A Proposed Solution to the 2015 RAS Problem Solving Competition
An Ensemble Classifier to Predict Track Geometry Degradation

Team: UTSO

Iván Cárdenas[a], Carlos Sarmiento[a], Gilberto Morales[a]

[a] *Centro para la Optimización y la Probabilidad Aplicada, Departamento de Ingeniería Industrial, Universidad de los Andes, Bogotá, Colombia*
{*id.cardenas470, ca.sarmiento1227, ga.morales413*}*@uniandes.edu.co*

Abstract

Railway operations are inherently complex and source of several problems. In the context of the 2015 RAS problem solving competition, this paper presents a solution approach which entails the construction of an ensemble classifier to forecast the degradation of track geometry. Our classifier is constructed by solving the problem from three different perspectives: deterioration, regression and clustering. We considered a different model from each perspective and our results show that using an ensemble method improves the predictive performance.

Keywords: Railroad Maintenance, Defects, Gamma Process, Logistic Regression, Support Vector Machines, Classification, Ensemble Algorithms

1. Introduction

Rail is a crucial mode of transportation in the United States. Railroads account for approximately 40 percent of intercity freight volume - more than any mode of transportation (AAR, 2015). In addition, Amtrak, the National Railroad Passenger Corporation transports an average of 86000 passengers every day. Analyzing track geometry defects is critical for keeping freight and passenger trains moving safely. According to the US Federal Railroad Administration Office of Safety Analysis (FRA, 2014), track defects are one of the leading causes of train accidents in the United States. For instance, among the 1747 train accidents that happened in 2012, 577 (33.03%) were caused by track defects, resulting in a total reportable damage of $ 102.9 million (Peng et al., 2013). Those defects are classified into two severity levels - red tags and yellow tags. Red tag defects violate FRA track safety standards and must be treated as soon as possible. Yellow tag defects satisfy FRA standards; however, they will eventually become red tag defects if they are not fixed. Ability to predict yellow tags which are potentially turning into red tags, before they are actually measured as red tag defects, allows railroads to more efficiently maintain the rail and remain in FRA compliance. Therefore, our main objective throughout this report is to present our ensemble solution that can forecast track degradation along time by solving the problem from three different perspectives; deterioration, regression and clustering.

*Universidad de los Andes, Carrera 1^a Este No. 19A - 40, Bogotá, Colombia.

2. Background and model preliminaries

2.1. Background

Every year, North American railroads spend millions of dollars on periodic rail inspection. A fleet of track geometry vehicles travel on the railroad network and examine rail tracks for external and internal rail defects using visual inspection and technologies such as induction and ultrasonic devices (Cannon et al., 2003). Additionally, they are equipped with Global Positioning System (GPS) to accurately identify the location where measurements are taken. These vehicles have the ability to identify around 40 different types of defects; however, only three specific types were analyzed: XLEVEL (XLE) is the difference in elevation between the top surfaces of the rails at a single point in a tangent track segment, SURFACE (SUR) exceptions are determined by depressions or humps in the rail surface and DIP (DIP) is the largest change in elevation of the centreline of the track within a certain moving window distance. DIP may represent either a depression or a hump in the track and approximates the profile of the centreline of the track. The following sections present in detail the proposed solution emphasizing data analysis, implemented models and obtained results.

2.2. Data Summary

Our model is based on a field dataset from 4 tracks including 4 different classes of railroads (II, III, IV, V). The database contains registers of the three types of defects (XLE, SUR, DIP) measured by the track geometry vehicles from 2007 to 2013. In total, there are approximately 6500 red tag defect records and 17500 yellow tag defect records. The limits that separate a yellow tag defect from a red tag defect for a specific railroad class were obtained from the Track and Rail and Infrastructure Integrity Compliance Manual (FRA, 2014). For modeling purposes, we defined ϕ_i as the amplitude in inches of the i^{th} record, t_i as the date in which the i^{th} record was registered and Y_i represents the color of the tag in the i^{th} record ($Y_i = 1$ if red tag or $Y_i = 0$ if yellow tag).

Additionally, the database contains records of the sum of total gross tons traveling across a particular segment of each track during a specific month. For modeling purposes, we aggregated the measurements by tracks and months. Then, we computed Θ_{kj} as the mean tonnage crossing across the track k in the month j.

