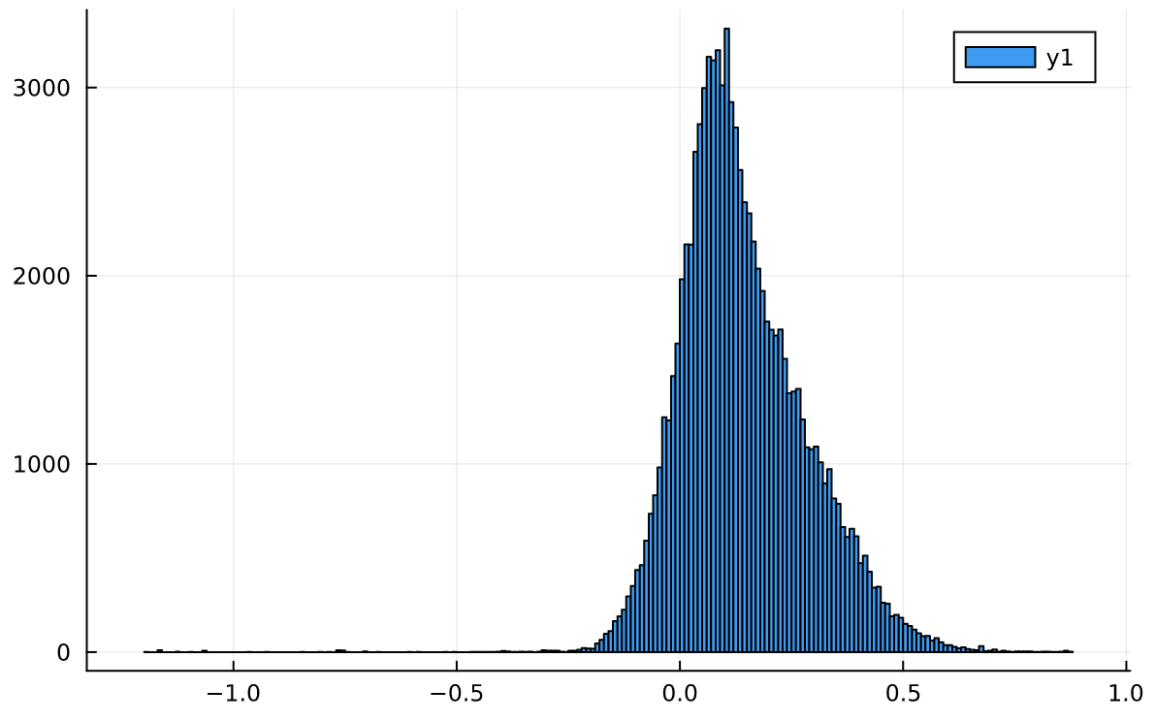


Figure 21: Gauge Histogram

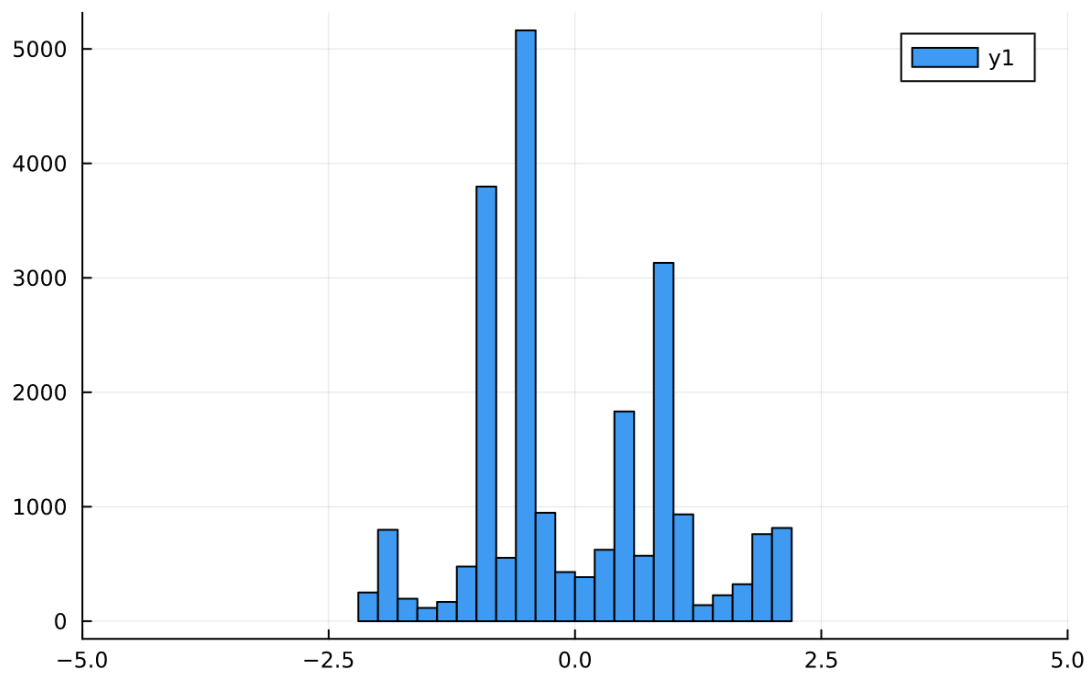


Figure 22: Filtered Curvature Histogram

We also looked at the positive and negative curvature separately. Despite our initial assumption the absolute value of curvature does not provide more accurate results, while the correlation between negative curvature and gauge is greater than the one for positive curves. The correlation was calculated using `cor` function in Julia. The Pearson correlation for positive curves and gauge is 0.27, while it is -0.42 for negative curves and gauge. Although we do not know what is causing this difference, this is an interesting finding and can be further discussed.

