Leal De-Rivas, BC.; Vivancos, J.; JOAQUÍN ORDIERES MERÉ; Capuz-Rizo, SF. (2017). Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models. Chemometrics and Intelligent Laboratory Systems. 160:32-39. doi:10.1016/j.chemolab.2016.10.015

Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models*

Beatriz Leal De Rivas[a], José-Luis Vivancos[b,c,d†],Joaquín Ordieres-Meré[e], Salvador F. Capuz-Rizo[b]

[a] Facultad de Ingeniería, Universidad Metropolitana, Terrazas del Ávila, Caracas, Venezuela.

[b] Departamento de Proyectos de Ingeniería, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain

[c] Centro de Reconocimiento Molecular y Desarrollo Tecnológico, Unidad Mixta Universitat Politècnica de València - Universitat de València, Camino de Vera s/n., 46022 Valencia, Spain

[d] CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Spain

[e] Research Group PMQ, Department of Industrial Engineering, Business Administration and Statistics, ETSII, Universidad Politécnica de Madrid (UPM), Spain

Abstract

Total acid number (TAN) has been considered an important indicator of the oil quality of used oils. TAN is determined by potentiometric titration, which is time-consuming and requires solvent. A more convenient approach to determine TAN is based on infrared (IR) spectral data and multivariate regression models. Predictive models for the determination of TAN using the IR data measured from ashless dispersant oils developed for aviation piston engines (SAE 50) have been developed. Different techniques, including Projection Pursuit Regression (PPR), Partial Least Square, Support Vector Machines, Linear Models and Random Forest (RF), have been used. The used methodology involved a five folder cross validation to derive the best model. Then a full error measure over the whole dataset was taken. A backward variable selection was used and 25 highly relevant variables were extracted. RF provided an acceptable modelling technology with grouped dataset predictions that allowed transformations to be performed that fitted the measured values. A hybrid method considering group of bands as features was used for modelling. An innovative mechanism for wider features selection based on genetic algorithm has been implemented. This method showed better performance than the results obtained using the other methodologies. RMSE and MAE values obtained in the validation were 0.759 and 0.359 for PPR model respectively.

Keywords

Aircraft turbine; engine oil; total acid number (TAN); FTIR; regression models

1. Introduction

Determining the condition of engine oil is critical for aviation safety and operation thereof. Therefore, periodic analyses of engine oils are mandatory. The conditions of aging fluids that

* In memoriam of Beatriz Leal De Rivas.

† Corresponding author at: Departamento de Proyectos de Ingeniería, Universitat Politècnica de València, Camino de Vera s/n 46022 Valencia, Spain. Tel.: +34 963877007; fax: +34 963879869. E-mail: jvivanco@dpi.upv.es

require regular monitoring are total acid number (TAN), viscosity index (VI), wear rate and depletion of antioxidants [1] . These parameters (TAN and VI) can be determined by standard physicochemical methods. However, the main motor oil physicochemical characteristics are covered by the American Standards for Testing and Materials (ASTM) guides. Viscosity index and TAN of oils are usually measured according to standard ASTM D445 [2] and ASTM D664 [3] , respectively. TAN determination is based on potentiometric titration with a base to a fixed endpoint. It is time-consuming and requires environmentally problematic solvents and reagents.

In recent years, efforts have been made to replace analytical methods based on the FTIR technique. In order to obtain analytical information in a rapid, non-destructive way, mid-infrared spectroscopy has been largely applied to study motor oils for different purposes: quantifying contaminants [4] [5] , or oxidation process [6] [7] , adulteration [8] [9] [10] , determining the antioxidant concentration [11] , determining physicochemical values (TAN [8][12] [13] , VI [14] [15] and TBN [12] [14]), and classifying them according to their origins [16] .

It is well-known that hydrocarbon oxidation is an autocatalytic process governed by the initiation of a free radical chain reaction, chain propagation, chain branching, and the termination of the radical chain reaction [17] . These multiple free radical pathways produce a complex mixture of possible oxygenated products, such as hydroperoxides, alkyl peroxides, dialkyl peroxides, alcohols, carboxylic acids, esters, ketones, diketones, aldehydes, hydroxyketones, ketoaldehydes and unsaturated oxygenated compounds [18] . All these molecules introduce functional groups that provide characteristic FTIR spectral bands. Therefore, oil oxidation should generate visible changes in the following vibration bands: (3100–3600 cm^{-1}; 2500– 3200 cm^{-1}; 1650–1730 cm^{-1}; 1680–1710 cm^{-1}; 1700–1740 cm^{-1}; 1050–1450 cm^{-1} and 1000–1250 cm^{-1}) [6] .

Recently, multivariate calibration has been applied to determine the antioxidant concentration [11] , and physicochemical values (TAN [8][12] [13] , VI [14] [15] and TBN [12] [14]), based on the FTIR technique. Multivariate calibration is an effective calibration method in which the chemical information (absorption, emission, transmission, etc.) of a set of standard mixtures recorded at different variables (wavenumbers) is related to the concentration of the chemical compounds present in the mixtures [19] . The popular form of calibration used in chemical analyses is univariate calibration, in which the chemical information of a set of solutions recorded at one variable (i.e., wavenumber) is related to the solute concentration in the solution. The most applied multivariate methods are classical least squares (CLS) [14] , principal-component regression (PCR) [14] and partial least squares (PLS) [7] [8] [11] [14] [15] .

In the present study, the TAN (total acid number) of turbine engine oils from military aircrafts was estimated by FTIR spectroscopy with multiple regression by using two different strategies. Instead of using the classic PCA for dimension reduction, we used both the PCA and the Independent component analysis (ICA), which slightly outperforms the PCA. By employing this backward variable selection, 25 highly relevant variables were extracted. After this variable selection, predictive models were built by different techniques, including Projection Pursuit Regression (PPR), Partial Least Square (PLS), Support Vector Machines (SVM), Linear Models, Generalized Linear Models (GLM) and Random Forest (RF). In addition, an innovative mechanism for wider features selection based on genetic algorithm has been implemented, which outperforms the previously used techniques.

2. Material and methods

2.1. Oil Samples.

Oil samples have been used in airplanes with fixed or reciprocating piston engines installed on a private aircraft (not including freight or passenger) with a maximum certified take-off weight of up to 5700 kg. These aircrafts are included in the *Schedule of Condition Based Maintenance* (SCBM) of the Venezuelan National Institute of Civil Aviation (Instituto Nacional de Aeronáutica Civil, INAC). INAC provides an alternative to extend the service up to 8 more years for these engines for a 12-year calendar time to perform reconditioning or overhaul, but which have not yet reached the limit in flight hours (1200–1700 hours) specified by the manufacturers.

Oils were used in engines of the brands Teledyne Continental Motors and Textron Lycoming. The owners of these aircrafts operate their units at fewer than 100 hours per year, and make engines work under tropical climate conditions (heat and moisture). This tends to accelerate oxidation processes and oil degradation, thus promoting corrosion and rust formation, mainly in those parts exposed to engine elements, such as cylinders. The SCBM program includes some preventive and predictive parameters, such as compression checks, boroscopic inspection, and the flow testing of cylinders, oil filters inspection, oil consumption verification, complete engine inspection, plus the prior inclusion of an oil analysis that the inspector should have before aircraft assessment begins.