2.3. Data Tidying

Data tidying consists of giving structure to the dataset to facilitate its analysis. Tidy data sets are easy to manipulate, model and visualize (Wickham, 2014). A considerable effort was spent cleaning data to get it ready for analysis, using a sequential process composed of three stages (Aggregation, Joining, Cleaning).

- **Aggregation**: The dataset was divided into 12 subsets, each of which containing all the records for a specific combination of track and defect type. To generate consistent spatial units and accommodate different modeling purposes, we divided each track in smaller segments called lots, following the strategy presented by He et al. (He et al, 2014) used for track deterioration analysis. Each lot is 0.02 mile (about 100 ft) in length. Then, we aggregate the registers within each lot. Given that there is more than one register per date per lot, we took the register of the maximum amplitude to represent the track segment condition for the inspection run under consideration.

- **Joining**: Given that we are interested in the evolution of the defects over time, a single record of the defect i is not enough to develop our analysis. Hence, we search the next record of the defect in time for each of the defects. Using the two consecutive registers for a specific defect and lot, we computed $\Delta_{\phi_i} = |\phi_{i+1}| - |\phi_i|$ as the difference between the absolute value of the amplitude of the $i+1$ and the i defect. Moreover, we calculate $\Delta t_i = t_{i+1} - t_i$ as the number of days in which the Δ_{ϕ_i} change in amplitude occurred. Finally, we look for the changes in the color tags of the defects defining the variable $\eta = 0$ if $Y_i = 0$ and $Y_{i+1} = 0$, $\eta = 1$ if $Y_i = 0$ and $Y_{i+1} = 1$, $\eta = 2$ if $Y_i = 1$ and $Y_{i+1} = 0$ and $\eta = 3$ if $Y_i = 1$ and $Y_{i+1} = 1$.

- **Cleaning**: From the constructed data set of joint registers, we remove the cases in which the following situations occur:

  - $\Delta_{\phi_i} < 0$: There are cases where the cumulative deterioration decreases over time. Removing these cases from the data set is equivalent to assuming that the defects always get worse along time in absence of a preventive or corrective maintenance.
  - $\Delta t_i > 365$: There are long periods of time without any record for a specific defect. Assuming that the track geometry cars inspect the tracks at least once a year, it is almost certain that during a $\Delta t_i > 365$ a preventive or corrective maintenance was done given the lack of records.
  - $\eta = 2$ **and** $\eta = 3$: Given that the proposed model must predict when yellow tag defects will reach red tag levels, the joint registers that initiate with red tag defects are not useful in our analysis.

  In conclusion, for each possible combination of tracks and defect types we developed a dataset containing the following variables: Δ_{ϕ_i} represents the change in amplitude of the defect i in Δt_i days, $\theta_{k\Delta t_i}$ as the mean tonnage crossing track k in that specific segment Δt_i of the year using the values of Θ_{kj} which represent the mean tonnage crossing track k in month j. Moreover, we computed α_i as the missing amplitude of the yellow tag defect i to reach the red tag level. Finally, the value of η considering $\eta = 0$ if $Y_i = 0$ and $Y_{i+1} = 0$ or $\eta = 1$ if $Y_i = 0$ and $Y_{i+1} = 1$.

3. Proposed model

In order to improve current track rectification decisions, this study aims to help existing railroads address the following question: Will a specific yellow tag defect reach red level before an interval of t days? We have analyzed this problem from three different perspectives, as follow.

- **Deterioration**: When a railroad network is put into use, physical changes to the system occur over time. These changes may be the result of internal processes; for instance, natural changes in material properties, or external processes, such as environmental conditions. Regardless of the cause, these changes may result in the loss of system capacity to perform its intended function (Sánchez-Silva, 2015). Under this perspective, we propose a ***gamma process*** to model the increasing amplitude of the defects over time.

- **Regression**: Regression analysis estimates the relationship between a dependent variable and a set of explanatory variables, and is widely used for prediction and forecasting (Phillips et al., 2015). For a specific defect the set of variables (α_i,Δt_i,$\theta_{k\Delta t_i}$) affect the probability of the defect being yellow or red after t days. In this context, we propose a ***binary logistic regression*** to interpret how our response variable Y_i reacts to the explanatory variables.