Is there a meaningful relation between the Degree of Curvature and Tie Condition?

So far, we know that there are two different tie materials in our data set with different condition rating ranges and overall characteristics. To perform a reasonable analyze, we split the dataset into two, one containing wooden ties, and the other including concrete ties. We also had to create a dataframe whit both geometry and Lrail data. To do so, we used Julia function innerjoin, and then filtered the dataset to eliminate no-curved track. Box plots for concrete tie is shown below in [23](#). Concrete tie rates ranges from 0 (good) to 1(fair), and 2 (bad). The median values showed with the lines inside each box indicate that bad ties correspond to higher curvature degrees compared to good ties. But, the range of curve for each tie rating tells us the good ties are not limited to parts of the track with only low degrees of curvature, and curvature is not the only participating factor in tie deterioration.

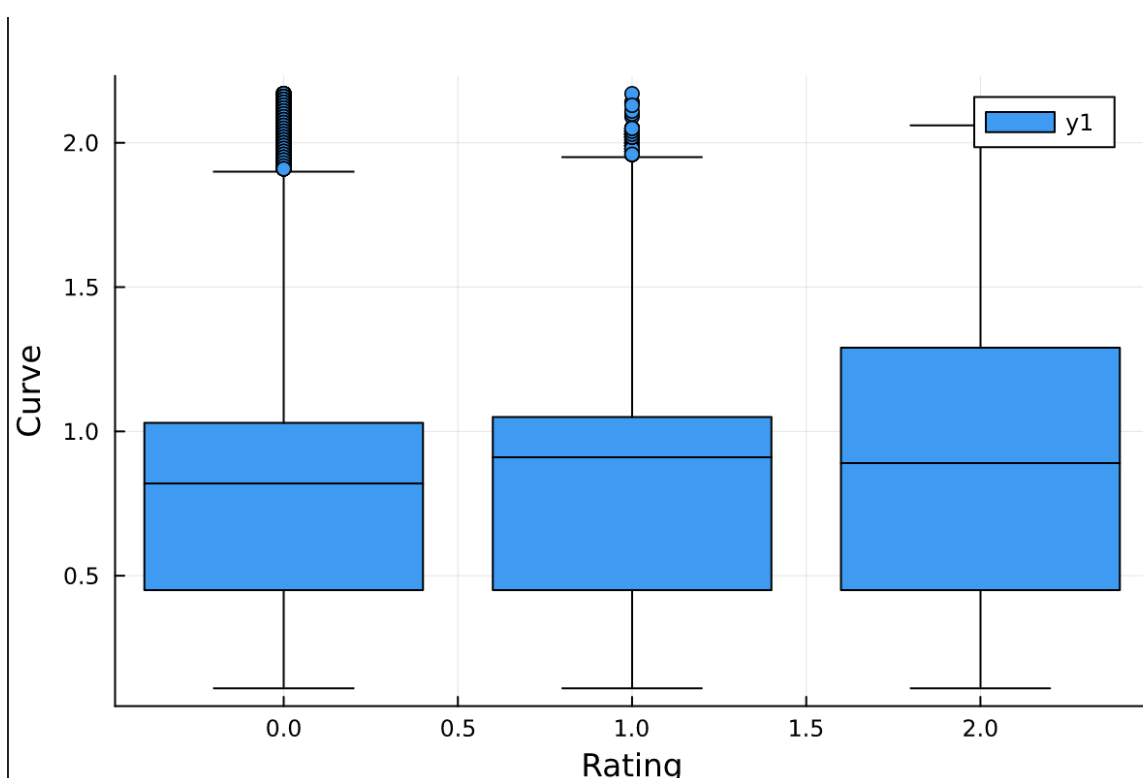


Figure 23: Concrete Tie Rateing vs Curvature

Box plots for wooden ties is shown below in [24](#). Wooden tie rates ranges from 0 (good) to 3 (bad). This plot shows that most of bad wooden ties are located in the parts of the railroad with a higher curvature degree compared to good wooden ties. In fact, the distribution of bad wooden ties suggest curvature is playing a stronger role in wooden tie deterioration compared to concrete ties.