Eighty ashless dispersant oils (AeroShell W 100, SAE 50) were collected from aviation piston engines used during the 2009–2012 period of engines of the brands Teledyne Continental Motors with operation times within the 600–2000 hour range and oils with a 50-hour operation. Table 1 shows the complexity of the samples in a variety of engines, times of operations and acidity.

Owners	Types of engines	No. samples	Average oil operation (hours)	Average engine operation (hours)	Period Sampling	TAN range (mg KOH / g oil)
1	6	17	30.7	1015.3	2009–12	0.28–3.53
2	2	9	7.3	660.8	2010–12	0.21–5.05
3	2	3	27.1	774.0	2010–11	0.18–1.76
4	5	8	23.8	601.1	2010–12	0.81–2.13
5	1	3	64.0	2092.0	2011–12	0.67–2.34
6	1	1	50.0	616.7	2010	1.30
7	1	1	60.6	1264.0	2010	1.24
8	1	1	50.0	1048.1	2011	2.95–2.95
9	1	2	33.5	1279.6	2010–11	0.94–1.57
10	2	3	52.8	1270.0	2012	1.08–1.95
11	1	1	47.4	742.9	2010	1.20
12	8	13	39.9	928.7	2010–12	0.64–3.14
13	1	5	33.1	1102.8	2010	0.49–1.22
14	3	7	39.6	1097.4	2010–12	0.98–2.07
15	3	6	37.5	1086.6	2010–11	0.48–1.96

Table 1. Oil samples

2.2. Infrared spectra.

Infrared spectra were obtained in a Fourier Transform spectrometer, Perkin Elmer model Spectrum 100, within a spectral range of 450–4000 cm^{-1}, with a resolution of 1 cm^{-1} and 16 scans

per sample, corresponding to 3551 data points per spectrum (original variables). The cells used were zinc selenide (ZnSe) transmission cells and path length was fixed at 0.1 mm.

The FTIR equipment was kept in a cabin at low humidity (under 45%) and was usually purged with nitrogen gas every 6 months. The background analysis and cleanliness of cells were performed between each determination in an estimated time of 5 minutes.

Oil samples were subjected to ultrasonic agitation for a time longer than 5 minutes, but shorter than 10 minutes, and 6 mL were extracted with a Pasteur pipette and analysed. No dilution was employed. Cells were cleaned with N-heptane before each determination. Spectra were collected in duplicate per sample and the % of transmittance and absorbance of the spectra were recorded with the FTIR software.

2.3. TAN Determination

TAN values were measured according to the ASTM D664 method [3] . For potentiometric titration, previously standardized potassium hydroxide (0.1 mol/L) was used to neutralize all the acidic constituents. Titrations were carried out in a mixture of toluene, isopropanol and water (volumetric ratio of 500:495:5). The amount of acidic constituents is given in mg of potassium hydroxide (KOH) per g of oil. A sample pretreatment was done by heating to 60°C and filtering through 100 µm mesh filters. Each analysis was done 3 times to determine an average. The standard deviation was established in 0.08 mg/KOH.

3. Results and discussion

3.1. Spectroscopic results

The FTIR spectra of the motor oil samples were recorded within the 450-4000 cm^{-1} range and are illustrated in Figure 1. To obtain clearer FTIR graphics, the full spectra were divided into three regions: 3000–2850, 1800–1500 and 1000–450 cm^{-1}. The spectra of the motor oils clearly overlapped within the entire spectral region and no certain wavenumber was found. Figure 1 shows three typical infrared spectra of engine oils which corresponded to oils with different acidities, low, medium and high, and characteristic changes due to acidity and oxidation.

Gracia et al. [6] reported a large absorption band between the absorption bands for (C=O) corresponding to esters (1740 cm^{-1}) and cyclic esters (lactones, 1780 cm^{-1}). These changes indicate typical oxidation products and aging products caused by contaminants in oil. Other authors have considered simplifying the spectrum and have eliminated several spectral ranges using criteria such as regions with total absorbance, regions with bands from strong C-H and C-C vibrations from hydrocarbons contained in base oil (typically 3100–2750 cm^{-1}, 1500–1300 cm^{-1}, and 800–700 cm^{-1}) and regions with no significant absorbances (typically 2750–2000 cm^{-1},1600–1500 cm^{-1}, and 700–600 cm^{-1}) [13] .

3.2. Data analysis methods

This section presents the data analyses carried out. The available dataset involved 76 oil samples. For all these samples, the FTIR spectra from 4000 to 450 cm^{-1} were available, as well as the TAN in accordance with ASTM D-664 [3] .

The research program focused on overcoming the main hurdle that the quantitative FTIR analysis of lubricants faced, the need for a reference oil. This stumbling block was overcome by applying

differential spectroscopy in conjunction with a specific stoichiometric reaction. By recording the oil spectrum twice, once before and once after reagent addition, and examining the differential spectrum obtained by subtracting one spectrum from another, quantifiable information on a particular constituent can be obtained.

The main goal in this section was to build up a model to help determine TAN within a confidence range because it can reduce the cost and time required for this analysis. This is relevant because this determination is a must in most of the aircraft regulatory body as it determines the acid contamination of used oils. Having a quick and averaged procedure for TAN estimations is greatly appreciated by maintenance staff.

To determine such models, and in accordance with the literature, there are two different main methodologies. The first looks at reducing the number of independent variables (Figure 1). In order to build an effective and robust classification model the spectra data (that have up to 3500 independent variables) should be reduced, given the limited number of spectra available. Two common ways to reduce data are spectra averaging (up to 8 times), which is a bias-prone technique in combination with the principal component analysis (PCA), which is the most universally applied [20] The second methodology looks at identifying the most relevant individual wavelengths to explain the most desired variable. Different strategies are available, like expert judgment, forward or backward variable selection, genetic algorithm-based variable selection, etc. [13] .

In this work, the above mentioned strategies were used to analyse the quality of the solution and a hybrid solution was introduced.

For the first methodology, instead of using the classic PCA for dimension reduction, we used both the PCA and the Independent component analysis (ICA), which slightly outperforms the PCA. After dimension reduction, a model based on non-lineal regression by Support Vector Machines (SVM) with a Gaussian kernel was implemented. However, the techniques based on variable selection outperformed this solution.

Many feature selection routines use a "wrapper" approach to find appropriate variables, such as an algorithm that searches the feature space and repeatedly fits the model with different predictor sets. The best predictor set is determined by some measure of performance (i.e., RMSE, R^2, classification accuracy, etc.). Examples of search functions are genetic algorithms, simulated annealing and forward/backward/stepwise selection methods. In theory, all these search routines can converge to an optimal set of predictors. Then, predictive models were built by different techniques, including Projection Pursuit Regression, Partial Least Square, Support Vector Machines, Linear Models, Generalized Linear Models and Random Forest. The used methodology involved a five folder cross validation to derive the best model. Then a full error measure over the whole dataset was taken.