- **Clustering**: Machine learning is a field within computer science focusing on the study and construction of algorithms that can learn from and make predictions on data (Kohavi et al., 1998). Support Vector Machines (SVM) are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. From this perspective, we propose a ***SVM model*** to identify how to cluster defects in two categories: yellow tags and red tags.

In addition, the deterioration of the railroad network can be seen from the perspective of materials science (mechanical perspective). In particular, rail rolling contact fatigue, which is caused by the repeated rolling contact of wheels on the rail, evolves into rail failures such as squats, flaking and head check cracks (Matsui et al., 2013). Given that we do not have enough data to propose and evaluate a fatigue crack propagation model expressed in terms of the acting stresses and the rail material, this perspective of the problem was not considered.

In order to take advantage of the benefits of each of the proposed approaches, we develop a single classifier model for each perspective to predict our response variable Y. In order to reduce the variance and the bias of the predictions, we will ensemble the models into a unified model using bootstrap aggregation and stacking techniques. In conclusion, our proposed solution follows the structure presented in Figure 1.

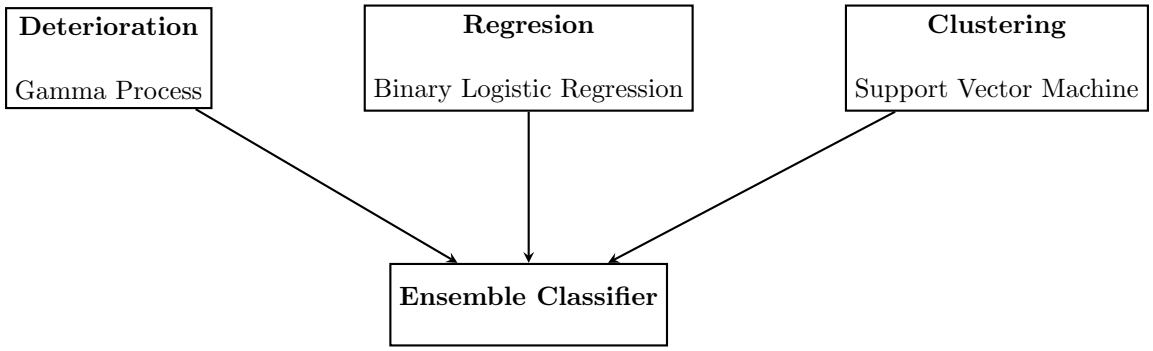

Figure 1: Solution Structure

3.1. Gamma Process

Since the introduction of the gamma process in the area of reliability in 1975, it has been increasingly used to model deterioration in terms of a time-dependent stochastic process for optimizing maintenance (Noortwijk, 2009). The gamma process is suitable to model gradual damage monotonically accumulating over time in a sequence of tiny increments, such as corrosion (Kallen et al., 2005), crack growth (Lawless et al., 2004) and concrete creep (Cinlar et al., 1977). In mathematical terms, the gamma process is defined as follows (Noortwijk, 2009). Recall that a random quantity Z has

a gamma distribution with shape parameter $v > 0$ and scale parameter $u > 0$ if its probability density function is given by $Ga(z|v,u) = \frac{u^v}{\Gamma(v)} z^{v-1} e^{-uz}$, where $z \geq 0$ and $\Gamma(a) = \int_{x=0}^{\infty} x^{a-1} e^{-x} dx$ is the gamma function for $a > 0$. Furthermore, let $v(t)$ be a non-decreasing, right-continuous, real-valued function for $t \geq 0$, with $v(0) = 0$. The gamma process with shape function $v(t) > 0$ and scale parameter $u > 0$ is a continuous-time stochastic process $\{X(t), t \geq 0\}$ with the following properties: (1) $X(0) = 0$ with probability one; (2) $X(\tau) - X(t) \sim Ga(v(\tau) - v(t), u)$ for all $\tau > t \geq 0$; (3) $X(t)$ has independent increments.

