APPLICATION OF MACHINE LEARNING ALGORITHMS TO PREDICT PAVEMENT PERFORMANCE

By:
Angad Arora
Mayank Gupta
Nidhi Desai

Advisors:
Dr. Roshanak Nateghi (Purdue University)
Dr. Arghavan Louhghalam (UMass Dartmouth)

Introduction and Motivation for the current Study

We are aware of the greenhouse emissions and rise of global warming. The contribution of transportation sector in the total US greenhouse emission is 28% (EPA, 2016). The U.S. roadway network poses significant impact on the environment, having more than four million miles of public roads, and with generation of 6526 million metric tons of Carbon Dioxide (CO_2) and total fuel consumption of nearly 168 billion gallons (FHWA, 2015). Few of the factors contributing to CO_2 emissions involves pavement condition, design, and its characteristics as they affect the fuel consumption of the vehicle (Gyenes and Mitchell, 1994; Chatti and Zaabar, 2012). The need for achieving sustainable transportation system requires maintaining nation's roadway network in good condition. As per the estimates of U.S. department of transportation, maintenance of national highways at their current condition requires an expenditure of nearly $95 to $109 billion during 2014 and 2020 respectively. There is limited financial resources from the government for infrastructure maintenance. Therefore, it is necessary to allocate resources where they are the most required. Predictive models which predict the condition of the road will help with identifying roads with higher environmental footprints and which requires maintenance (Louhghalam et al., 2016). There are various factors that contribute to the pavement performance, most prominent of them being weight of the vehicle, climatic conditions, material properties of pavement and vehicular traffic (Louhghalam et al 2016). The excess weight of the vehicle makes the contact slope even steeper, that correlates to the effort of the vehicle needed to move over the road leading to excessive heat dissipation (Louhghalam et al 2014). This loss of energy comes at the expense of excessive fuel usage leading to more emissions. Climatic conditions considered in the study include temperature, precipitation, wind speed, moisture etc. The factors include in the material properties of pavement are top layer stiffness, top layer thickness, subgrade modulus and viscosity (Louhghalam et al., 2014). There are other factors such as average daily traffic especially average daily truck traffic, vehicle speed, which affects the pavement performance. The measure of pavement performance is descried in terms of IRI (Roughness Index).

The current literature has come up with physical models towards predicting IRI but while doing so, they have incorporated limited number of factors. The motivation of data driven study is to verify that the factors considered in the physical models in previous studies are in fact backed up by predictive analysis too. The data set considered for our project is taken from Long-Term Payment Performance (LTPP) dataset (https://infopave.fhwa.dot.gov/) which is a huge repository of data collected over two decades.

Data Visualization:

The variables that were available can be generalized to the broad categories of:

Structure : Includes Maintainance, Roadway Information, Pavement Layers, Joints, Shoulders, Drainage , Reinforcement, etc.

Climate: Includes Precipitation, Wind , Freeze, Moisture, Temperature, Water Table, Solar Radiation, Humiidity

Performance: Includes Roughness Index, Logitudnal Profile and Temperature, Deflection, Distress

Traffic: Weight, Kesal Year, Volume

Selection of the Response Variable:

To measure the roughness index of the road, there were several variables that are mentioned as a part of the data. In the earlier times, the RMSVA (Root Mean Square Vertical Acceleration) was used to measure the Pavement Performance, but with time IRI became the main scale of measurement.

However, in the LTPP data there were 3 types of IRI mentioned viz Mean IRI, Left Wheel IRI and the Right Wheel IRI. Our study considered mean IRI as the performance parameter as the values of the other two IRI are highly collinear with mean IRI as shown in figure 1 The labels indicate SHRP IDs corresponding to different pavement sections.

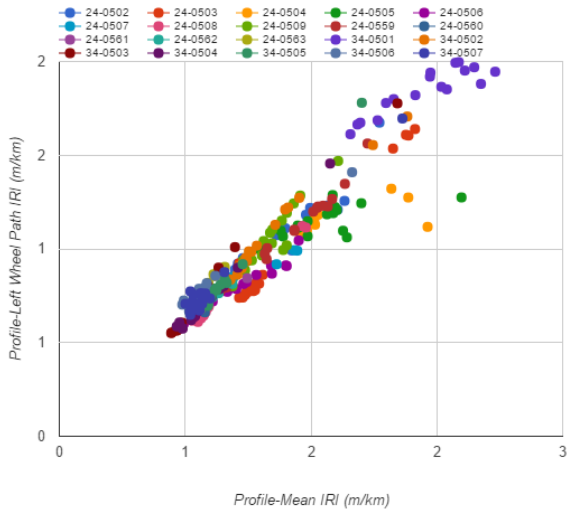

Figure 1: Graph showing collinearity between mean IRI and left wheel IRI

Also, the initial analysis of the IRI is done to get the possible range of values and also with respect to time as shown in figure 2 (Left) and 2 (Right) respectively. The study gathered Mean IRI data from PAVEMENT MONITORING MODULE within the LTPP dataset. in figure 2 (Right), the labels indicate SHRP IDs corresponding to different pavement sections.

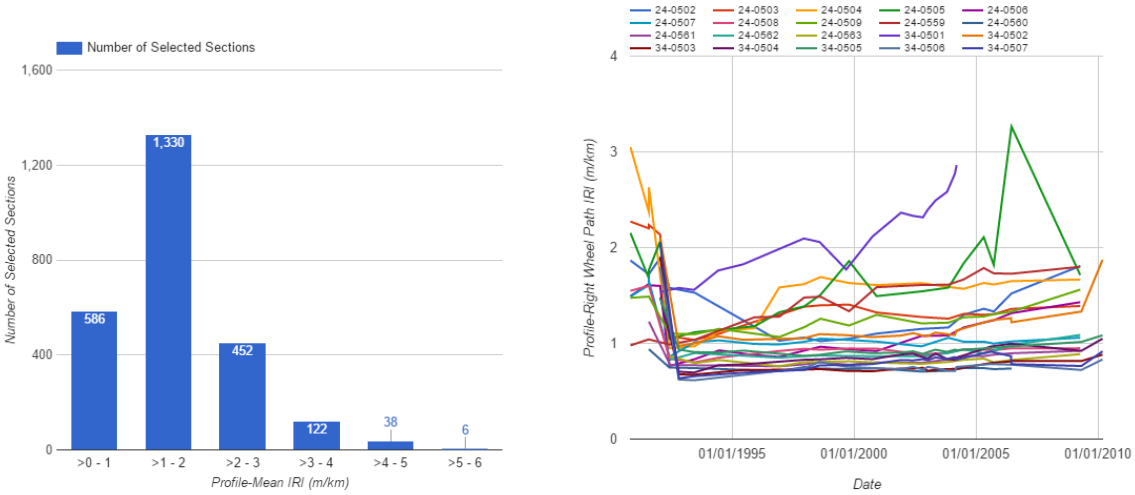

Figure 2: (Left) Shows the histogram of IRI values across Texas State, (Right) shows the variation of IRI with Time.

Hence, the scope for this study is to have data driven analysis to come up with the significant variables that effect the IRI. In the next section, the choice of variables used for this study is mentioned. Initial analysis of mean IRI (MRI) over the state of Texas is shown in Figure 5 illustrating the areas with maximum, minimum and average MRI range.