Since the authors were not satisfied with the application of the existing techniques, they wanted to combine both strategies in order to produce a better one. The main motivation was being convinced that not only a few variables (absorbance at different wavelengths) would be exclusively related to the TAN property but specific regions in the spectra as well. Actually, their opinion was that the relevance will arise when comparing the ratio between specific regions being sensible to the effect being studied, and other regions where such sensitivity does not happen.

One feature to be considered hereinafter will be described in equation 1.

$$feature_i = \int_{\lambda_1^i}^{\lambda_2^i} \left(1 - \frac{f(\lambda)}{F_{cont}^i}\right) \cdot d\lambda \quad (1)$$

Where f(λ) denotes the absorbance at the wavelength λ and F^i_{cont} is the average absorbance between $\lambda^i_{cont;1}$ and $\lambda^i_{cont;2}$. It is assumed that $(\lambda^i_{cont;1}, \lambda^i_{cont;2}) \cap (\lambda^i_1, \lambda^i_2) = \varnothing$.

In the following subsections, specific details about the applied strategies will be provided.

3.2.1. Dimension Reduction

An independent component analysis (ICA) was used, its definition being found in different papers [21] [22]

According to the work by Gonzalez-Marcos [23] the ICA slightly outperforms the PCA for data compression when used for regressing specific functions, especially when the Signal to Noise ratio is relatively low. Therefore, the ICA with different dimensions, between 5 and 10, has been explored, and models have been regressed in order to estimate the TAN value from the projected space. As presented later in this work, the performance of this strategy was unfortunately not very high, given the limited amount of available samples and the internal variability of the sample set.

Instead of trying to reduce data dimensions by projection, we used backward stepwise variable selection. In this work, we used the backward variable selection and we extracted 25 highly relevant variables (see Figure 2). Then predictive models were built with different techniques, including Projection Pursuit Regression, Partial Least Square, Support Vector Machines, Linear Models and Random Forest. The used methodology involved a five folder cross validation to derive the best model. Then a full error measure over the whole dataset was taken. The selected variables are shown in Table 2 (in cm^{-1}).

Variable Number	**1**	**2**	**3**	**4**	**5**
0	X1711	X1710	X1709	X1712	X1713
5	X1715	X1456	X1714	X469	X1707
10	X1706	X1708	X2924	X450	X454
15	X1717	X1716	X1631	X452	X1705
20	X722	X1718	X2954	X451	X953

Table 2. Wavelength of the relevant variables regarding the data projection for ICA.

Interpretations of the meaning of such wavelengths can be found, like carboxylic acids with ν (O–H) overtone [7] , ν (O–H), δ (C–O–H), and δ (O–H); aldehydes with ν (C=O); unsaturated hydrocarbons with ν (C=C), and several nitro compounds with ν (NO_2). The aging of some detergents may give signals in the fingerprint region. Oxidation products from aromatic compounds increase the intensity of ν (C=C). All these compounds influence the TAN value [13]

3.2.2. Random forests model

Random forests [24] , which are a combination of tree predictors using both the bagging and randomization approaches (ensemble of models), have received considerable attention in recent years and they combine many trees to form a forest for analyses. An individual tree represents a model describing the characteristics of an input feature present in a subset of the whole dataset. The accuracy of RF has proven competitive with many other data-mining techniques ([24] [25]).

Biau and Devroye [26] discussed the links between the layered nearest neighbour estimate and RF estimates, and proved the universal consistency of the bagged nearest neighbour method for regression and classification. The results indicate that besides RF's comparable accuracy and mathematical simplicity, they are computationally fast and robust to noise. In our case, the forest was designed to use 500 trees and each tree randomly involved three input variables.

As the main goal was to predict TAN values, regression was considered the proper problem type to be dealt with. Then several regression techniques were tested to determine which was the most successful to be used hereinafter. RF is a useful tool for regression studies and has the potential to model linear and non-linear multivariate calibration as it offers good behaviour if compared to many other techniques [27] . Our analysis also supports the same conclusion, even when accuracy is significantly higher (see supporting).

RF accuracy is represented in Figure 3. It can be concluded that Random Forest provides acceptable modelling technology with grouped dataset predictions to allow the performance of transformations to fit the measured values. The standard method [3] indicates that for the oils used in Potentiometric Titration, reproducibility can reach an error of 44%, and the developed model can predict within these limits for almost all the samples, except for a few samples (less than 10%) in which they were higher.

In Figure 4, it becomes clear that the size considered in this particular application for the Random Forest technology is properly dimensioned as the error becomes stabilized. Indeed in order to speed it up, just 400 trees seems enough, but stability is clearly exhibited.

Individual variables were considered relevant in this analysis, according to the strategy adopted in this work, which was the backward stepwise variable selection. Permutation importance, and similar schemes for variable selection and for providing statements of the "significance" of a predictor variable (instead of a merely descriptive ranking of the variable importance scores), are common. Sometimes, however, they produce low statistical significance. This is because some authors have proposed the backward elimination strategy, which we have used here.

A relevant factor is the importance of the variables. As it is possible to take a predictive measure (Mean Square Error: MSE) with the original data set, we can see that node impurity is related to each variable. Then at each split, how much this split reduces node impurity is calculated (for regression trees, indeed, the difference between the residual sum of squares (RSS) before and after the split), and then with the 'permuted' dataset, and a comparison can be made over them somehow. In particular, it is expected that the original MSE is smaller, so the difference can be taken. Finally, in order to make the values comparable over variables, they are scaled (Table 3).

Figure 5 presents some of the trees involved in the decision process based on the ensemble method. It is possible to realize that variables involved changes in different trees but also the sequence changes. The p values for every decision node were plotted making possible to understand uncertainty related to the sequence of decisions as well as the normalized range of the dependent y-value (predicted TAN). As theory establishes appropriate combination of decisions promoted by different trees builds the final prediction, but it also makes possible to establish variable relevance in relationship with the dependent variable. This effect was summarized in Figure 6.

Variable	%IncMSE	IncNodePurity
X1711	3.5	1.7
X1710	2.5	1.3
X1709	4.4	1.7
X1712	4.4	1.5
X1713	4.3	1.7
X1715	3.2	1.7
X1456	2.0	2.6
X1714	3.4	1.6
X469	1.5	1.9
X1707	2.6	2.0
X1706	3.6	1.7
X1708	4.0	1.7
X2924	0.6	3.6
X450	1.9	1.4
X454	-1.5	1.9
X1717	3.1	1.5
X1716	4.6	1.8
X1631	3.8	1.7
X452	4.5	3.0
X1705	2.7	1.8
X722	2.5	1.6
X1718	3.3	1.4
X2954	0.7	1.7
X451	2.3	2.2
X953	1.8	1.4

Table 3. Importance of relevant variables.

3.2.3 Band-based featured regression models.

As previously mentioned (section 3.2.2) the success rate for TAN estimation was not good enough even though Figure 4 exhibits a rather good performance. Unfortunately, it was just the average figure for the full set of data after using the cross-validation technique but it is not enough when sets of data never seen before are considered.

The authors believe that such situation is due to the extremely sensitiveness of the models against absorbance at specific wavelengths. In order to increase robustness, regions instead of single spot signals have been considered. Also, a concept of ratio between responses for different regions was considered just to increase the robustness against different gains from different hardware.