Let $X(t)$ denote the cumulative deterioration at time t, $t \geq 0$, and let the probability density function of $X(t)$, in accordance with the definition of the gamma process, be given by $f_{X(t)}(x) = Ga(x|v(t),u)$ with expectation and variance $E(X(t)) = \frac{v(t)}{u}$, $Var(X(t)) = \frac{v(t)}{u^2}$, respectively. In order to apply the gamma process model in the rail network problem, for each track and defect type, we aggregated all the records of Δ_{ϕ_i} representing the change in amplitude of the defect in Δt_i days into a data set. This procedure is equivalent to assuming that all the defects in a particular combination of track and defect type follow the same pattern of deterioration. In conclusion, we estimated 12 gamma processes (one for each combination of track and defect type). Each data set consists of inspection times t_i, $i = 1, ..., n$, where $0 = t_0 < t_1 < t_2 < ... < t_n$, and corresponding observations of the cumulative amounts of deterioration x_i, $i = 1, ..., n$, where $0 = x_0 \leq x_1 \leq x_2 \leq ... \leq x_n$. Considering a gamma process with shape function $v(t) = ct^b$ and scale parameter u, the parameters c and u are estimated by the method of maximum likelihood initially presented by Cinlar et al. (Cinlar et al., 1977). We assume that the value of the power b is initially known and equal to 1, but c and u are unknown. However, c and u can be obtained by maximizing the logarithm of the likelihood function of the increments. The likelihood function of the observed deterioration increments $\delta_i = x_i - x_{i-1}$ $i = 1, ..., n$ is presented in equation 1.

$$L(\delta_1, ..., \delta_n | c, u) = \prod_{i=1}^{n} f_{X_{(t_i)} - X_{(t_{i-1})}}(\delta_i) = \prod_{i=1}^{n} \frac{u^{c[t_i^b - t^b{}_{i-1}]}}{\Gamma(c[t_i^b - t^b{}_{i-1}])} \delta_i^{c[t_i^b - t^b{}_{i-1}]-1} e^{-u\delta_i} \tag{1}$$

By computing the first partial derivatives of the loglikelihood function of the increments with respect to c and u, the maximum-likelihood estimates $\hat{c}$ and $\hat{u}$ can be solved from equation 2.

$$\hat{u} = \frac{\hat{c} t_n^b}{x_n} \quad and \quad \sum_{i=1}^{n} [t_i^b - t_{i-1}^b] \{\psi(\hat{c}[t_i^b - t_{i-1}^b]) - log\delta_i\} = t_n^b log(\frac{\hat{c} t_n^b}{x_n}) \tag{2}$$

The maximum-likelihood method to estimate the parameters c and u can be extended to estimate the parameter b as well. The parameter b then must be determined by numerically maximizing the likelihood function L (Nicolai et al.,2007). A particular gamma process was estimated for each track and type of defect, the parameters are shown in Table 1.

In this way, we calculated the probability that a yellow tag defect remains tagged as yellow after t days using the cumulative distribution function of the gamma process estimated. Basically, recall that we defined $X(t)$ as the deterioration in the amplitude of the defect after t days. Assuming that $X(t) \sim Ga(v(t) = ct^b, u)$, we computed $P(Y = 0) = F_{X(t)} = (P(X(t) \leq \alpha)$. Thereby, we are computing the probability that the deterioration of the defect in these t days does not exceed the missing amplitude to reach red tag level. On the other hand, the probability that a yellow tag defect reaches the red level after t days is estimated using the complement, so $P(Y = 1) = 1 - P(Y = 0)$. Finally, we estimated an