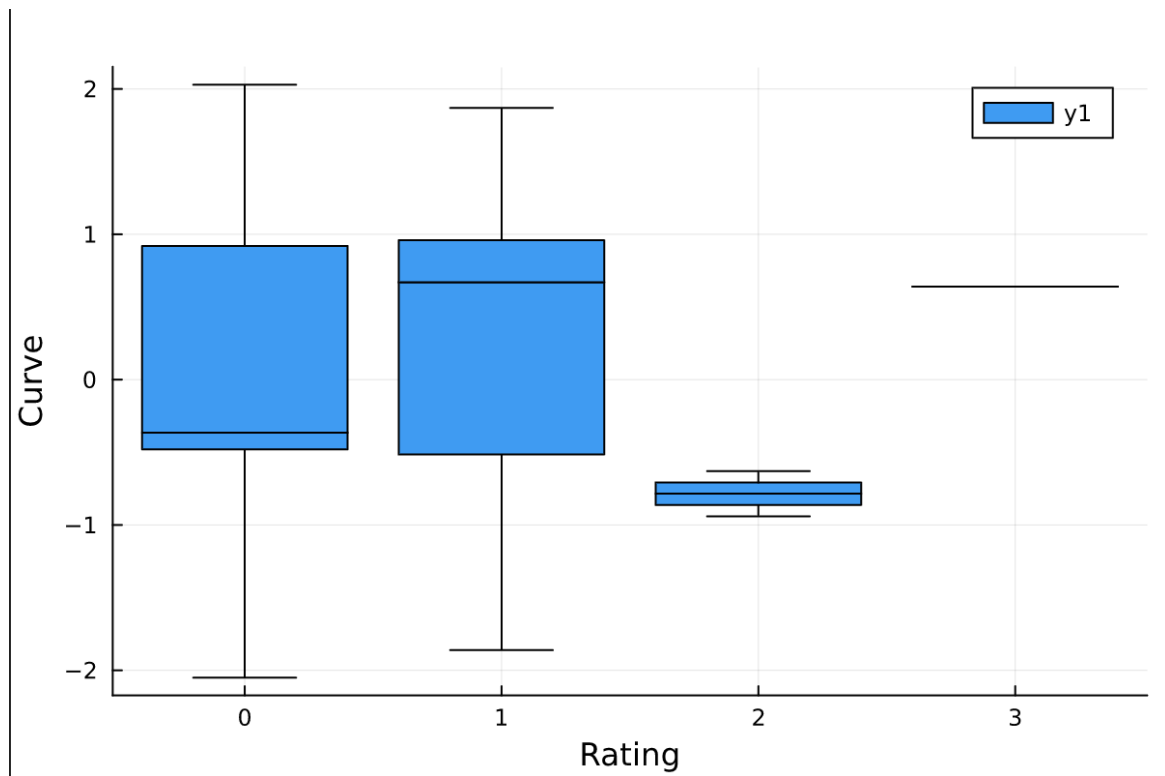


Figure 24: Wooden Tie Rating vs Curvature

Is there a meaningful relation between the Degree of Curvature and Number of Spikes per Tie?

We know that spikes are only used for wooden ties, therefore we filtered all wooden ties. Next, we know that we have the number of spikes in each Region of interest but we need the total number of spikes per tie, so we created a new column with the total number of spikes per tie in it, and added it to our filtered data frame using Julia function `hcat`. Figure below shows a box plot of number of spikes per tie and curvature. What we learn from the maximums shown with lines above each box is that for greater curvature degrees, more spikes are used in each tie, which is consistent with our expectations. But the median, shown with lines inside each box do not follow the same pattern. But remember, the number of ties with more than 8 spikes is significantly lower than the number of ties with 4 to 8 spikes.