Therefore, the concept of feature described in previous sections was introduced and regions of 10 cm^{-1} were defined. The algorithm involves checking different potential regions for signal as continuum type of signals. In order to avoid confusions, any positive intersection between signal and continuum regions was forbidden. Finally, and in order to gain generalization capability, it was allowed to consider a vector of features as the basis for developing the decision model. In this practical case, vectors of ten features have been considered.

After describing the design criteria established to build the new modelling approach, it is necessary to indicate that the idea is to assess different type of models (the same being used in

supporting) on the new features. Now, a wide set of potential features to select from is available. To make possible a convenient way of solving procedures, the authors have introduced a Genetic Algorithm by selecting features that improve the quality of a linear regression based on those features. The specific selected criterion for the model was the Bayes Information Criteria (BIC).

In practical terms, the authors have developed an implementation of the R (http://www.r-project.org) package Genetic Algorithm (GA) to be run in an High Performance computing the UPM owns (CESVIMA), making possible the feature identification.

According to Gracia et al. [6] interpretations of the meaning of signal ranges and continuum ranges (Table 4) can be found as well as ketones/aldehydes, carboxylic acids and esters with ν (C=O) (4 and 8 in Table 4) in the signal range versus no signal or saturated hydrocarbons with δ (C–H) in the continuum range. And according to Abbas et al. [28] interpretations of the meaning of signal ranges and continuum ranges (Table 4) can be found; aromatic compounds with ν (C–O) (6 and 9 in Table 4) in the signal range versus aliphatic compounds $(CH_2)_n$) and aromatic compounds with ν (C=C) in the continuum range. Additionally aromatic compounds with δ (C–H) (5 in Table 4) can be found in the signal range versus no signal in the continuum range.

	Signal range (cm^{-1})	Continuum range (cm^{-1})
1	1110–1119	2180–2189
2	2225–2234	710–719
3	2165–2174	1380–1389
4	1820–1829	2020–2029
5	815–824	1330–1339
6	1270–1279	715–724
7	1565–1574	745–754
8	1880–1880	1365–1374
9	1015–1024	1470–1479
10	1135–1144	1465–1474

Table 4. Feature list identified by the GA.

After identifying the features, 66% of laboratory samples were randomly selected for training the algorithms and 33% were preserved for independent data validation. As those samples were never used during the modelling epoch, model building started over with the training samples by using the same kind of models used (see Supporting).

As the TAN values for the validation samples were known, after producing the models, validation in terms of accuracy was performed. The naïve assessment or reference point was considered as the χ^2 distance criteria between full spectra. An estimation based on the ICA projection method was also performed. In these both cases, the TAN value was selected from the closest spectrum available in the training set.

The criteria representing the error of both RMSE and MAE were used (Figure 6 and section B in Supporting). RMSE values in the validation were from 1.197 for GLM model until 0.759 for PPR model; meanwhile MAE values in the validation were from 0.733 for GLM model until 0.359 for for PPR model. Although RMSE values are very similar, when validation was used the results obtained from PPR model were better. The PPR model was able to predict TAN values reducing the RMSE by 37% (compared to GLM model) and the MAE by 51% (compared to GLM model).

It can be observed that ICA outperforms the very first approach, which is distance, and that the Project Pursuit Regression technique (see Figure 8) preserves maximum generalization capability over other techniques. In those, the initial advantage for bagging used by Random Forest progressively reduces as the features now are much more aggregated than in the previous strategy. In the TAN range lower to 2.5 (mg KOH / g oil) the model had the chance of learning with more samples and it had a better prediction. On the other hand, in the TAN range over 2.5 (mg KOH / g oil) with a lower quantity of samples, the model does not have the elements required in order to learn and interpolate.

Conclusions

Total acid number (TAN) has been considered an important indicator of oil quality of used oils. TAN is determined by potentiometric titration. A more convenient approach for the determination of TAN is based on infrared (IR) spectral data and multivariate regression models.

The spectra of motor oils markedly overlapped within the entire spectral region and no certain wavenumber can be found. These spectra of engine oils corresponded to oils with different acidities, low, medium and high, and showed characteristic changes due to acidity and oxidation. These changes indicate typical oxidation products and aging products caused by contaminants in oil.

We have built predictive models to determine TAN using IR data measured from ashless dispersant oils developed for aviation piston engines (SAE 50). Different techniques, including Projection Pursuit Regression, Partial Least Square, Support Vector Machines, Linear Models and Random Forest, have been used. Instead of using the classic PCA for dimension reduction, we have used both the PCA and the Independent component analysis (ICA), which slightly outperforms the PCA. After dimension reduction, a model based on non-linear regression by Support Vector Machines (SVM) with Gaussian kernel was implemented. In this work, we have used the backward variable selection and we have extracted 25 highly relevant variables. Then predictive models have been built by different techniques, including Projection Pursuit Regression, Partial Least Square, Support Vector Machines, Linear Models and Random Forest. The methodology used involved a five folder cross validation to derive the best type of model. Then a full error measure over the whole dataset was taken. When wider ranges of spectra were considered as feature candidates and the forecast methodology extends to consider fully separated sets for training and validation, including the five cross validation strategy for training, the situation changes slightly, as Random Forest downgrades its performance but Project Pursuit Regression still keeps its performance levels. RMSE and MAE values obtained in the validation were 0.759 and 0.359 for PPR model respectively. The PPR model was able to predict TAN values reducing the RMSE by 37% (compared to GLM model) and the MAE by 51% (compared to GLM model).

It appears as evident that benefits can be expected from increasing the density of data, especially for training of models since otherwise the risk for outliers significantly grows.

Acknowledgements

The authors would like to thank Roland Torres of the Universidad Metropolitana for his collaboration in oil sample processing. BLDR acknowledges financial support from the Venoco Company. The authors also thank the Universidad Politécnica de Madrid for granting access to the CESVIMA (http://www.cesvima.upm.es/) HPC infrastructure. We would also like to thank

the author Beatriz Leal de Rivas (in memoriam), for her efforts to conform this team of researchers from different areas of expertise, and we want to dedicate this work to her loving memory.

References

[1] A.M. Petlyuk, R.J. Adams, Oxidation Stability and Tribological Behavior of Vegetable Oil Hydraulic Fluids, Tribol. T. 47 (2004) 182-187.

[2] ASTM D445-12 Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity), Book of Standards Volume: 05.01, 2012.

[3] ASTM D664-11a Standard Test Method for Acid Number of Petroleum Products by Potentiometric Titration, Book of Standards Volume: 05.01, 2011.

[4] E.-P. Ng, S. Mintova, Quantitative moisture measurements in lubricating oils by FTIR spectroscopy combined with solvent extraction approach, Microchem. J. 98 (2011) 177-185.

[5] C. Besser, N. Dörr, F. Novotny-Farkas, K. Varmuza, G. Allmaier, Tribol. Int. 65 (2013) 37-47.

[6] N. Gracia, S. Thomas, P. Bazin, L. Duponchel, F. Thibault-Starzyk, O. Lerasle, Combination of mid-infrared spectroscopy and chemometric factorization tools to study the oxidation of lubricating base oils, Catal. Today. 155 (2010) 255-260.