Track 1 - XLE		**Track 2 - XLE**		**Track 3 - XLE**		**Track 4 - XLE**	
$\hat{b}$	1.2008	$\hat{b}$	1.1076	$\hat{b}$	0.9367	$\hat{b}$	0.7861
$\hat{u}$	1.377	$\hat{u}$	1.5022	$\hat{u}$	1.5488	$\hat{u}$	0.9397
$\hat{c}$	0.0010	$\hat{c}$	0.0022	$\hat{c}$	0.0176	$\hat{c}$	0.0292
Track 1 - SUR		**Track 2 - SUR**		**Track 3 - SUR**		**Track 4 - SUR**	
$\hat{b}$	0.8743	$\hat{b}$	0.9921	$\hat{b}$	1.0047	$\hat{b}$	0.9570
$\hat{u}$	0.5146	$\hat{u}$	0.6948	$\hat{u}$	0.6843	$\hat{u}$	0.6312
$\hat{c}$	0.0246	$\hat{c}$	0.0118	$\hat{c}$	0.0094	$\hat{c}$	0.0153
Track 1 - DIP		**Track 2 - DIP**		**Track 3 - DIP**		**Track 4 - DIP**	
$\hat{b}$	0.9658	$\hat{b}$	1.1449	$\hat{b}$	1.0309	$\hat{b}$	1.1705
$\hat{u}$	0.5909	$\hat{u}$	0.7216	$\hat{u}$	0.7120	$\hat{u}$	0.8320
$\hat{c}$	0.0128	$\hat{c}$	0.0024	$\hat{c}$	0.0069	$\hat{c}$	0.0017

Table 1: Gamma Process parameters

optimal threshold (π) in order to establish the final prediction using the ROC curve (Ballings et al., 2015). In conclusion, if $P(Y = 1) > \pi$ the final prediction would be red tag, otherwise the prediction would be yellow tag.

3.2. Binary Logistic Regression

In diverse statistical problems, usually the variable of interest is qualitative, which means that it can only fit into one of a finite group of categories. Predicting a qualitative response for a new observation is a process known as classification (James et al.,2013). Binary logistic regression is a widely used technique, for the case of a dependent variable with two possible outcomes or categories. Because it is capable of computing the probability that the response variable is classified into one of the two possible outcomes, given a set of independent variables or predictors $(x_1, x_2, ..., x_p)$. For this reason it is extremely helpful in very different studies. For example, logistic regression can be used for computing a spatial prediction of landslide hazard using factors such as slope, altitude, aspect, etc.(Hong et al., 2015). On the other hand, in the modelling of wildfires, logistic regression has been used to model the human-caused wildfire risk estimation (del Hoyo et al., 2011). In reliability engineering logistic regression has also been widely applied. For instance, in a crack recognition algorithm for road pavements that used a binary logistic regression model that included morphological characteristics of the road as predictors (Yoo et al.,2015).

In the case of this paper, binary logistic regression allows us to compute the probability that a yellow tag will become a red tag before an interval of t days, based on the logistic function that gives outputs between 0 and 1 for any input value. So having p independent variables the probability of having a red tag is presented in equation 3

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}}, \tag{3}$$

where $(x_1, x_2, ..., x_p)$ is the set of independent variables (predictors) and $(\beta_1, \beta_2, ..., \beta_p)$ are the correspondent regression coefficients.

As it was explained in Section 2, a data tidying process was performed, so that a complete set of predictors could be computed. We developed 3 different models, one for each defect (XLE, DIP and SUR). In each model the factors included are: (1) α, which is the missing amplitude (in inches) of the yellow tagged defect for reaching the red tag level, (2) $\theta_{k\Delta t}$,

which is the mean tonnage travelling across the track k during a Δt interval of days, (3) Δt which is the interval (in days) between measures and (4) k which is the correspondent track of the defect evaluated.

Regression coefficients were found using the maximum likelihood method, in which we seek estimates of β_0 to β_p such that the predicted probability ($\hat{p}(x_1, ..., x_p)$) of red for each yellow tagged defect corresponds as closely as possible to the defect's observed outcome after Δ_t days. In other words, it is choosing the estimates of $\hat{\beta}_0$ to $\hat{\beta}_p$ that maximize the likelihood function presented in equation 4.

$$\ell(\beta_0, ..., \beta_p) = \prod_{i:y_i=1} \hat{p}(x_i) \prod_{i':y_{i'}=0} (1 - \hat{p}(x_{i'})) \tag{4}$$

The estimated coefficients are shown in Table 2. It can be observed that given the track is a qualitative independent variable, 3 dummy variables must be implemented in the model.