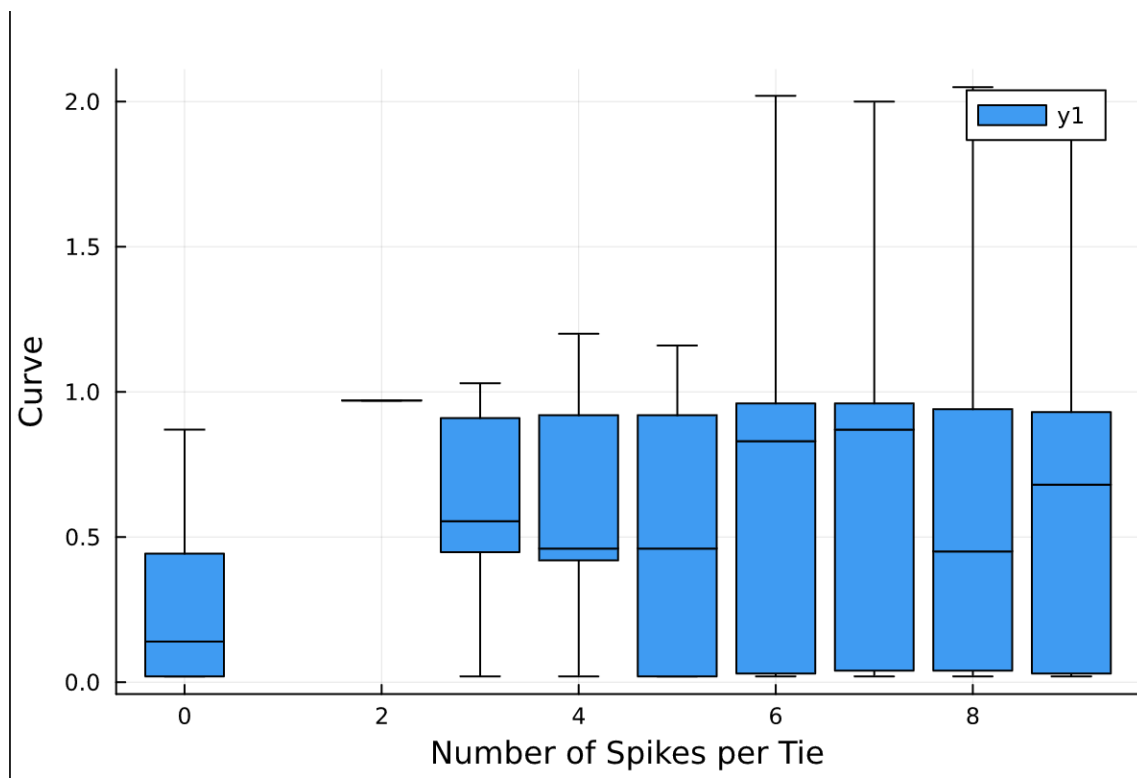


Figure 25: Number of Spikes per Tie vs Curvature

Predictive Modeling

As our next step, we want to use this dataset and the information we gained about it to develop a predictive model. Our goal is to identify track degradation predictors and their contribution weight to the overall degradation. By doing so, we can know about the probable defects in track before they actually happen, which is categorized as preventive maintenance practice. This maintenance method is more cost-efficient and can aim the railroad companies in better allocating their resources. Based on our results we decided to move on with the relationship between Gauge, Curve, Tie Condition, Number of Spikes, and maximum Askew Angle.

We tried to develop a predictive model based on what we have learned about our dataset in the Exploratory Data Analysis section. In the EDA phase, we analyzed the dataset by plotting the distribution of different conditions of various elements of the track, followed by attempting to find potential correlations between different attributes of the track elements. The results suggest that most of the elements under study are not correlated to each other. However, our team believes that there could be a meaningful relation between the track components' condition or existence and gauge. As mentioned before, gauge is the distance between two rails, which has a standard value that needs to be maintained. Wide gauges would cause derailment, and that is why gauge is considered to be among one of the most important features of the track when it comes to inspection, or defect prediction. Therefore, we decided to have gauge as the output of our predictive model. Having a model that accurately predicts gauge could be a helpful tool for the railroad industry since it will decrease the need for geometry data collection. We followed four approaches to create the predictive model, including two layer Neural Network model, three layer Neural Network model, Decision Tree, and Random Forest. The accuracy of all models are unanimously low, which suggests there is not a considerable correlation between the input values, and the output data. More detailed explanation and suggested path forward is provided in the following sections of this project deliverable.

Data Cleaning

First step towards creating a predictive model is having a proper data frame with all the input values lined up. It is worth noting that we separated concrete and timber ties into two data frames, and analyzed them separately. As it is mentioned in the EDA section of the project, the dominant tie material in this track is timber. Since concrete and timber behave differently in a significant way, we thought the correct approach is to divide them. To create the clean data frame, we selected the most relevant track features as our input data. The output data is the gauge which is a number between 0 and 1. As mentioned in the EDA phase of the project, this is the difference between the existing gauge value in the track, and the standard gauge. Numbers closer to 0 mean the gauge is better compared to the ones with higher values. Also, the negative values indicate the gauge is tighter than the normal value. A list of the input and output data for each tie material is provided below:

1. Wooden Tie:
 - Input Data:
 1. Rating : Condition of Tie- Ranked from 1 to 4
 2. Askew Angle
 3. Degree of Curve
 4. Total Number of Spikes
 5. Total Number of Anchor
 6. Frac_Def_TiePlate
 7. Frac_Def_Fasteners
 - Output Data:
 1. Gauge

2. Concrete Tie:
 - Input Data:
 1. Rating
 2. Askew_Angle
 3. Curve
 4. Frac_Def_Fasteners
 - Output Data:
 1. Gauge

Regression

Neural Network Modeling

Our first approach was creating a regression model of the relationship between the input values, and the gauge. We used multiple Neural Network models, with different combinations of activation functions. We split the dataframe into two dataframes. 80 percent of the input values, were randomly selected to be used for training the model, and the remaining 20% were used to evaluate the accuracy of the model. Initially, we created a two layer NN model, with two dense layers, followed by two activation functions, one at each layer. We tried two different types of combination for activation functions to see which one works better. The loss function that was applied to the data, was Mean Square Error (MSE). Below, is the sequence of the activation functions used:

1. Two Layer Neural Network model
 - Dense -> ReLU -> Dense -> ReLU
 - Dense -> Tanh -> Dense -> Logit

We calculated the Root Mean Square Error (RMSE) using the test data for the accuracy. The RMSE for the best fitting model, was about 0.107 and 0.11 for the wooden ties and concrete ties respectively. Since the mean value of the output data is about 0.15, the RMSE value is not considered to be sufficient. Figure [26](#) and Figure [27](#) represent the scatterplot of the model performance on the test data, for wooden ties and concrete ties respectively.