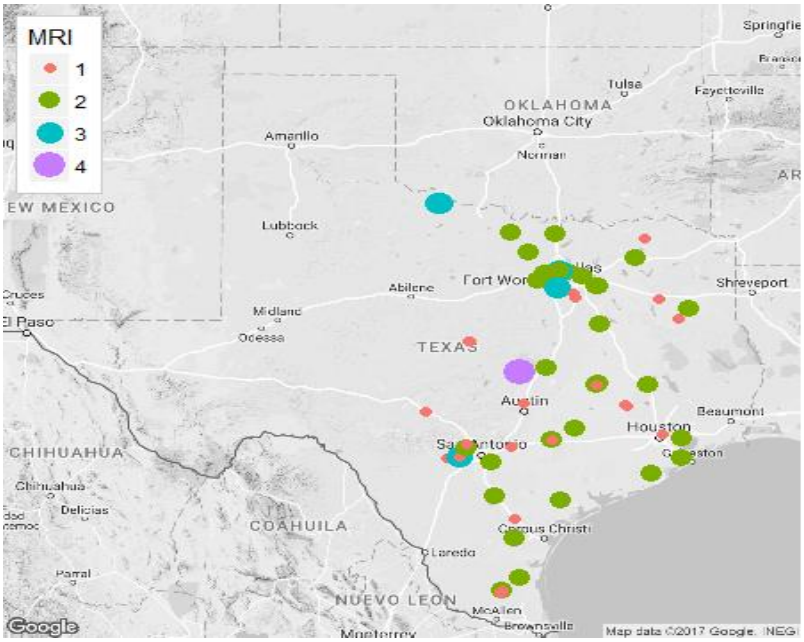

Figure 3: Shows the distribution of MRI over the state of Texas, 1 denotes values between 0-1, 2 denotes values between 1-2, 3 denotes values between 2-3 and so on.

Choice of variables for the Study:

The following table list the names of the variables we are considering for this study, as well as their mean, minimum and maximum values.

Variable Name	Minimum	Mean	Maximum
Mean IRI	0.536	1.3810	3.927
Slope Variance	0.6193	4.358	25.56
Thickness	10.2	26.004	43.4
Day Precipitation	0	0.721	28.4
Mean Day Temperature	-3.9	20.229	31.4
Max Day Temperature	4.3	26.64	38.7
Min Day Temperature	-12.3	13.66	25.6
Max Day Wind Speed	0	9.30	24.10
AADTT	30.0	736.99	2770.0
Mean Day Humidity	30.0	67.131	96.0
Kesal Year	4.0	352.278	4326.0

Skewness in the data can be observed by the violin below. The following shows the distribution of the variables data through violin plots.

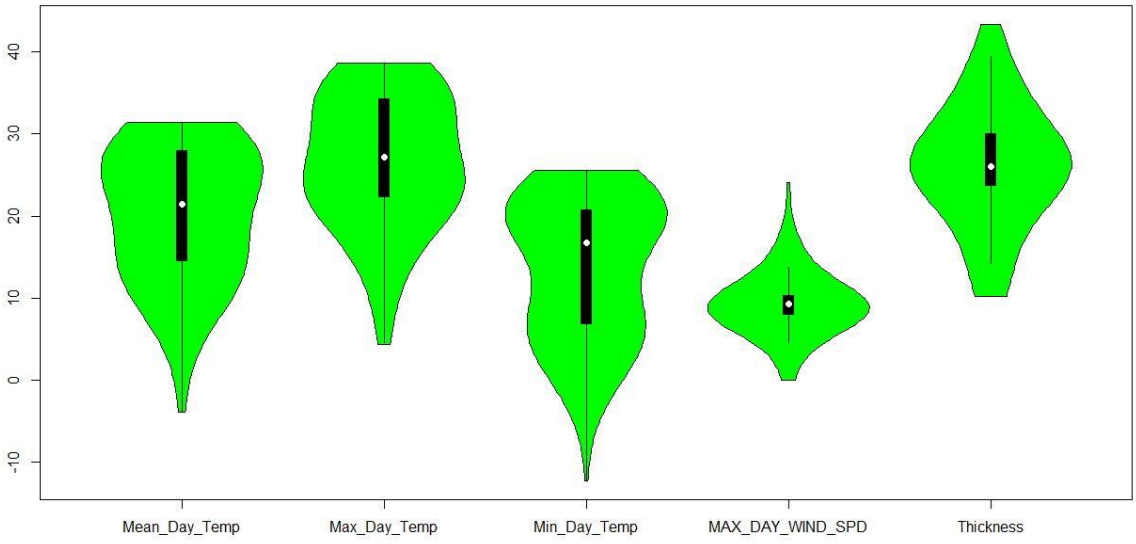

Figure 4: Violin plots for temperature, wind speed and thickness variables.

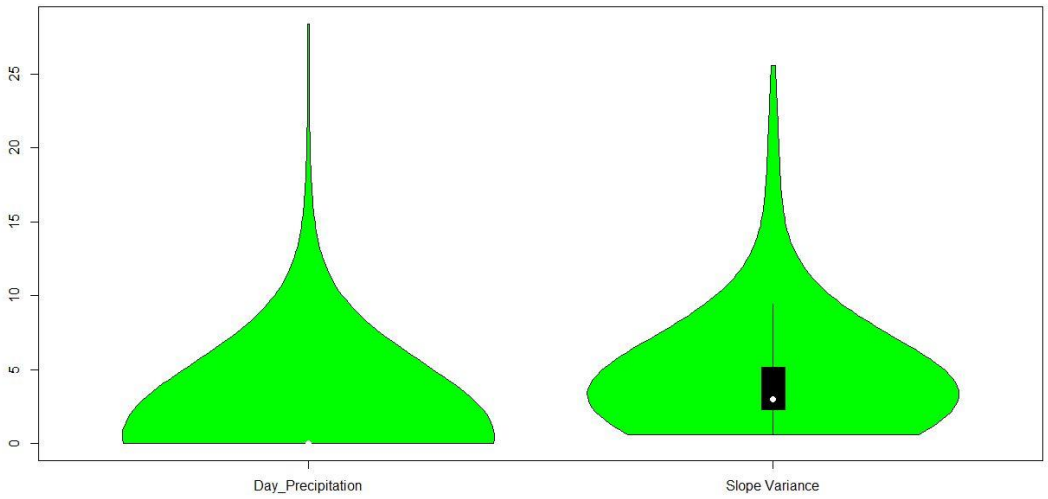

Figure 5: Violin plots for day precipitate and slope variance variables.

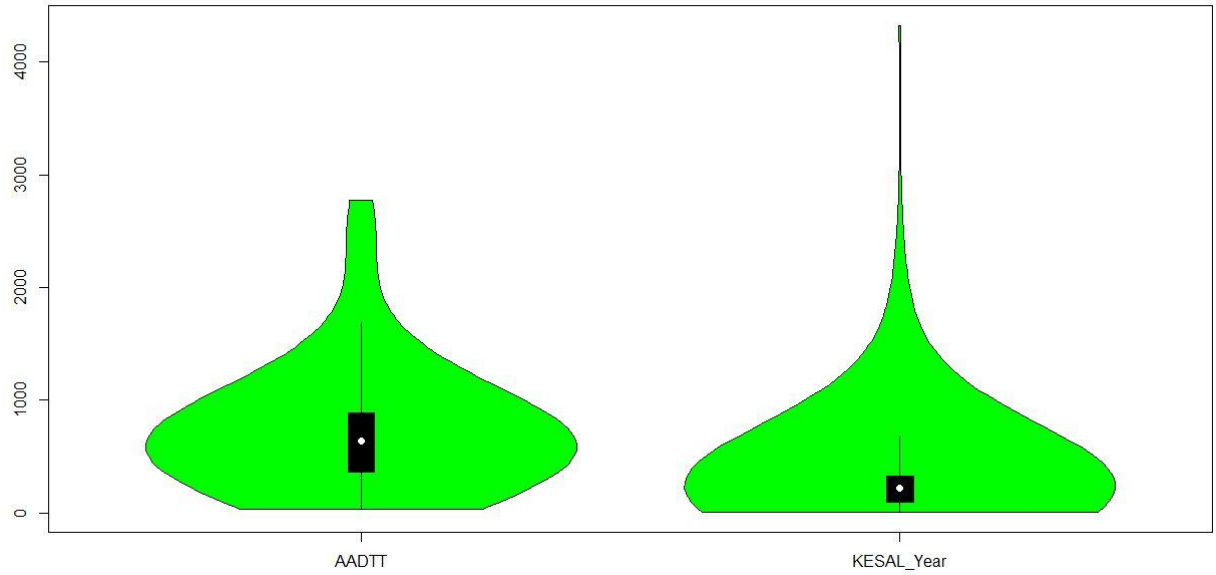

Figure 6: Violin plots for AADTT and Kesal Year variables.