[7] A. Villar, S. Fernández, E. Gorritxategi, J.I. Ciria, L.A. Fernández, Optimization of the multivariate calibration of a Vis–NIR sensor for the on-line monitoring of marine diesel engine lubricating oil by variable selection methods, Chemometr. Intell. Lab. 130 (2014) 68-75.

[8] M.A. Al-Ghouti, Y.S. Al-Degs, M. Amer, Determination of motor gasoline adulteration using FTIR spectroscopy and multivariate calibration, Talanta 76 (2008) 1105-1112.

[9] M.A. Al-Ghouti, L. Al-Atoum, Virgin and recycled engine oil differentiation: A spectroscopic study, J. Environ. Manage. 90 (2009) 187-195.

[10] V. Macián, B. Tormos, Y. A. Gómez, J. M. Salavert, Proposal of an FTIR Methodology to Monitor Oxidation Level in Used Engine Oils: Effects of Thermal Degradation and Fuel Dilution, Tribol. T. 55 (2012) 872-882.

[11] M.J. Adams, M.J. Romeo, P. Rawson, FTIR analysis and monitoring of synthetic aviation engine oils, Talanta 73 (2007) 629-634.

[12] F. R. van de Voort, J. Sedman, D. Pinchuk, An Overview of Progress and New Developments in FTIR Lubricant Condition Monitoring Methodology, Journal of ASTM International 8(5) (2011) 103344, DOI: 10.1520/JAI103344.

[13] Y. Felkel, N. Dörr, F. Glatz, K. Varmuza, Determination of the total acid number (TAN) of used gas engine oils by IR and chemometrics applying a combined strategy for variable selection, Chemometr. Intell. Lab. 101 (2010) 14-22.

[14] M.A. Al-Ghouti, Y.S. Al-Degs, M. Amer, Application of chemometrics and FTIR for determination of viscosity index and base number of motor oils, Talanta. 81 (2010) 1096-1101.

[15] J.W.B. Braga, A.A.D.S. Junior, I.S. Martins, Determination of viscosity index in lubricant oils by infrared spectroscopy and PLSR, Fuel. 120 (2014) 171-178.

[16] J. Zieba-Palus, P. Koscielniak, Differentiation of motor oils by infrared spectroscopy and elemental analysis for criminalistic purposes, J. Mol. Struct. 482–483 (1999) 533-538.

[17] A.A. Gridnev, S.D. Ittel, Catalytic Chain Transfer in Free-Radical Polymerizations, Chem. Rev. 101 (2001) 3611-3659.

[18] B.N. Barman, Behavioral differences between group I and group II base oils during thermo-oxidative degradation, Tribol. Int. 35 (2002) 15-26.
[19] H. Martens, T. Naes, Multivariate Calibration, John Wiley & Sons, New York, 1989.
[20] R.M. Balabin, R.Z. Safieva, Motor oil classification by base stock and viscosity based on near infrared (NIR) spectroscopy data, Fuel 87(12) (2008) 2745-2752.
[21] P. Comon Independent component analysis—a new concept? Signal Process. 36: (1994) 287–314
[22] C. Jutten, J. Hérault, Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture, Signal Process. 24 (1991) 1–10.
[23] A. González-Marcos, L. M. Sarro, J. Ordieres, A. Bello. Evaluation of data compression techniques for the inference of stellar atmospheric parameters from high resolution spectra. Mon. Not. R. Astron. Soc. Under review
[24] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[25] M.R. Segal, Machine learning benchmarks and random forest regression. (2004) Available at http://repositories.cdlib.org/cbmb/bench.
[26] G. Biau, L. Devroye, On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, J. Multivariate Anal. 101 (2010) 2499–2518.
[27] J. B. Ghasemi, H. Tavakoli, Application of random forest regression to spectral multivariate calibration. Anal. Methods. 5(7) (2013) 1863–1871.
[28] O. Abbas, C. Rebufa, N. Dupuy, A. Permanyer, J. Kister, Assessing petroleum oils biodegradation by chemometric analysis of spectroscopic data. Talanta. 75(4) (2008) 857–871.

Figure captions

Figure 1. IR spectra of three typical used oil samples, low acidity (0.86 mg KOH) (thin line), medium acidity (2.02 mg KOH) and high acidity (3.53 mg KOH) (thick line), showing characteristic changes in the spectrum due to acidity and oxidation.

Figure 2. Relevance of variables to identify the number of relevant variables.

Figure 3. Predicted, using a Random Forest type model, vs. measured values for TAN values.

Figure 4. Error evolution against the number of trees included in the ensemble regressor.

Figure 5. Trees involved in the decision process in the Random Forest model for a) tree number 1; b) tree number 10 and c) tree number 20.

Figure 6. Performance of models for validation dataset. a) RMSE error estimation and b) MAE error estimation. Continuous line for $\chi 2$ regression and dashed line for ICA regression

Figure 7. Forecasting of PPR model when validation data was considered. It becomes clear that more samples are required for high TAN values, otherwise, they will become outliers.