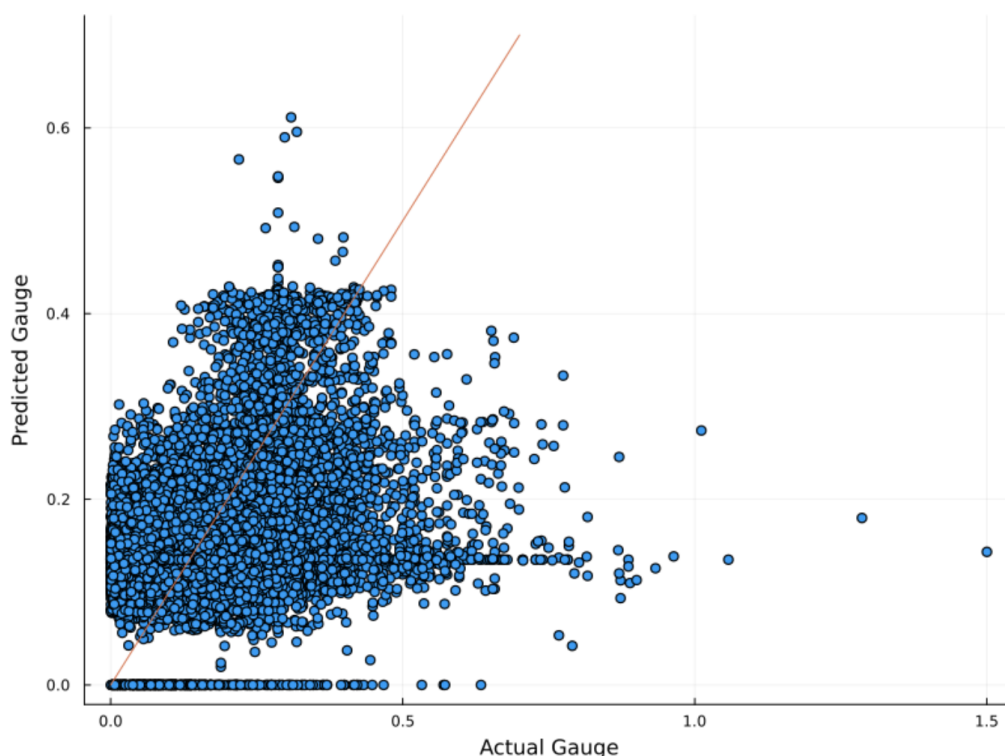


Figure 26: Two Layer Neural Network Regression for Wooden Ties

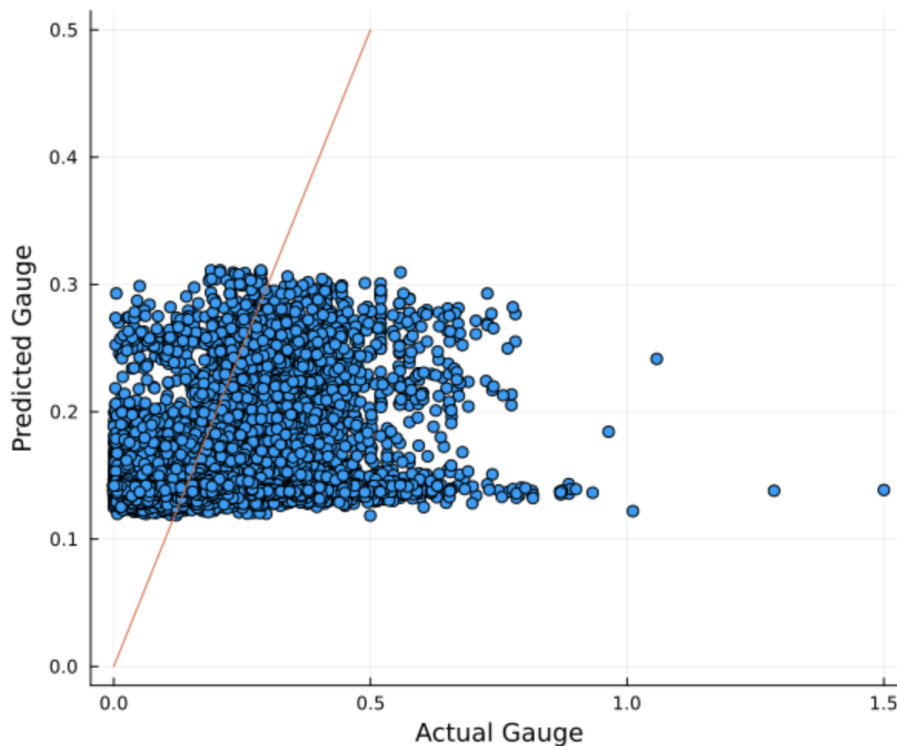


Figure 27: Two Layer Neural Network Regression for Concrete Ties

We followed two methods to improve the accuracy of the model. First, we deleted some of the input values that based on the results of the exploratory analysis phase of the project, had a lower impact on the gauge, including Rating, Frac_Def_TiePlate and Frac_Def_Fasteners. Then we Tried to label the data as “bad and”good” samples. To do so, we set a threshold for the output data that labels observations with gauge values more than 0.75, “bad”. Only 300 observations out of 180000 were labeled “bad”. Given this, we decided to continue with gauge values, instead of labels. None of the methods mentioned above, provides improvement in model accuracy.

