# Mid-Way Project Report CS 771A

## Classification of Railway Track Geo-Defects

*Team members:* Deepak Singh 151091; Mimisha Khamrai, 151098; Kritika Verma, 151096; Mridul Kanti Das, 151099; Archit Mantri, 14127.

***Introduction:*** As this is a Mid-Way Project Report for the course CS771A and Sir has instructed that this should contain the progress so far, list/description of methods we have tried, experimental results (if any), any roadblock we have faced during the project work and plans for the rest of the semester, we will provide all the above things by this Report by partitioning them into three following parts:

- What we have done up to now (which are completed)
- What we are doing now a days for the project (currently going on)
- What are the plans for the rest of the semester

We will provide the details of what we have done and brief about what we are doing and what we will do in the coming days.

***What we have done Up to Now??*** This part can be further partition into 3 parts:

**a)** Summary of the data and imputation of missing values
**b)** Data pre-processing (from unstructured data to structured data)
**c)** Method we have applied up to now and results.

**(a).**First of all we imported all the *.csv* files in **R Statistical Software**. After that we used 'summary' function in R and analysed the results. We observed that all the things were good except *Tonnage Data***.** In that we observed that for year 2007 and 2008 all the data on total Car/Train East/West were 0. But we observed that at these places we have non-zero values for total tonnage passed which seemed us as an error so we treated them as Missing Values. We removed the data for Year 2007 and 2008 and tried to fit a model from the available data. But after further analysis we also observed that some of the values are missing for Years other than 2007 and 2008. So firstly we tried to find out values missing for the Years other than 2007 and 2008. For imputation of these missing values firstly we tried to use linear regression analysis but due to bad (non-linear) patterns in scatter plot between tonnage passed and total number of Cars/trains East/West to West/East (To see the patterns anyone can use our Shiny App made by us) we were not able to use any of the above mentioned factors for imputation of missing values using Linear Regression Analysis. So we tried to see the pattern for total number of trains (sum of trains East to West and West to East) with total tonnage passed we saw the almost linear pattern. So we decide to use the total tonnage passed and total trains passed. But after fitting the regression line we saw that there was a large intercept term as 96.722 (which implies that even if the total passed from a point is zero our model will give 96 trains passed through that point which is not good). So we thought to switch to other method and looked for **Time Series Analysis** and after analysing the results we were satisfied by the obtained results except at some points where there are missing values continuously for 3-4 months. We again made those values as Missing Values. At last to find out these missing values and missing values in the years 2007 and 2008 we used the **Linear Regression Analysis.** After doing all we have the non-missing data in our hand. For more on finding missing values please have a look at our code in that we have written all these things as comments.

**(b).**After finding the missing values when we proceed to our project we felt a heavy need to pre-process the data. In this process we have removed the repeated defect on the same day and also the repeated defect by the milepost (in the FAQ they have given that any defect point within the 200 feet from the previous test's defect point will considered as same defect points and also the tests within the 7 days will be considered as same). For doing this we have firstly made unique defect positions, which were at least 200 feet away from each other, on a given line segment number. After that we assigned unique defect positions to each of the given defect positions. We thought that this pre-processing was necessary for any of the method we would like to implement. Apart from this we have done the pre-processing of the data based on the methods we were likely to implement. We have done almost all the things (pre-processing and implementation) of one of the method, details of the same we will provide in the next section.

**(c).**We have applied an extension of Gamma Process cited papers [1], [2], [3], [4]. In this we have used shape parameter of the as function of some variables. In our case these variables were the dependent variable which we wanted to use. In this method we will find the probability that a *Yellow* defect with given characteristics will turn into *Red* Defect after a specified time interval in days. We have used **shape parameter =**$c*t^{\beta X}$ where **c** is a constant and **X** is vector of dependent variables and **β** vector of coefficient of dependent variables (for this we cited [5]). By using this form of shape parameter we were able to use as many dependent variables as we needed. For parameter estimation we have used the **Maximum Likelihood Estimation** technique. In this we have assumed that property of each defect positions are independent to each other which is obvious assumption as the defect characteristic at a point will not affect the defect characteristic at different position. In this we also assumed that the deterioration process is different for each combination of Line Segment Number and Defect Type (have to fit 12 gamma process). So we divided our data into 12 dataset after that we have to do some processing of these datasets in order to work with them. These processing are related to how to use *Tonnage Data* and *Inspection Data.* So from *Tonnage Data* we have included *total tonnage* and *total number of Trains passed* from a given defect position. We have used the *Inspection Data* only in Pre-processing. At a defect position we collected the history of that defect position. And we found the number of inspections done in between two consecutive testing dates. We have taken only those observation for which the count of inspection is 0. This is valid as in case of non-zero count of Inspection between two dates $D_1$ and $D_2$ implies that in between $D_1$ and $D_2$ they have done the inspection but they were not able to detect a defect at that position which implies that there is some repairing in between $D_1$ and $D_2$. We have used the validation dataset to decide on the value of threshold probability. We have to use the same data set of *Tonnage* which we have used in training of the model. Before prediction for the given (By INFORMS) testing dataset we have tested our model on self-made *Testing Dataset* (partitioned the given training data in 70%, 15%, 15% as **training, validation** and **Testing).** Up to toady we have tested our model for *Line Segment Number 1* and Defect Type *Surface* and got the **accuracy** as **73%.**

***What we are doing??:*** After doing the above now we have focused ourselves to implement **Advanced Machine Learning Techniques**. We are trying to implement **Logistic Regression, Decision Tree, Perceptron and Cluster Analysis** to solve the above problem. For these methods we are citing [6], [7]. We have done pre-processing for **Logistic Regression** and **Decision Tree** and currently coding to implement these methods using Computer in **R Statistical Software** and to find the model accuracy. We will complete these methods by coming weekends. In case of **perceptron** and **Clustering** up to now we are wondering about how to implement these methods for our data set. But we have read the above mentioned reports and found some fair idea about implementation of these methods. We are expecting we will be able to implement these two methods by next Weekends.

***Plans for the rest of the semester:*** while discussing **Expectation Maximization [8]** topic in the class, Sir hinted about sequential data models and HMMs classification. And as in our data we have to analyse the Tag of the defect as the time passes, this can be seen as sequential data. So we collected more information about this method using Google Search. We found that it has a broad use in the area of Predictive Modelling. We found a thesis on ***Using Hidden Markov Models for fault diagnostics and prognostics in Condition Based Maintenance systems*** [9]. We also observe that **Xiaodong Zhang, Roger Xu and Team** has used HMM for ***integrated fault diagnostic and prognostic approach for bearing health monitoring and condition-based maintenance*** **[10]** (only can be accessed from IITK Network). **Miloud Sedira and Ahmed Felkaoui** has used HMM in ***Rotating machinery Diagnostic Using Hidden Markov Models (HMMs)*** **[11].** One of our team member is reading these papers and will report shortly in 2-3 days that will we be able to use the HMM for our problem or not. Based on his affirmative answer we will collectively try to implement HMM for our Problem. For prediction of testing data given (By INFORMS) we are thinking to use **Ensemble Classifier.** To get ideas about using **Ensemble Classifier** we are following the class notes of 26-Oct-2016 and also this report [12] and this paper [13] by ***Zhi-Hua Zhou*** to get an inspiration about why to use **Ensemble Classifier.**

*As here we are given very limited space to write our work so if anyone want to see our work in detail we would like to suggest to have a look at our GitHub Repository. **(For codes please visit GitHub repository).***