The explanation of the various variables is given below:

1: Speed: Dissipated energy in suspension due to roughness must be compensated by the engine power to maintain a constant speed. Figure 7 shows initial level analysis for the speed variation in the state of Texas. The study gathered Mean IRI data from SPECIFIC PAVEMENT STUDIES MODULE within the LTPP dataset.

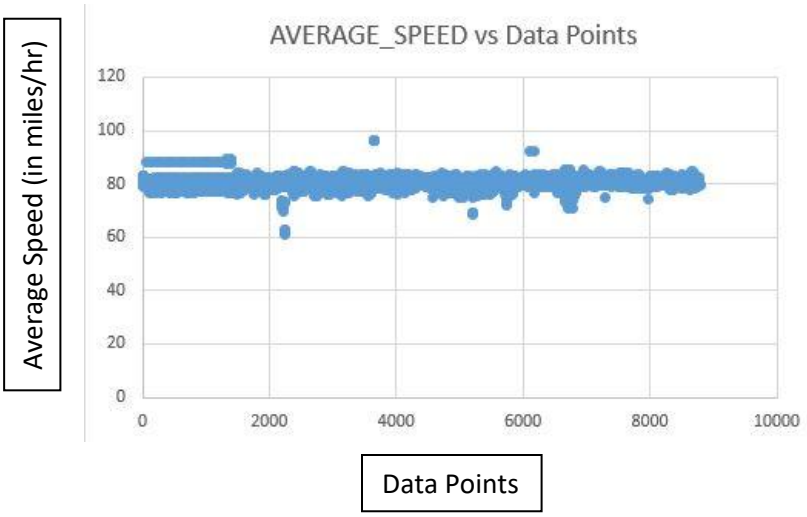

Figure 7: (Left) Shows the bar plot of average vehicle speed across Texas State, (Right) shows the variation of average speed with 8000+ data points.

To get the accurate speed limits corresponding to individual SHRPID (unique ID for the pavement sections), data from the Coordinate axis in LTPP is combined with the ArcGIS data https://www.arcgis.com/features/index.html/ to get the exact longitude and latitude positional match.

2: AADTT (Average Annual Daily Truck Traffic):

Data information from the daily truck traffic can infer about the average weight or loading a pavement has to go through and hence it could have a significant relation with the MRI (Louhghalam et al., 2016).

$$\frac{Eh^3}{12}\frac{d_W^4}{dx} + mV^2\frac{d^2w}{\partial x^2} + kw = p$$

Here the pavement is modeled as a viscoelastic beam on an elastic foundation subjected to an axle load travelling with a constant speed V in a steady state condition. The differential equation of beam's displacement w, in a moving coordinate system is given above.

As can be seen from the equation the vehicle weight and speed have a very strong relation with the pavement surface characteristics and it will be worthwhile to include such parameters in this study.

Figure 8 shows some of the descriptive data analysis of the AADTT (Average annual daily Traffic). It is clear from the initial analysis that the load has been on an ever increase on the pavement and hence the future pavement should be designed based on twin challenges of improving performance (MRI) as well as accommodation of additional load. The study gathered AADTT data from TRAFFIC MODULE within the LTPP dataset.

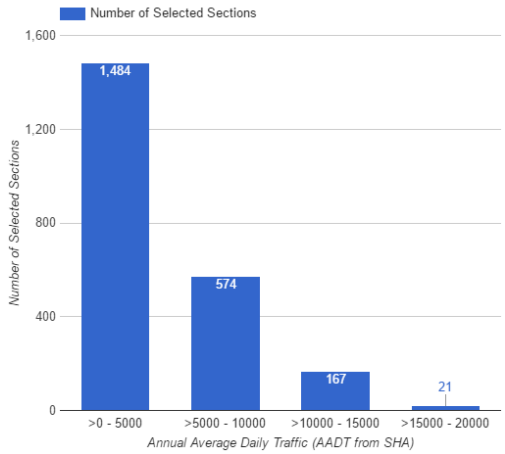

Figure 8: Shows the bar plot of average annual daily truck traffic across Texas State

3: Total Pavement Layer Thickness:

This variable is indicative of the pavement layer total thickness. As in equation (1), the variable h corresponds to the total thickness of the structure. It visualizes the thickness has direct relation with the energy dissipated by the vehicle (Louhghalam et al., 2016) :

$$\delta\,\varepsilon \propto \left(V^{-1} * P^{2} * E^{[-1/4]} * h^{[-3/4]} * k^{[-1/4]}\right) \ldots\ldots (1)$$

$\delta\varepsilon$ is a parameter directly related to dissipation and the equation above necessitates the need to add this parameter to the data driven study. There have been pavements that are old in construction, so every time, the repair happens, another layer is added on the pavement and that naturally makes old pavements to have more thickness as compared to recent constructions. The thickness is the summation of base and sub-base layers and is measured in inches. Initial exploratory analysis is shown in figure 9. The study gathered thickness data from SPECIFIC PAVEMENT STUDIES MODULE within the LTPP.

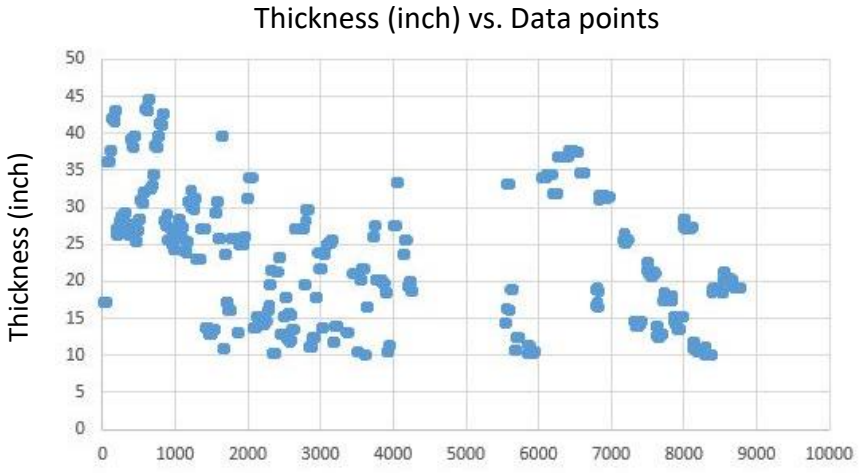

Figure 9: Shows the scatter plot of base thickness among pavements in the state of Texas with the data points.

4: KESAL_year:

ESAL here stands for Equivalent Single Axle Load, KESAL is 1000 ESAL. Therefore, KESAL_year is estimate of 1000 ESALs/year. This variable deal with smaller vehicles which were not accounted for in earlier variable (AADTT which accounts for higher axle traffic). As vehicle weight is an important parameter to get the loading condition for the pavement, it becomes imperative for this study to incorporate all kinds of loading on the pavements.

These ESAL estimates are provided only for sites which have an acceptable sample of axle load measurements contained in the LTPP database in the indicated year. The axle load sample is expanded to an annual estimate using a time-based multiplier. Descriptive analysis of this variable is shown in figure 9. The study gathered KESAL_year data from TRAFFIC MODULE within the LTPP dataset.