The second approach is similar to the first one, but it has three dense layers instead of two, to increase the complexity of the model and lead to a better performance. However, the results were close to the results of the first approach, and the accuracy is still very low. We calculated the RMSE for the test data to evaluate the model performance. The best three layer model has an RMSE of 0.104, which is insufficient. We also used two different combinations of activation functions, which is listed below:

2. Three Layer Neural Network model
 - Dense -> ReLU -> Dense -> ReLU -> Dense -> ReLU
 - Dense -> Tanh -> Dense -> Tanh -> Dense -> Logit

Figure 28 illustrates the scatter plot of the best fitting three layer NN model for the test data. The results of the regression section shows that the models are constantly underestimating, and the predicted values are always lower than the actual values.

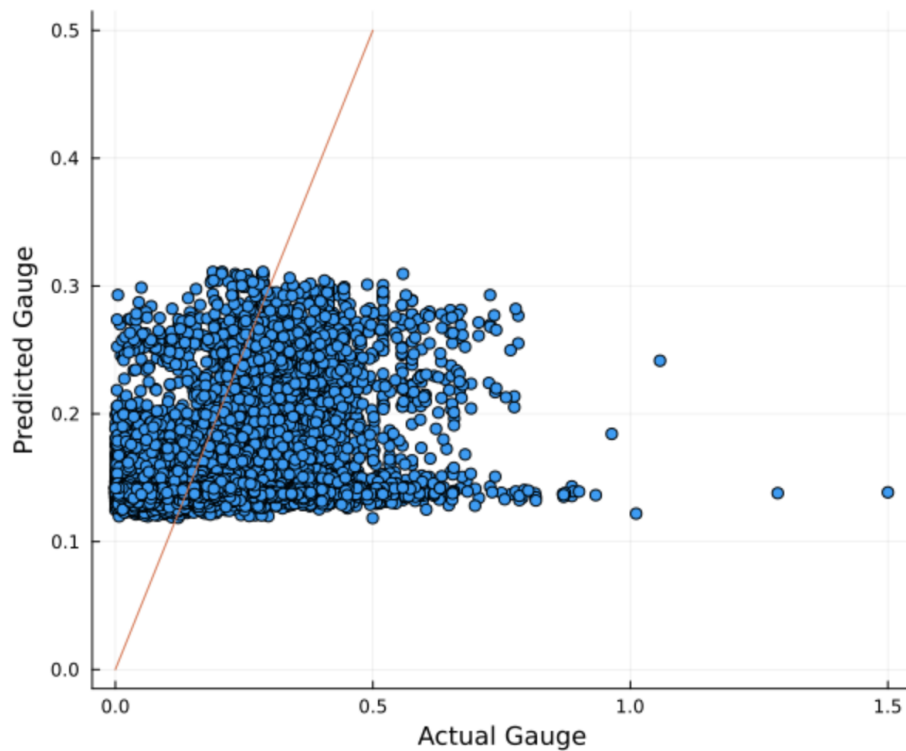


Figure 28: Three Layer Neural Network Regression for Wooden Ties

Decision Tree

The third approach is regressing the input values and the gauge with a Decision Tree model using an existing julia package, "DecisionTree". In the first two approaches, we created the Neural Network model ourselves, meaning we defined the functions and put them together to form a model using the programming knowledge we gained during this course. In this approach and the next one, we used the functions defined within the packages to create the model. This method provided better results than the previous ones, however the accuracy is 0.097 which is still unsatisfactory.

Random Forest

The last approach is using a Random Forest regression model to predict the gauge. We used an existing julia package, DecisionTree, to create the model. This method gives the best results compared to the other models, and the accuracy is 0.095, which is not acceptable. Figure 4 and figure 5 represent the results for wooden ties and concrete ties respectively.