Defect Type	β_0 Intercept	β_1 - α Amplitude	β_2 - $\theta_{k\Delta_t}$ Tonnage	β_3 - Δ_t Interval	β_4 Track 2	β_5 Track 3	β_6 Track 4
XLE	1.149	-29.204	0.096	-0.001	0.216	-0.188	0.289
SUR	0.598	-20.21	0.083	-3.64E-04	0.486	0.768	-
DIP	1.687	-20.277	0.018	0.007	-0.556	-0.712	-0.752

Table 2: Binary Logistic Regression Parameters

Finally, the probability that a given defect turns red after an interval of Δ_t days is computed using the logistic function. The defect type (XLE, SUR, DIP) will define which one of the 3 logistic regression models must be used. Then, the current amplitude of the defect, the tonnage that will travel across the section and the track where the defect is located are used as input values for the model. Lastly, the probability of red is computed using the logistic function and the estimated coefficients shown in Table 2. As in the case of the gamma process, we estimated an optimal threshold (π) in order to establish the final prediction using the ROC curve.

3.3. Support Vector Machines

Support Vector Machines (SVM) allows cluster classification through a separation kernel function. The binary classification problem was initially created by V. Vapnik and the AT&T research group (Cortes et. al, 1995). We tested the most common kernel functions: linear, polynomial, radial and sigmoidal functions(Guyon et. al.,2002). The purpose of the kernel function or hyperplane is to separate in the most effective way the clusters of each class. In order to obtain this separation, the algorithm maximizes the distance from both clusters simultaneously, while ensuring that each cluster remains in a different side from the hyperplane. As most datasets contain outliers, we must include a slack ξ_i to each data point that will activate in the case the data point does not belong to the cluster. We must also include a penalty C in the case the slack is activated in order to create the most efficient hyperplane and not a modification due to the outliers.

Figure 2 presents the linear kernel ($wX - b = 0$) and the related optimization problem for a dataset with classes -1 and 1. In this algorithm, we assign the value -1 to the class $Y = 0$ in order to accurately perfom the convex optimization problem. The information in this dataset is classified in a class vector $Y \in \{-1, 1\}^m$ and a characteristic matrix $X \in \mathbb{R}^{n \times m}$

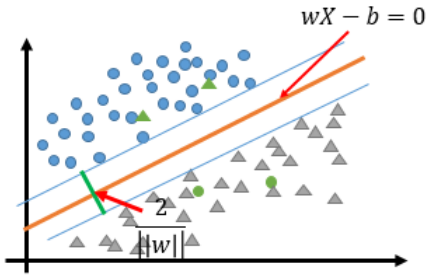

Figure 2: SVM linear kernel and Optimization Problem

Defect	Amplitude	Tonnage	Days	Track 1	Track 2	Track 3	b
DIP	-15.435	0.0003	0.004	0.492	0.023	-0.018	0
SUR	-15.311	0.051	0.003	-0.318	0.08	0.238	0
XLE	-24.398	0.062	-0.002	0.035	0.381	0.131	0

Table 3: Support Vector Machines Coefficients

with n characteristics to classify each data point (α, which is the missing amplitude (in inches) of the yellow tagged defect for reaching the red tag level, $\theta_{k\Delta t}$, which is the mean tonnage travelling across the track k during a Δt interval of days, Δt which is the interval (in days) between measures and k which is the correspondent track of the defect evaluated). This in order to obtain the weights (w) and b coefficients. The optimization problem associated with this kernel function is presented in equations 5-7.

$$\min_{w\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^{+m}} \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \qquad \text{s.t.} \tag{5}$$

$$Y_i(X_i w - b) \geq 1 - \xi_i \qquad i = 1, \cdots, m \tag{6}$$

$$\xi_i \geq 0 \qquad i = 1, \cdots, m \tag{7}$$

For each the defects we tested several penalization (C) values, in order to determine the ideal kernel function for each of the possible defects. For all the defects we obtained an ideal lineal kernel with the coefficients (w and b) presented in table 3.