5: Slope Variance: The slope variance is the measure of unevenness of the road. This unevenness can be measured using mechanical profilograph. Slope variance is the second spatial derivative of height.

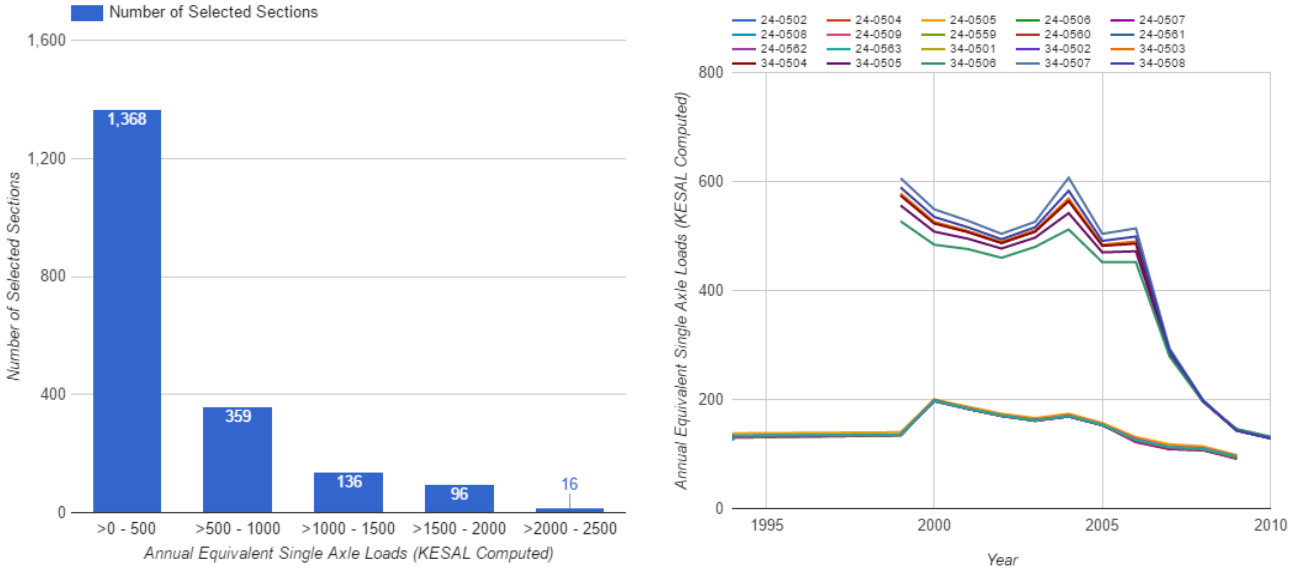

Figure 10: (Left) Shows the bar plot of equivalent single axle load/1000 across Texas State, (Right) shows the variation of the same with time (annually)

6: Climatic Factors: The data for these variables is gathered from CLIMATE MODULE within the LTPP dataset.

Temperature:

The literature has stresses upon the temperature dependence of dissipation. The literature used activation energy concept for the relaxation time as analogous to dissipation.

$$\log_e a_T(T) = U_c[\frac{1}{T} - \frac{1}{T_{ref}}]$$

$a_T(T)$ is the shift factor accounting for the acceleration or deceleration of the relaxation time i.e. it accounts for the dissipation when the temperature T is different from a reference temperature T_{ref} and U_c is constant at 2700 K for concrete pavements.

This study incorporated Temperature data of 3 types Mean, Maximum and Minimum. The reason for having these as separate variables is that apart from mean, we are interested in extremes of the temperature because it is the range that determines the severity of impact as well as fatigue carrying capacity of pavement and hence can have significant effect on the performance.

Precipitation and Humidity:

This study included Precipitation and Humidity as factors because with these into account, the pavement will be subjected to moisture and it is worth noticing the impact of moisture on the mechanical properties of the Pavement and hence the performance. Exploratory analysis of Precipitation and Humidity is shown in Figure 11.

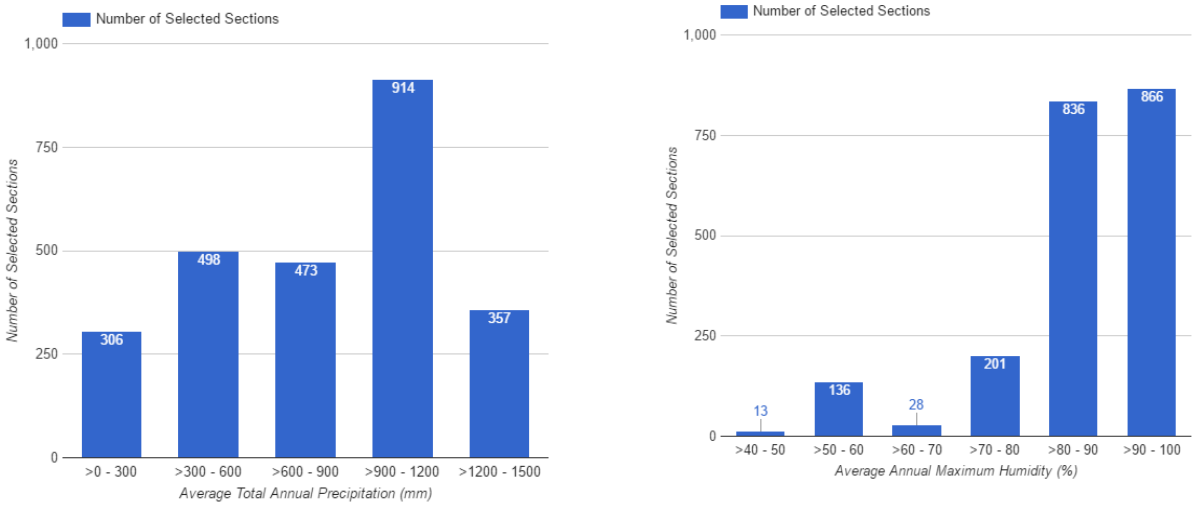

Figure 11: (Left) Shows the Average Annual precipitation and (Right) Average Annual Maximum Humidity with sections (the regions in the state of Texas)

Maximum Wind Speed:

If the wind speeds are significant, they can greatly affect the force due to air drag and hence impact the energy required to maneuver the vehicle. Figure 12 shows the average wind speeds with sections.

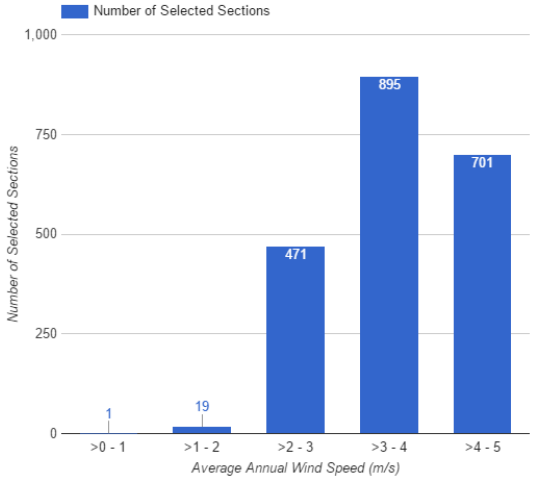

Figure 12: Shows the Average Annual wind speed with sections (the regions in the state of Texas)

Correlation among Variables:

The correlation plot (Figure 13) shows the strong correlation among all temperature variables and between AADTT and KESAL_year. The temperature correlations seem obvious as they are measures of a single entity. However, this does not mean that the study should ignore one because the work will also focus on the range of temperature and impact of extremities in the temperature scales. The same line of reasoning also goes for AADTT and KESAL_year, as both are a measure of traffic on the roads, but both give a different kind of loading, AADTT has more loading per area of the pavement whereas KESAL_year gives the total volume of small vehicles. For the rest of the variables, almost no correlation exists.