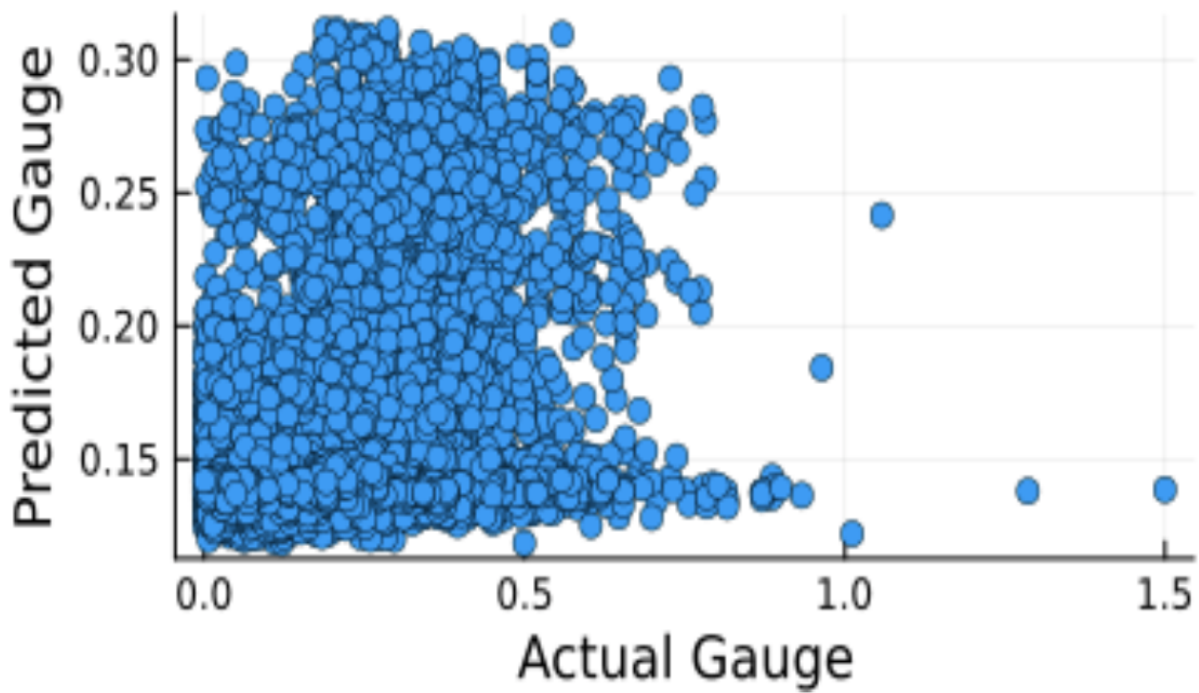


Figure 29: Random Forest for Wooden Ties

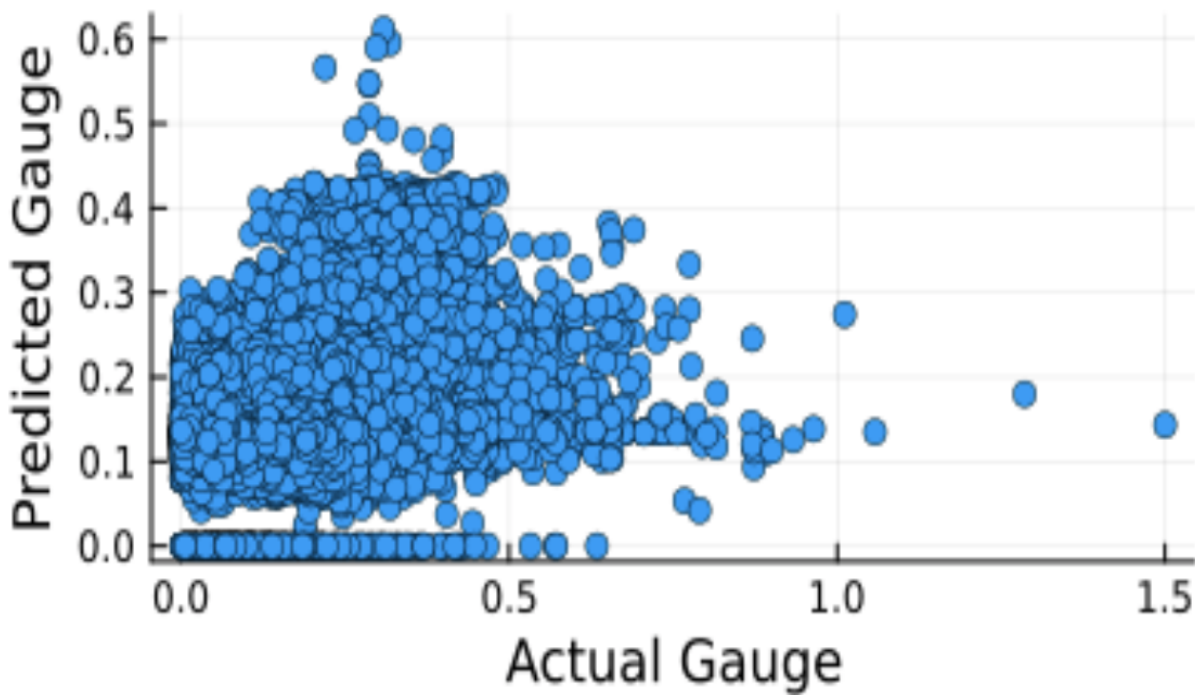


Figure 30: Random Forest for Concrete Ties

Classification

After failing with the regression models, we tried classification models. For having a classification model we need to create labels for the data set. As it was mentioned in the regression section, setting a threshold which considers gauge values more than 0.75 to be “bad”, did not provide enough samples to work with. The deal with this issue, first, we changed the threshold to 0.5. The new “bad” sample size was 3000 for wooden and about 700 for concrete ties, which is better than the previous one, but still not enough compared to our dataframe size, which is 180000 samples for wooden and

67000 for concrete. Then, we randomly selected 3000 and 700 samples from the “good” observations for wooden and concrete, respectively. Lastly for each material type (wooden or concrete), we combined good and bad samples, and created a new dataframe, which was no longer imbalanced. We shuffled the rows, using the shuffle function to make sure there is no pattern in the new dataframe. Same as the NN models, we split the data into two dataframes, 80% as the training data and the remaining 20% was used to test the accuracy of the model. We have a total number of 6000 samples for the wooden ties, including 4800 train samples, and 1200 test samples. These numbers are 1400 for the whole sample of the concrete, 180 for the test data and 1120 for the train data.