Figure 1

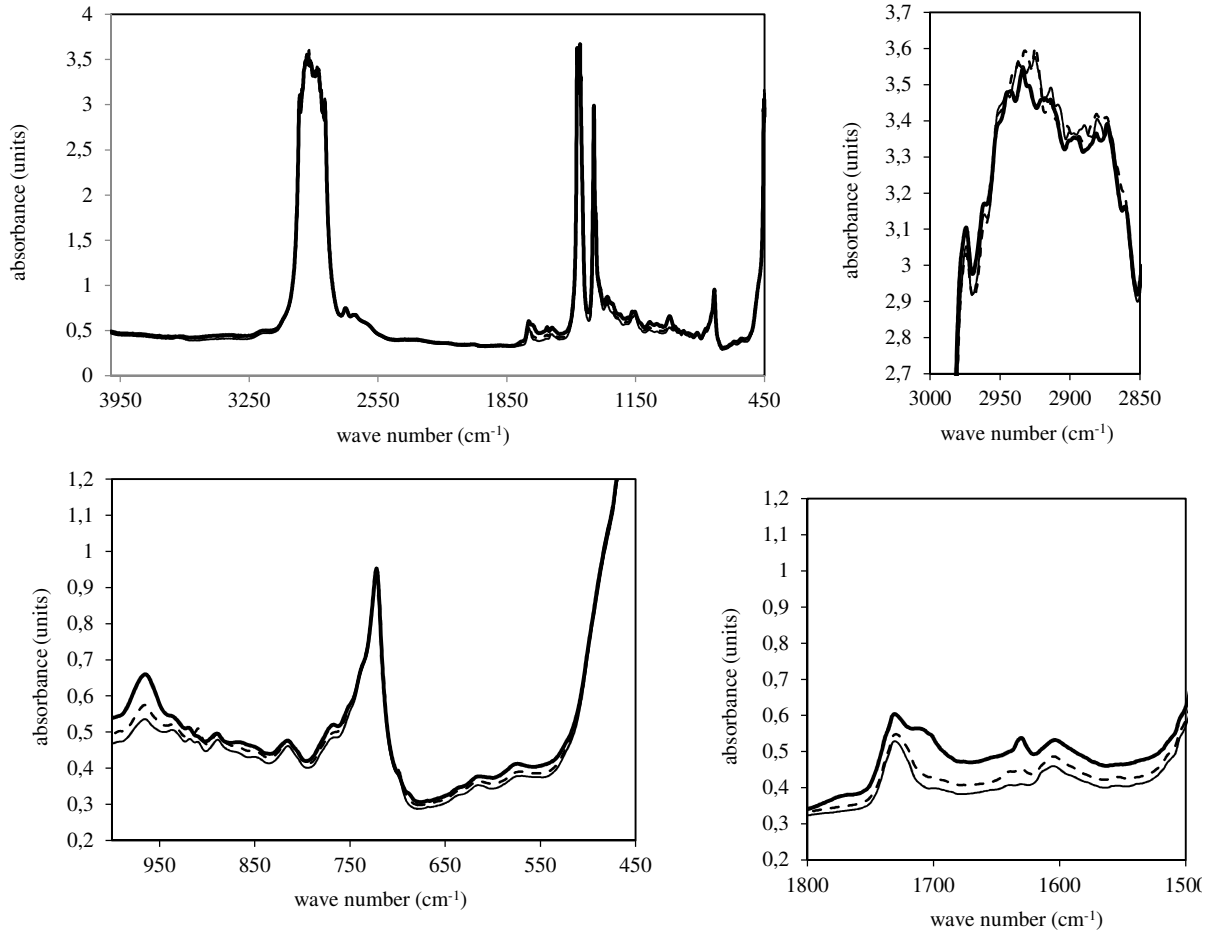

absorbance (units)
wave number (cm-1)
3950
3250
2550
1850
1150
450
3000
2950
2900
2850
950
850
750
650
550
1800
1700
1600