Principle Component Analysis:

Figure 14 shows the PCA biplot with the first two principle components. The first two principle components explain 50% of the data. As we can see from the PCA biplot, slope positively effects and the temperature variables negatively affect the first principle. The variables AADTT and Kesal Year are in same direction which suggest that these variables are correlated, which can be also seen from the correlation plot. The second principle component is negatively affected by AADTT, Kesal Year, Precipitation and Thickness.

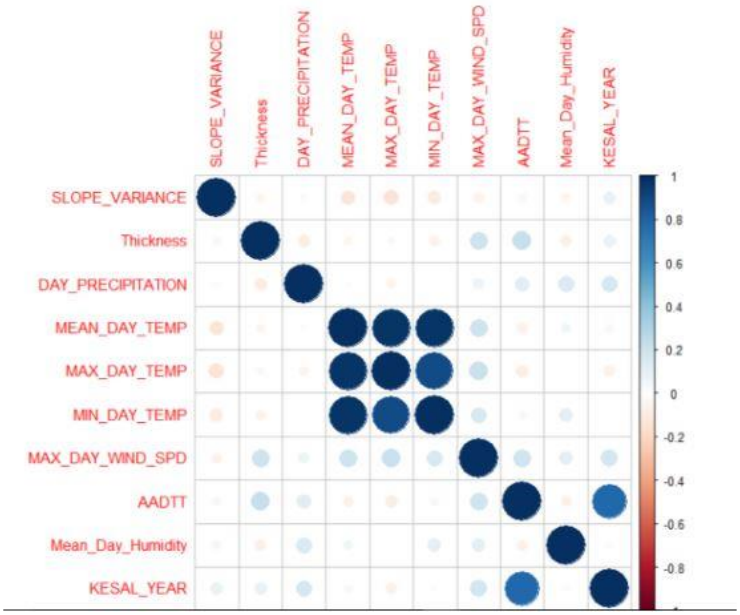

Figure 13: Shows the correlation plot among variables

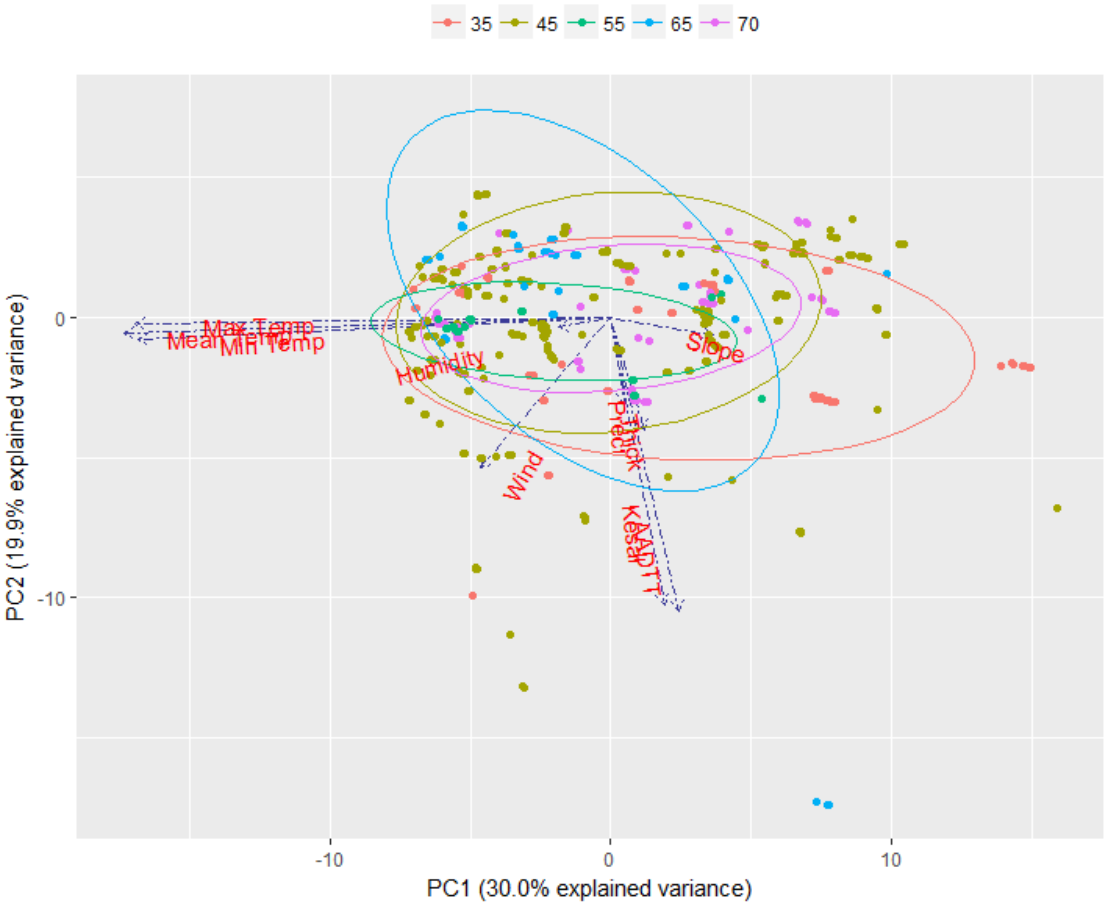

Figure 14: PCA biplot

Now we have completely explained the data set for our study and we will explain different methodology for predictive modelling in the next section.

Methods for predictive modelling

Predictive data analytics involves statistical approach of building models which learn from the available data. There are different ways through which the models can learn. It involves supervised learning and unsupervised learning.

In supervised learning, we have a set of variables which may be termed as inputs, $X^T = (X_1, \ldots\ldots, X_P)$ and these inputs impacts one or more outputs $Y = (Y_1, \ldots\ldots, Y_m)$. Our goal in case of supervised learning is to use the set of inputs that predict the values of outcomes. (Hastie, Tibshirani, & Friedman, 2009). In case of unsupervised learning, our goal is to draw the conclusions from the data set without the knowledge of response variable. (Hastie et al., 2009). The methods of supervised learning involves linear regression, general additive models, tree based models where we have knowledge of response variable and set of predictors effecting the response. Unsupervised learning involves clustering, principal component analysis etc.

In supervised learning, data learning process aims at finding f that will help in estimating Y well. We have;

$$Y = f(x) + \varepsilon$$

where ε is an irreducible error.

There are several ways to estimate this f which will be discussed later in the section. The method of estimating this f are broadly categorized into parametric modeling and non-parametric modeling. (James, Witten, Hastie, & Tibshirani, 2013)

In parametric modeling, we make assumptions on function f. For example, in case of linear regression, we assume that f will be linear. If we do not make any explicit assumptions about f, it is called non-parametric modelling. The aim here is to estimate the function which is close to the data points and isn't too rough or wiggly as well. The flexibility (complexity) of the models is dependent on whether the chosen method is parametric or non- parametric. The interpretability of the model decreases with increasing complexity.

In the next section, we explain all the different methods which can be used to estimate f.

Supervised Learning Models:

Linear Regression Model:

This a parametric model in which the response, Y is estimated as:

$$Y = X^T\beta + \varepsilon$$

This model is built under set of assumptions given below:

- Relationship between X and Y is assumed to be linear
- Errors have constant variance
- The observations are independent of each other
- Errors are independent and identically distributed and $\epsilon \sim N(0, \sigma^2)$
- Errors follow normal distribution (normally distributed)

The objective in this setting is to find the optimal value of parameters β using the methods of least squares which explains our data (Hastie, Tibshirani, & Friedman, 2009). The main disadvantage of linear model is that it is not flexible and we may be unable to capture the actual behavior of the data.