3.4. Ensemble Classifiers

Ensembles methods solve problems that are (1) statistical, (2) computational, and (3) representational in nature (Dieterrich,2000) by averaging models . Many different classifiers can be learned from specific combinations of data, representations, objective functions and optimization methods. Combining the predictions of multiple classifiers can reduce the variance because the results are less dependent on peculiarities of a single model. Moreover, ensemble classifiers can reduce the bias because a combination of multiple classifiers may learn a more expressive concept class than a single classifier (Zhou, 2012). In order to take benefit of these advantages of ensemble classifiers, we analysed two techniques:

- **Bootstrap Aggregating**: Bagging, short for bootstrap aggregating, is considered one of the earliest ensemble schemes (Breiman, 1996). Bagging is intuitive but powerful, when a certain number of calssifiers are generated, these individuals are combined by the majority voting scheme. Given a testing instance, different outputs ($Y = 0$, $Y = 1$) will be given from the trained classifiers (Gamma process, Binary Logistic Regression, SVM), and the output voted by the majority will be considered as the final decision. Since the number of models is odd, there is no possibility for a tie.

- **Stacking**: Stacking has a two level structure: level-0 (base-level) classifiers and a level-1 (meta) classifier (Wolpert, 1992). During the process of classifying a new instance ($Y = 0$, $Y = 1$), the trained base-level classifiers will give their individual predictions, and the predictions will be considered as the input of the meta-classifier to generate the final decision. In this way, we propose two stacked models. The first uses binary logistic regression as the meta-classifier, while in the second the meta-classifier is SVM. In both models, the other two single classifiers compose the base-level.

4. Results

In order to evaluate the accuracy of the proposed models, using the available data, we constructed a testing set with the real response for each type of defect. Given that false positives and false negatives count the same, our evaluation function will be the accuracy (error rate) rather than the sensitivity (true positive rate) or the specificity (true negative rate) of the model. Accuracy is the rate of correct predictions made by the model over a data set. Cross-validation is a method for estimating the accuracy of a classifier by dividing the data into k mutually exclusive subsets (folds) of approximately equal size. The classifier is trained and tested k times. Each time it is trained on the data set minus a fold and tested on that fold. The accuracy estimate is the average accuracy for the k folds. In Table 4, we present the overall accuracy results obtained by a cross validation procedure with $k = 4$. GPR stands for Gamma Process, BLR stands for Binary Logistic Regression, SVM stands for Support Vector Machine, BAG stands for Bootstrap Aggregating, SBL stands for Stacking with Binary Logistic Regression and SSV stands for Stacking with Support Vector Machine.

Single Classifiers			**Ensemble Classifiers**		
GPR	**BLR**	**SVM**	**BAG**	**SBL**	**SSV**
XLE: 72.15%	XLE: 73.17%	XLE: 74.55%	XLE: 74.61%	**XLE: 74.71%**	XLE: 73.99%
SUR: 77.05%	SUR: 79.25%	SUR: 78.73%	**SUR: 81.28%**	SUR: 80.79%	SUR: 79.50%
DIP: 75.76%	DIP: 81.62%	DIP: 76.28%	DIP: 79.45%	DIP: 81.15%	**DIP: 82.56%**

Table 4: Overall Accuracy Results

We observed that the best results were obtained through the ensemble classifiers. In each defect, at least one of the ensemble classifiers overpasses the single algorithms. Nevertheless only in the SUR defect, all three ensemble classifiers overpass the three single algorithms, having the best result in the Bootstrap Aggregating algorithm. In the case of the XLE defect the best results were obtained through the Stacking with Binary Logistic Regression. Finally, for the DIP defect we obtained the best results through the Stacking with Support Vector Machines.

5. Concluding remarks

In order to solve the defect classification prediction problem, we tested multiple approaches considering deterioration (using gamma processes), regression (using binary logistic regression) and clustering (support vector machines). After considering each algorithm individually and testing them with real data given by the original problem's training data set, we considered ensemble classifier approaches in order to merge the algorithms. At least one of each ensemble classifier was the best in each of the given defects, resulting in choosing the Stacking with Binary Logistic Regression for the XLE defect, the Bootstrap Aggregating for the SUR defect and the Stacking with Support Vector Machine for the DIP defect.