Figure 2

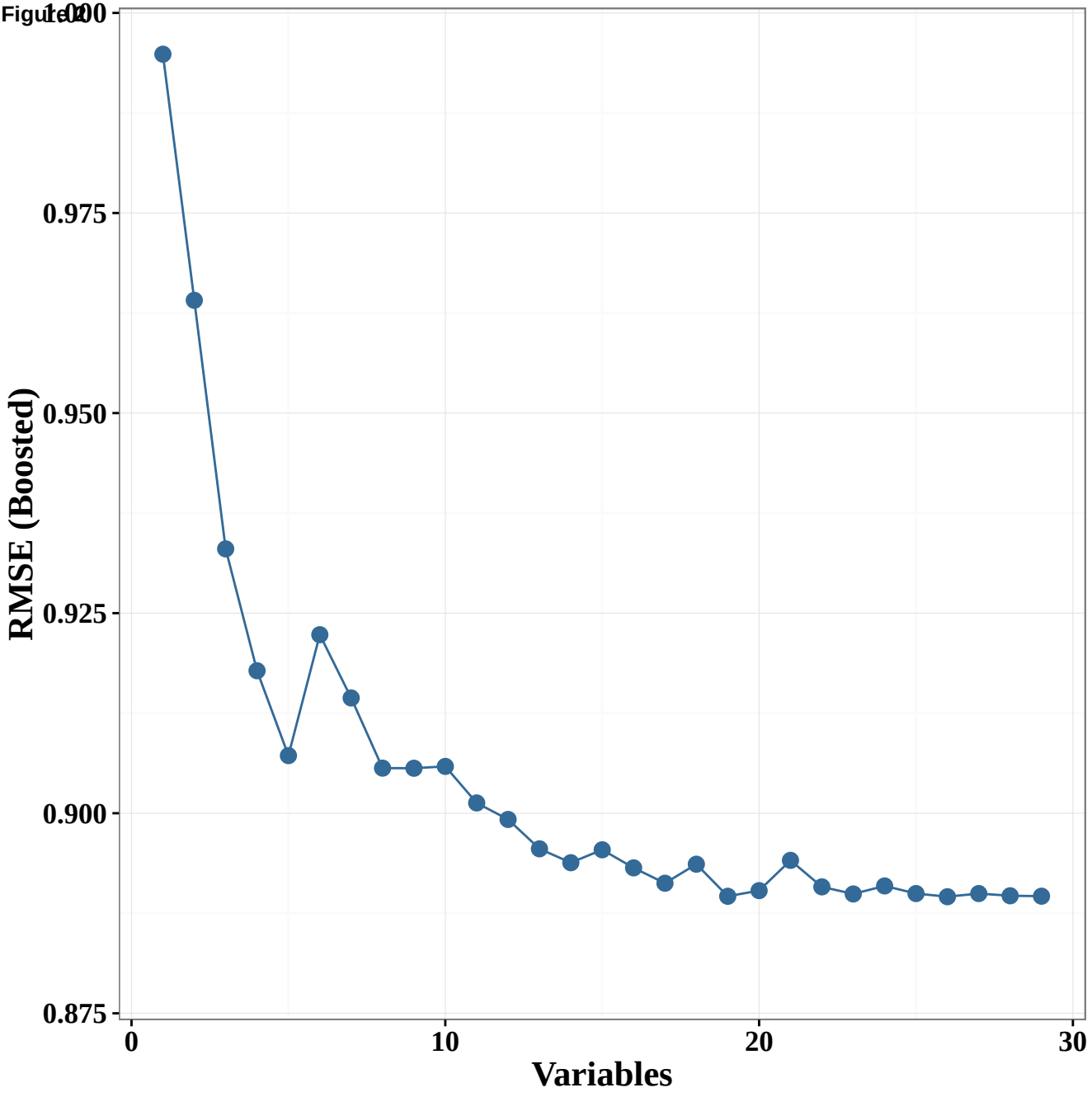

1.000
0.975
0.950
0.925
0.900
0.875
RMSE (Boosted)
0
10
20
30
Variables

Figure 3

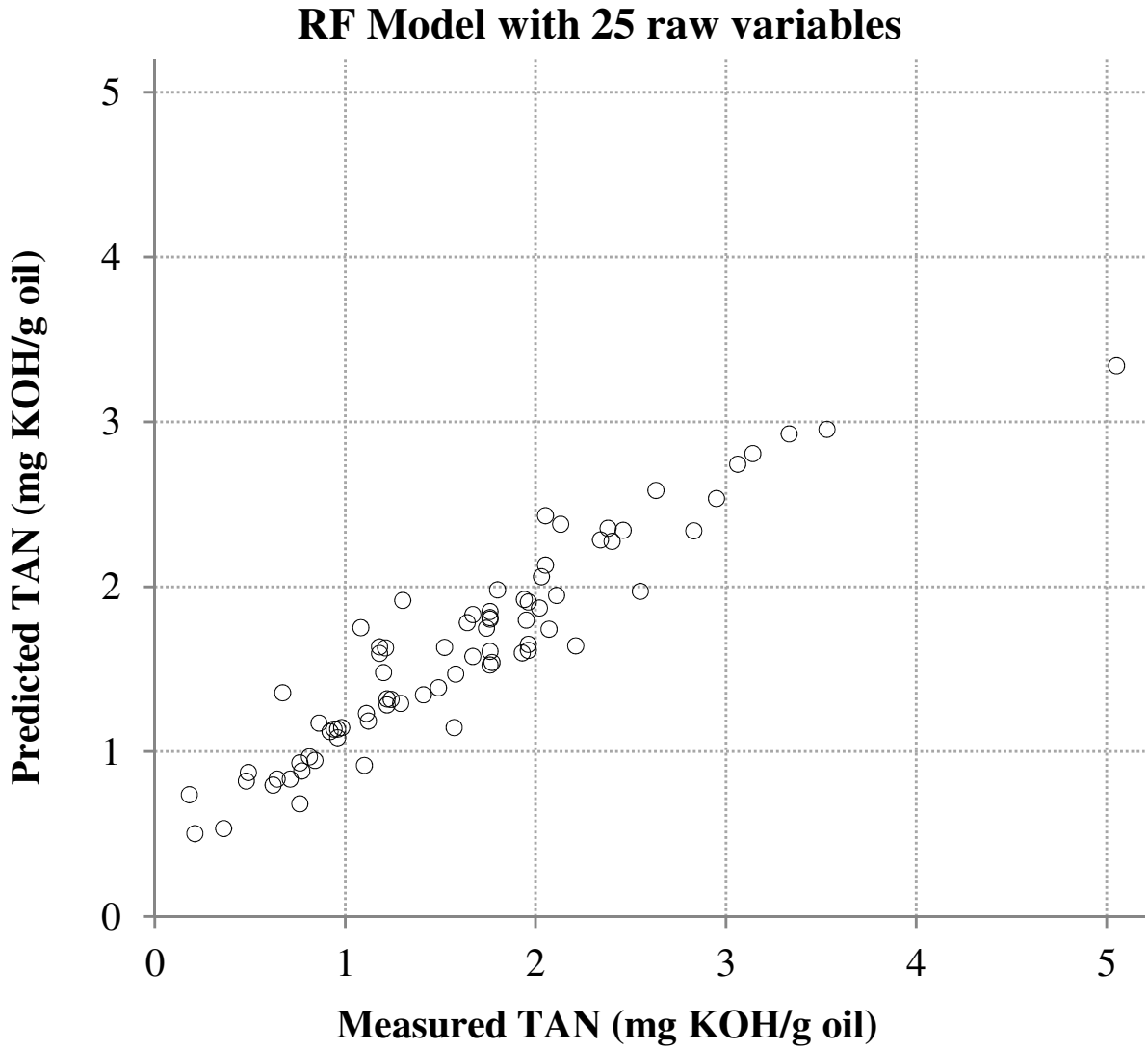

RF Model with 25 raw variables
Predicted TAN (mg KOH/g oil)
Measured TAN (mg KOH/g oil)
0
1
2
3
4
5