Another class of methods called shrinkage method in linear regression, where regression coefficients are penalized which help us to deal with the issue of multicollinearity and in selection of best subset of variables as well. These techniques are called ridge regression and the lasso. In these shrinkage methods, our goal is to minimize the function of the form: (Hastie, Tibshirani, & Friedman, 2009)

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \theta \sum_{i=1}^{p} f(\beta_i) \ldots. (2)$$

where θ is the tuning parameter. The value of $f(\beta_i) = |\beta_i|$ for lasso method and $f(\beta_i) = \beta_i^2$ for ridge regression. The lasso regression helps in selecting the best subset of variables as well by giving us the sparse solution.

Generalized Additive Models

In linear models, the assumption of linearity make these models quiet less flexible. It may be the case that linear model is not sufficient to explain the data every time. Therefore, there is a need to go beyond linearity and relax the assumptions of rigid model and allow non-linear relationship between predictors and response variable. This is done by introducing smoothing functions, which can capture which can help provide information that is not revealed using traditional linear models. The general form of the additive models is given as:

$$Y = \beta_0 + \sum_{j=1}^{P} f_j(X_j) + \epsilon$$

where $f_j(X_j)$ can be any smoothing function for a given predictor. (Hastie et al., 2009)

Multivariate Adaptive Regression Splines

It is a non-parametric method that is suitable for high dimensional data. It is a method in which piecewise linear basis functions of the form $(x - t)_+ \; and \; (t - x)_+$ are formed at each observation for all the predictors. Each of the basis function is taken into consideration and best set of basis function is chosen that produces the largest decrease in the training error (Friedman,1991).

$$Y = \beta_0 + \sum_{m=1}^{P} \beta_m h_m(X)$$

Where $h_m(X)$ is a function from set C of candidate functions or a product of two or more such basic functions.

But this may lead to overfitting in the data. Therefore, in order to avoid overfitting, we penalize the model for every extra basis function and finally chose the model which gives us the best value of general cross validation (GCV) which is defined as,

$$GCV(\theta) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\theta(x_i))^2}{(1 - \frac{M(\theta)}{N})^2}$$

where $M(\theta)$ is the effective number of parameters in the model that penalizes the value of GCV (Hastie et al., 2009).

Tree based Methods

There are some methods which involves making segments of the predictor space into a number of simple regions. These are called tree based methods. They are defined for both regression and classification problems. There are different methods tress based methods which can be employed. Classification and Regression trees (CART) employ recursive binary splitting approach which is a greedy method that tries to create the best split in each step. (Breiman et al., 1998). Tree based methods provide easy visualization of the result with better interpretability as compared to linear models. The regions $R_1, R_2, \ldots\ldots R_j$in the tree based methods are chosen in such a sense that it minimizes the value of residual sum of squares which is given as,

$$\sum_{j=1}^{J} \sum_{i \epsilon R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the j^{th} box.

The recursive binary splitting works in a sense that a predictor X_j is chosen and the cutpoint s is decided such that it splits the predictor space into regions $\{X|X_j < s\}$ and $\{X|X_j > s\}$ and then tries to minimize the RSS. The tree based methods may produce good predictions on the training set but they may overfit the data. To avoid overfitting, we tend to prune the tree such that neither the tree is too complex, nor it leads to overfitting.

The tree based method discussed above may suffer from high variance and hence we may adopt a method called bagging (James et al., 2013) where instead of choosing whole sample, we choose various bootstrapped sample and fit the trees on these samples and get the final prediction by averaging the over the different samples. If we calculate $\hat{f}^1(x), \hat{f}^2(x), \ldots\ldots \hat{f}^B(x)$ using B separate training sets, then final prediction is given as

$$\hat{f}_{avg}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x)$$

The method of bagging discussed above may suffer from a problem if the predictors are correlated with each other. Therefore, we use the method called random forest (Breiman et al., 1998) where we build our model by choosing a random subset of m predictors from p predictors each time where $m \approx \sqrt{p}$. These m different predictors are chosen for each bootstrapped sample and then the final prediction is done by averaging the prediction over all the samples just like in bagging.

There is another approach in tree based methods called Bayesian Additive Regression Trees (BART) which uses sum of trees to approximate the function $f(x)$ which estimates Y. The basic idea is that it imposes a prior which regularizes the fit by keeping the effect of individual tree small. In other words, BART works in a way boosting algorithm works where the models are learnt sequentially using the weak learners from the previous models. (Chipman et al., 2010). There are several advantages of using tree based methods; (i) Trees are very easy to explain to people, even simpler than linear regression (ii) Trees are easily interpretable and hence used where the aim to infer through data driven modelling.

Unsupervised Learning Models:

Principle Component Analysis (PCA):

PCA is an unsupervised learning method where the aim is to find the principle directions in the multi-dimensional space of dimension equal to the total number of variables. PCA does not consider the effect of the variables on the response variable rather which linear combination of variables has the most variance. PCA is particularly useful for finding patterns in high dimensional data. Therefore, it is a useful statistical technique which has applications in various fields such as face recognition, image, and video compression. Even though the data considered in this study is not high dimensional, applying PCA will help us understand which variables affect the variance.

The 1st principle component can be found by solving the following maximization problem:

$$\max_{u1} Var(u_1^T X)$$

$$s.t. u_1^T u_1 = 1$$

The subsequent principal components can be found out using maximization problem of the same form as given above.

Our Results:

To understand the relationship between the response variable i.e. Mean Response Index (MRI) and predictor this study used nine statistical models: Generalized Linear Model (GLM), General Additive Model (GAM), Classification and Regression Trees (CART),

Bagged CART, Random Forest (RF), Multivariate Additive Regression Splines (MARS), Bagged MARS, Bayesian Additive Regression Trees (BART). Ordinary Least Square (OLS) has not been used since the data is not normally distributed. This study considers two parameters for assessing performance of the models: cross validation error and R-square. R-square represents how good our model fits the training dataset cross validation error gives the predictive accuracy of the models.

The following table shows the results for various models.

Model Name	**10-fold Cross Validation error**	**R-square**
Mean only model	0.57	-
GLM	0.23	0.84
GAM	0.22	0.92
CART	0.22	0.88
Bagged CART	0.21	0.9
Random Forest	0.06	0.99
MARS	0.18	0.91
Bagged MARS	0.17	0.93
BART	0.1	0.99
SVM	0.26	0.8