6. References

1. Association of American Railroads (AAR), Economic and Public Benefits, Available online at https://www.aar.org/todays-railroads, September 2015.
2. Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R., 2015, Evaluating multiple classifiers for stock price direction prediction, Expert Systems with Applications, 42, 7046-7056.
3. Breiman, L., 1996, Bagging predictors, Machine Learning, 24, 123-140.
4. Cannon, D., Edel, K., Grassie, S., Sawley, K., 2003, Rail defects: An Overview, Fatigue & Fracture of Engineering Materials & Structures, 26, 865-886.
5. Cinlar, E., Bazant, Z., Osman, E., 1977, Stochastic process for extrapolating concrete creep, Journal of the Engineering Mechanics Division, 103, 1069-1088.
6. Cortes, C., & Vapnik, V, 1995. Support-vector networks. Machine learning, 20, 273-297.
7. del Hoyo, L., Isabel, M., Martínez, F., 2011, Logistic Regression Models for Human-caused Wildfire Risk Estimation: Analysing the Effect of the Spatial Accuracy in Fire Occurrence Data, European Journal of Forests Reserve, 130, 983-996.
8. Dietterich T., 2000, Ensemble methods in machine learning, Proceeding of the First International Workshop on Multiple Classifier Systems, 1-15.
9. Federal Railroad Administration, Track and Rail and Infrastructure Integrity Compliance Manual, Volume II Track Safety Standards, Chapter 1 Track Safety Standards, January 2014.
10. Federal Railroad Administration Office of Safety Analysis, Safety Data, Available online at http://safetydata.fra.dot.gov/officeofsafety, September 2015.
11. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. 2002. Gene selection for cancer classification using support vector machines. Machine learning, 46, 389-422.
12. He, Q., Li, H., Bhattacharjya, D., Parikh, D., Hampapur, A., 2015, Track geometry defect rectification based on track deterioration modelling and derailment risk assessment, Journal of the Operational Research Society, 3, 392-404.
13. Hong, H., Pradhan, B., Xu, C., Bui, D., 2015, Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines, Catena, 133, 266-281.
14. James, G., Witten, D., Hastie, T., Tibshirani, R., 2014, Introduction to Statistical Learning. New York: Springer.
15. Kallen, M., Noortwijk J.M., 2005, Optimal maintenance decisions under imperfect inspection, Reliability Engineering & System Safety, 90, 177-185.
16. Kohavi, R., Provost, F., 1998, Glossary of terms, Machine Learning, 30, 271-274.
17. Lawless, J., Crowder, M., 2004, Covariates and random effects in a gamma process model with application to degradation and failure, Lifetime Data Analysis, 10, 213-227.
18. Matsui, M., Kamiya, Y., 2013, Evaluation of material deterioration of rails subjected to rolling contact fatigue using x-ray diffraction, Wear, 304, 29-35.
19. Nicolai, R., Dekker R., Noortwijk J., 2007, A comparison of models for measurable deterioration: An application to coatings on steel structures, Reliability Engineering & System Safety, 92, 1635-1650.
20. Noortwijk, J.M., 2009, A survey of the application of gamma processes in maintenance, Reliability Engineering & System Safety, 94, 2-21.
21. Osuna, E., Freund, R., & Girosi, F, 1997, Training support vector machines: an application to face detection. In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on (pp. 130-136). IEEE.
22. Peng, F., Ouyang, Y., Somani, K., 2013, Optimal routing and scheduling of periodic inspections in large-scale railroad networks, Journal of Rail Transport Planning & Management, 3, 163-171.
23. Phillips, J., Cripps, E., Lau, J., Hodkiewicz, M., 2015, Classifying machinery condition using oil samples and binary logistic regression, Mechanical Systems and Signal Processing, 60, 316-325.
24. Sánchez-Silva, M., Klutke, G., 2015, Reliability and Life-Cycle Analysis of Deteriorating Systems, Springer Series in Reliability Engineering, 1, 348.
25. Wickham, H., 2014, Tidy Data, Journal of Statistical Software, 59, 10.
26. Wolpert, D., 1992, Stacked Generalization, Neural Networks, 5, 241-259.
27. Yoo, H., Kim, Y., 2015, Development of a Crack Recognition Algorithm from Non-routed Pavement Images using Artificial Neural Network and Binary Logistic Regression, Journal of Civil Engineering, 0, 1-12.
28. Zhou, Z., 2012, Ensemble Methods: Foundations and Algorithms, Chapman & Hall Series in machine learning & Pattern Recognition, 1, 236.