Neural Network

As what we had for the regression part, we used our own three layer neural network model for predicting the Gauge as “Good” or “Bad” (0 or 1). The sequence of this three layer is Dense -> Tanh -> Dense -> Tanh -> Dense -> Logit. The loss function is the Binary Cross Entropy in this part. The accuracy which is the total number of correct predictions divided by total number of data, is 84% for the Wooden ties and 64% for the concrete ties. The results for the Wooden ties can be acceptable but it is not really good for the concrete. Therefore, we also used Decision Tree and Random Forest methods for wooden and concrete, respectively. It is worth to mention that for all of these models in the classification part, the results of the models are not binary. Therefore, before calculation the accuracy, we quantized them to 0 or 1 using this logic that if the result is less than 0.5, we consider it as 0 and otherwise, it is 1.

Decision Tree

For the Wooden ties, we tried a decision tree model using DecisionTree package in Julia. The accuracy of this model after quantizing is 83% which is so close to our Neural Network model.

Random Forest

For concrete ties, we considered a random forest model using the same package in the Julia. The accuracy of this model after quantizing is 70% which is better than our Neural Network model.

The figure (Sadaf) shows the results for the Wooden and concrete ties in the NN, Decision Tree and Random Forest models.

Discussion

The goal of this project was to find a relationship between the track components’ health conditions, and the track geometry. In the exploratory analysis phase, we learned that there is some extent of correlation between some specific track elements, and the gauge values. Based on this finding, we implemented multiple regression models to predict the actual values of the gauge based on the track components’ conditions. As discussed in the regression section, we were not able to achieve this goal. However, we were able to predict the defective gauge values using the classification method. Since the motivation of this research is to make the maintenance scheduling more efficient for railroad companies, we only need to know if there is a defective situation somewhere in the track or not, and we do not need to know the actual values of the gauge. Given this, we are confident to say we were able to achieve the goal of this project by identifying the hazardous gauge values. Further work can be done on defining a more precise threshold for the model, to be able to prioritize the maintenance required for different gauge conditions. Currently, we have two main categories, defective, and healthy gauges. It can be expanded to different levels of defectiveness.

Neural Networks, Decision Trees, and Random Forests use different mathematical calculations to regress the data, and each of them has its practical applications. We tried all of them to account for all of the possibilities that exist regarding finding a meaningful relation between the chosen track features, and the gauge. Unfortunately, none of the regression methods could provide a satisfactory level of accuracy and as we can see in the corresponding figures, none of them illustrate any patterns. However, we achieved relatively good results implementing classification models. Besides having good results, we would discuss some of the possible reasons why the regression models did not perform as expected, to provide some insight for future projects. Our first hypothesis is that maybe there is not an efficient way of predicting the actual gauge values based on the track data we currently have. To be more specific, there may be other factors that have higher impacts on the gauge, or at least, the features we picked are not correlated enough to the gauge. One probable solution to this issue is changing the data set or collecting different types of data from the track, but sometimes this is not possible to obtain a whole new data set. Some of the items that have a potential impact on the gauge are listed below, which can be used for making the regression method work better:

1. Ballast condition
2. Lateral strength of the track provided by the shoulder ballast
3. The longitudinal strength of the track provided by crib ballast
4. Vertical loads including static, dynamic, and impact loads
5. Speed of the trains (It does not directly affect the gauge, but since the loads transferred from the wheels to the track are heavily dependent on the speed, it could be a good idea to study speed individually.)
6. A better solution could be collecting multiple data sets from the same section of the track over time. This helps monitor the track condition with the passage of time, which provides more insight into the pattern of the defections on the track, which in our case is the gauge. Adapting this method, we can track the gauge over time, and identify the sections of the track in which the gauge is getting wider aggressively, and by analyzing track features of those sections we can identify the elements with dominant impacts on the gauge, and new models could be applied to them to check for potential correlation, and ultimately having a working predictive model. It is also worth noting that if the current dataset had a time stamp, it was possible to compare the gauge situation with just one dataset.

We believe trying to predict the actual values of the gauge, is not possible using the current data. Based on that, we decided to implement classification models. We labeled the data into “bad” and “good” samples, and as mentioned before, the total number of bad samples was very low compared to the good ones. The fact that the data was imbalanced, could be another reason why we were not successful in creating a model with our initial data frame. Furthermore, we believe changing the threshold from 0.75 to 0.5 played a big role in having better results in the classification section.

We were able to answer our initial question using the classification method, and the errors we have are considered to be at satisfactory level.

References