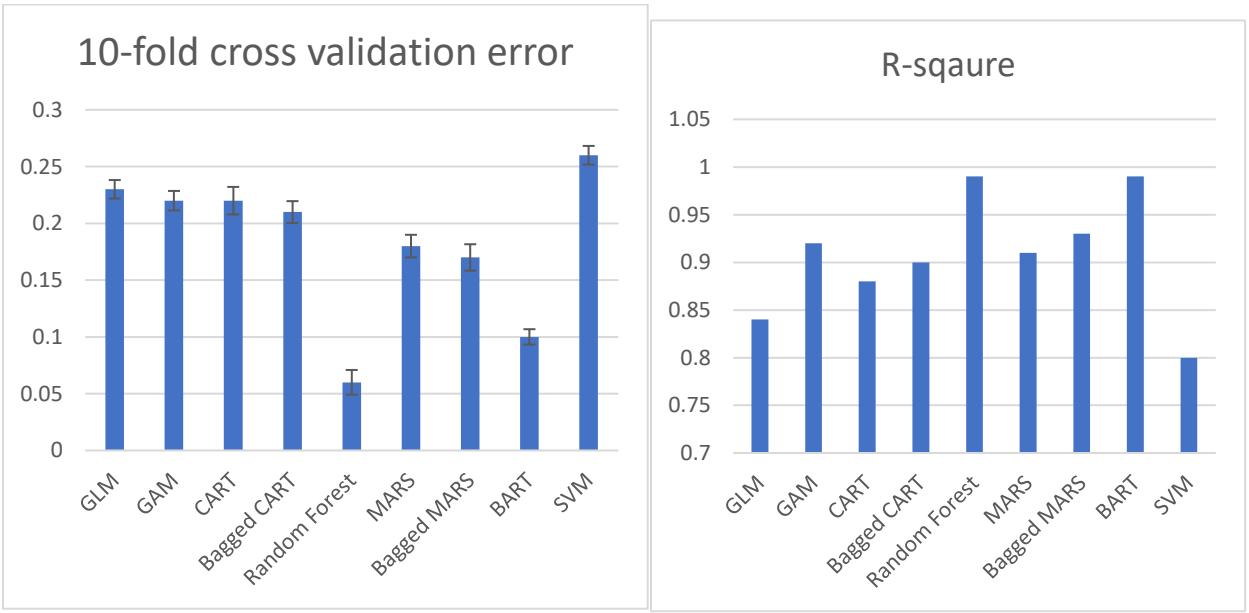

Fig 15: 10-fold cross-validation error and R-square for different models

The study took Random Forest as apt model as its CV error is lowest and R square is very high. We statistically compared Random Forest with all the models using Wilcox test which gives low p-values, implying the validity of Null Hypothesis, i.e. Random Forest can accurately predict the Mean IRI given the 11 response variables.

Model 1	Model 2	P-value
Random Forest	GLM	2.09E-05
Random Forest	GAM	1.81E-05
Random Forest	CART	1.02E-05
Random Forest	Bagged CART	1.04E-05
Random Forest	MARS	1.08E-05
Random Forest	Bagged MARS	1.08E-05
Random Forest	BART	2.17E-05
Random Forest	SVM	1.05E-05

Inference

The variable importance can be seen using VSURF. Variable importance (VI) is defined as follows:

Let $j \in \{1, \ldots., p\}$. For each out of bag sample we permute at random the j-th variable of the data. Variable importance of the j-th variable = mean increase of the error of a tree after permutation. The more the error increases, the more important is the variable.

VSURF results:

Variable selection using random forest (VSURF) ranks the variables by sorting the RF score of importance (VI - averaged from 50 runs) in descending order in the top left corner. VI mean is higher than the threshold value for all 11 variables.

The top right graph shows the standard deviations of VI and is used to estimate a threshold value for VI. This threshold (given by the red horizontal line) is set to the minimum prediction value given by a CART model fitting this curve (given by the green piece-wise constant function). Since all the variables have average VI exceeding this threshold, all will be retained. The sequence of variable importance is: Slope Variance, Kesal Year, AADTT, Max day temperature, Max day Wind speed, Thickness, Speed Limit, Mean day temperature, Min day temperature, Mean day Humidity, Day precipitation.

The bottom left graph provides variable ranking based on better interpretation. We see that the error decreases quickly. It reaches near minimum when the first nine true variables are included in the model and then it remains nearly constant. The selected model contains all variables except Day precipitation.

The bottom right graph provides variable ranking for prediction purpose which involves Slope Variance, Kesal Year, AADTT, Max day temperature, Thickness, Speed Limit and Mean day humidity.

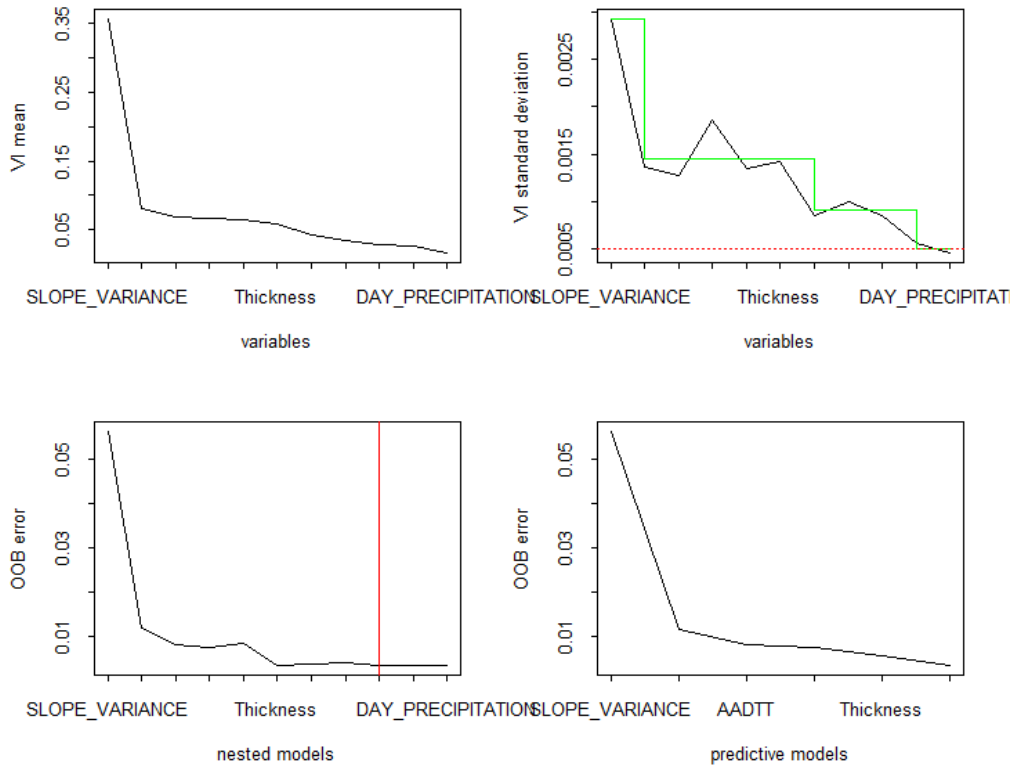

Fig 16: Variable importance plot for Random Forest using VSURF

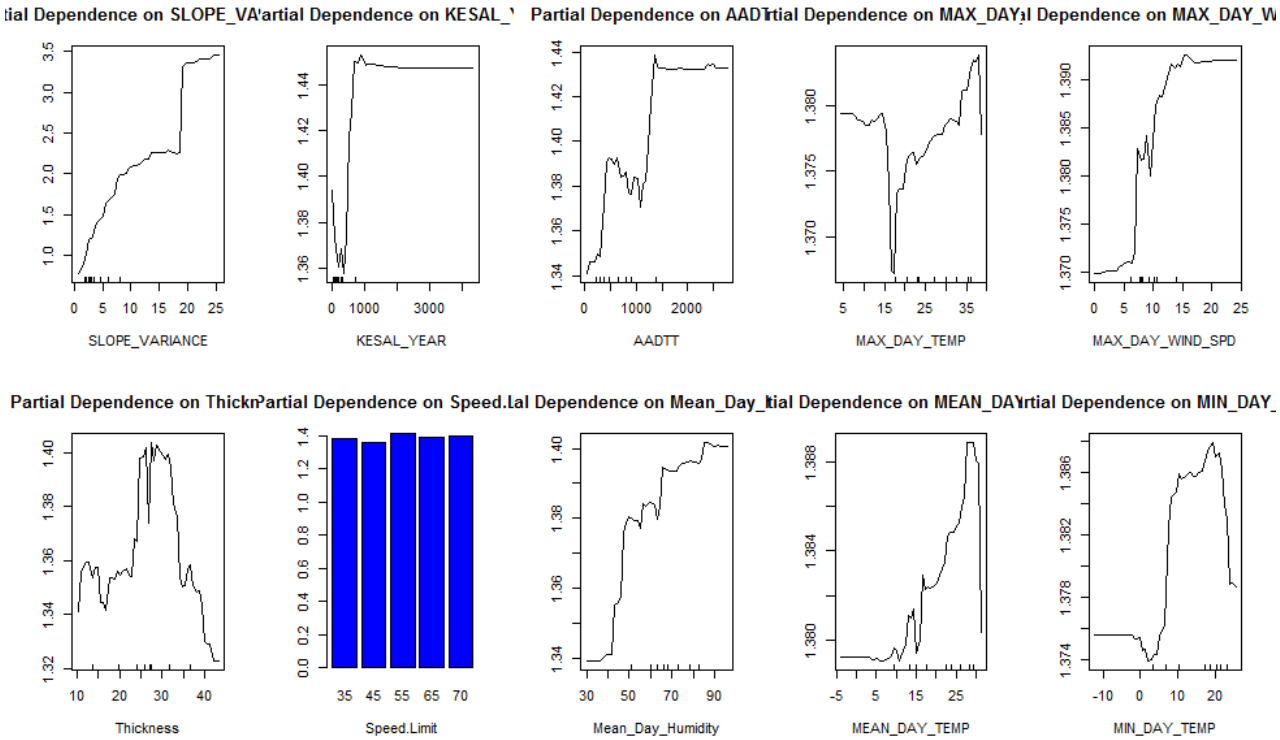

Fig 17: Partial Dependency plots

Inference from Partial Dependency Plots:

The following points can be understood using partial dependency plots:

A: Slope relation with MRI: As the slope increases, the energy required by the vehicle to overcome more resistance (high slope) increases, which is in direct correlation with MRI, hence MRI also increases as shown on the Partial Dependency plots.

B: KESAL_YEAR with MRI: The MRI increases sharply with the increase of small single axle vehicles.

C: AADTT with MRI: As the AADTT increases, the MRI for the pavement increases. There is a little period in which the MRI seems to remain constant but with further increase in heavy axle traffic, the MRI increases sharply and then goes constant. This may be because the sections might become immune to any further increase in MRI or due to maintenance. This trend can be partially explained through pavement design and the data collection methodology. The data was collected for both city roads (where the heavy axle vehicle traffic on an average is low) and highways (where the heavy axle vehicle traffic on an average is high). Initial increase in MRI is explained by pavement characteristic as the pavement is designed to perform better under a certain load. MRI goes constant for some period because the roads prone to less heavy vehicle traffic may have obtained their optimum operating conditions, but as AADTT further increases (this is specific to increase of heavy axle traffic on the Highways) the MRI for those pavement increases as the high-speed pavements at that point may still not have achieved their optimum operating conditions.

D: Temperature with MRI: From the 3 plots of temperature (Min, Max and Mean) vs MRI. It can be easily inferred that optimum mean temperature range is (-5-15 ˚C), minimum temperature range (-15-5 ˚C) and maximum temperature range is (15-20 ˚C). As the temperature increases the MRI increases, hence the hotter conditions have much severe impact on the MRI.

E: Wind Speed with MRI: The plot clearly shows a direct relation – increasing wind speed increases the MRI. This can be explained from earlier mentioned case as the wind speed increases, the Air Drag Force increases, that makes the vehicle do more effort for maneuver, hence the dissipation increases which is directly related to the MRI.

F: Thickness with MRI: The plot can be explained by the data collection by dates. Initially with lesser thickness of pavement layer, the MRI is good, this is primarily for the newly constructed pavements, that have better material properties as compared to older constructions. This is also confirmed from the latter part of the plot, as the MRI decreases with increasing thickness layers, this is specifically for older pavements, that have been refurbished over time by adding additional layers of material.

G: Humidity with MRI: The plot gives very direct relation as the humidity increases, the MRI increases, this is true from the fact that the moisture content will impact the layer properties of the pavement and hence may deteriorate the performance.

For the chosen model, we show finally show the graph of actual vs fit and actual vs predicted to see how well our model worked.

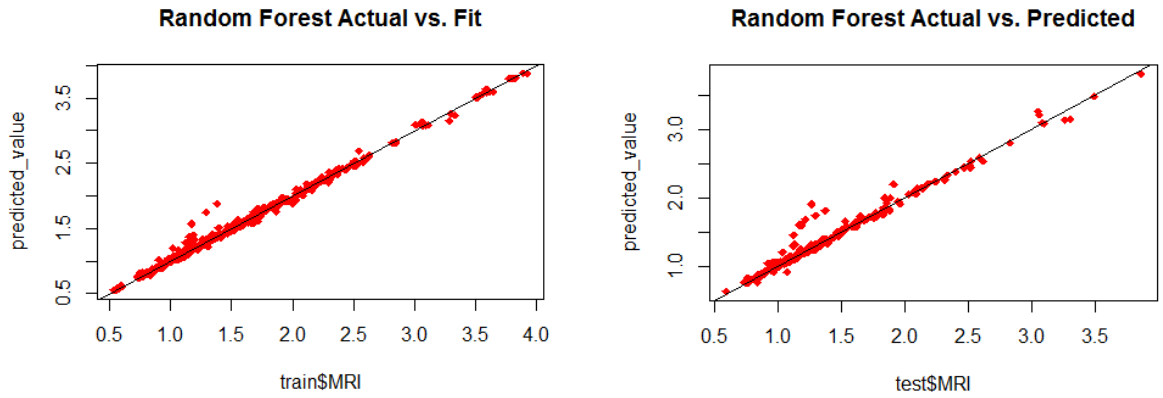

Fig 18: Plot of actual vs fit and actual vs fitted

Summary and further research

The study has been done for the state of Texas using the data across different roads in the state. We fitted various parametric and non-parametric models and based on the prediction accuracy and the value of R-square, random forest seems to be the best model that explains the data. The list of important predictors is also reported that explains the variation of mean MRI. The study has just considered one state where there is not much variation in the climatic conditions. To perform the sensitivity analysis for the climatic conditions, it would be better if more observations from across the regions of U.S are considered. The dips in the MRI may be because of regular maintenance of roads. It will be interesting to exclude the observations corresponding to the maintenance period and observe the behavior of predictors with MRI.

References:

EPA. (2016). Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990-2014. Www.Epa.Gov, 1–481. https://doi.org/EPA 430-R-12-001

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Elements, 1, 337–387. https://doi.org/10.1007/b94608

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R. https://doi.org/10.1007/978-1-4614-7138-7

Louhghalam, A., Akbarian, M., & Ulm, F.-J. (2014). Scaling Relationships of Dissipation-Induced Pavement-Vehicle Interactions. Transportation Research Record: Journal of the Transportation Research Board, 2457(1), 95–104. https://doi.org/10.3141/2457-10

Louhghalam, A., Akbarian, M., & Ulm, F.-J. (2016). Carbon management of

infrastructure performance: Integrated big data analytics and pavement-vehicle-interactions. Journal of Cleaner Production, 142. https://doi.org/http://dx.doi.org/10.1016/j.jclepro.2016.06.198

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General) Journal of the Royal Statistical Society. Series A (General J. R. Statist. Soc. A, 13517213(3), 370–384. https://doi.org/10.2307/2344614

Chipman, H., George, I., & McCulloch, R. (2010). BART: Bayesian additive regression trees. The Annals of Applied Statistics., 4, 266–298

Friedman, J. (1991). Multivariate adaptive regression spline. The Annals of Statistics, 19, 1–141.

Contribution of each team member:

Angad Arora: Literature review, understanding dataset, preliminary data visualization, selection of variables, understanding inferences.

Mayank Gupta: Literature review, understanding dataset, gathering, and cleaning the required data, selection of variables, understanding inferences.

Nidhi Desai: Literature review, understanding dataset, gathering, and cleaning the required data, models fitting, understanding inferences.