Figure 4

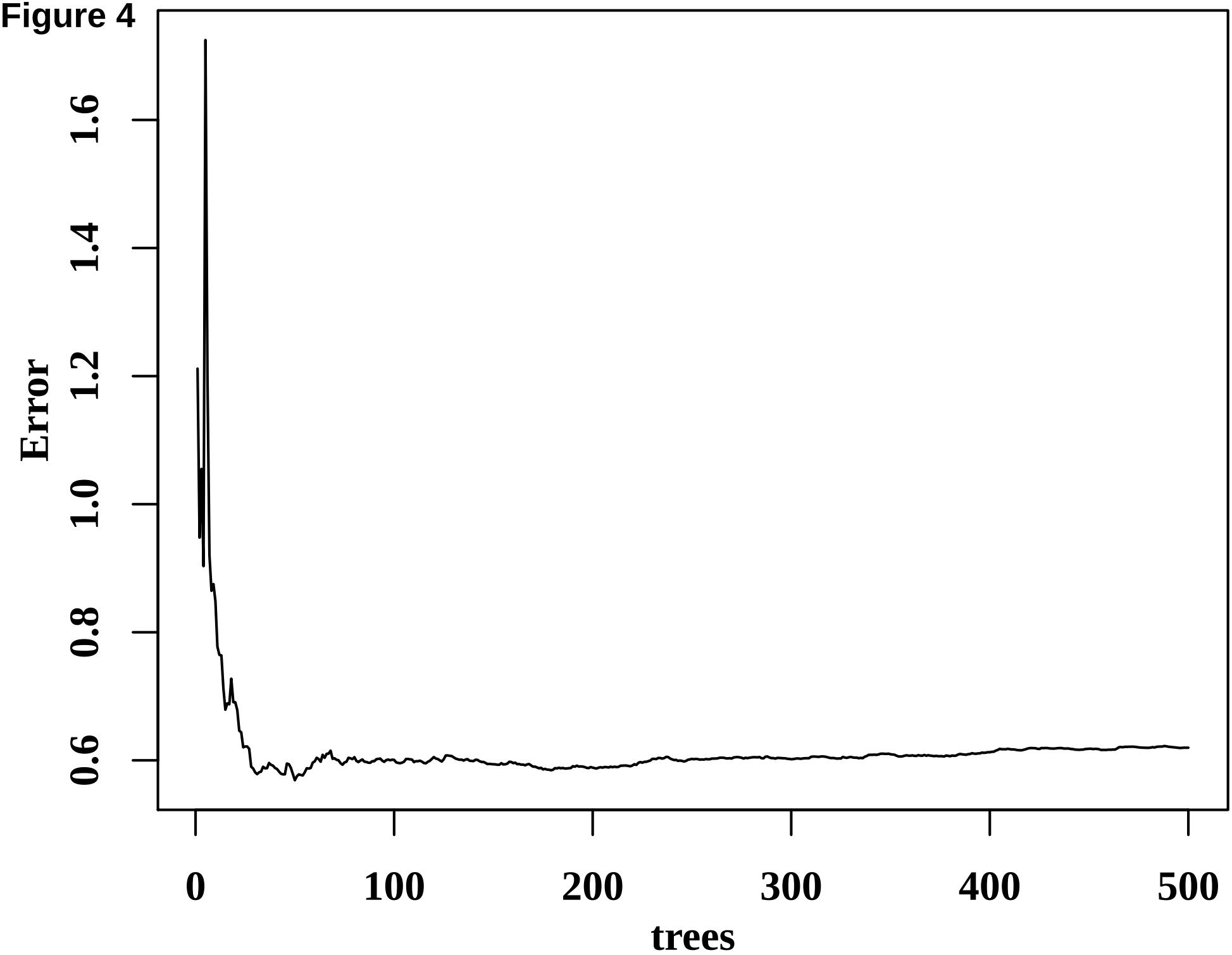


Figure 5a_Trees number 1 Random Forest

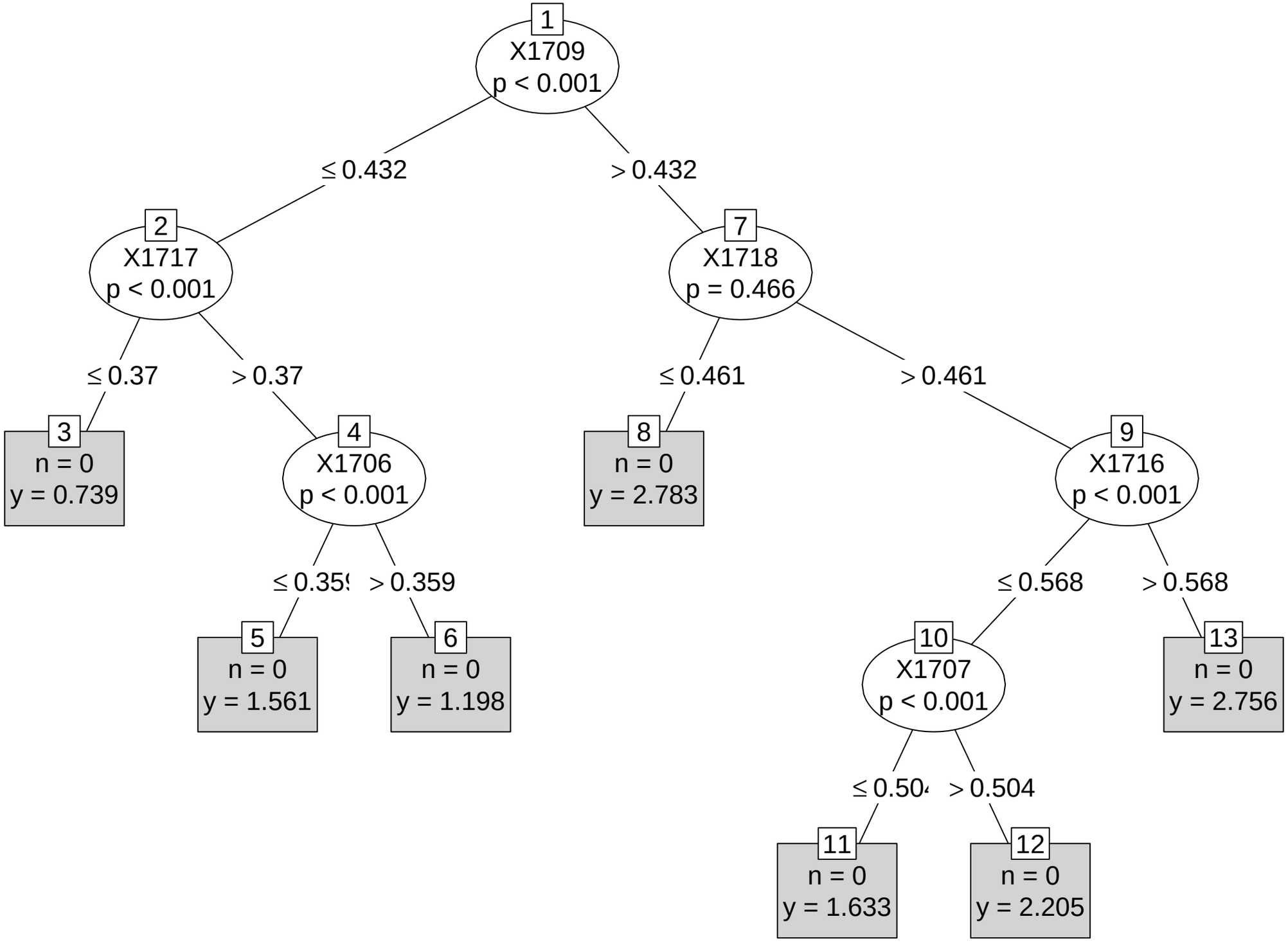


Figure 5b_Trees number 10 Random Forest

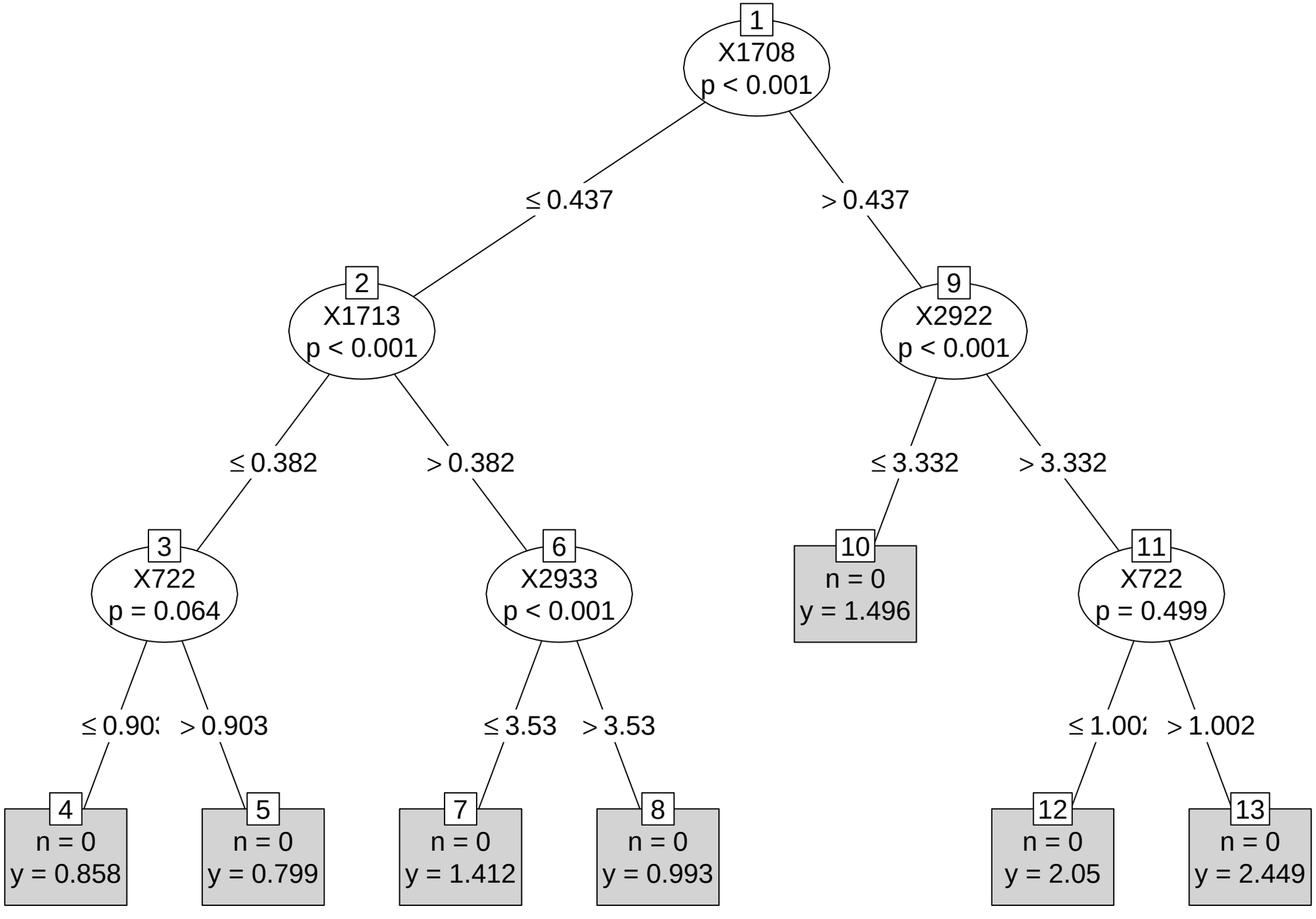


Figure 5c_Trees number 20 Random Forest

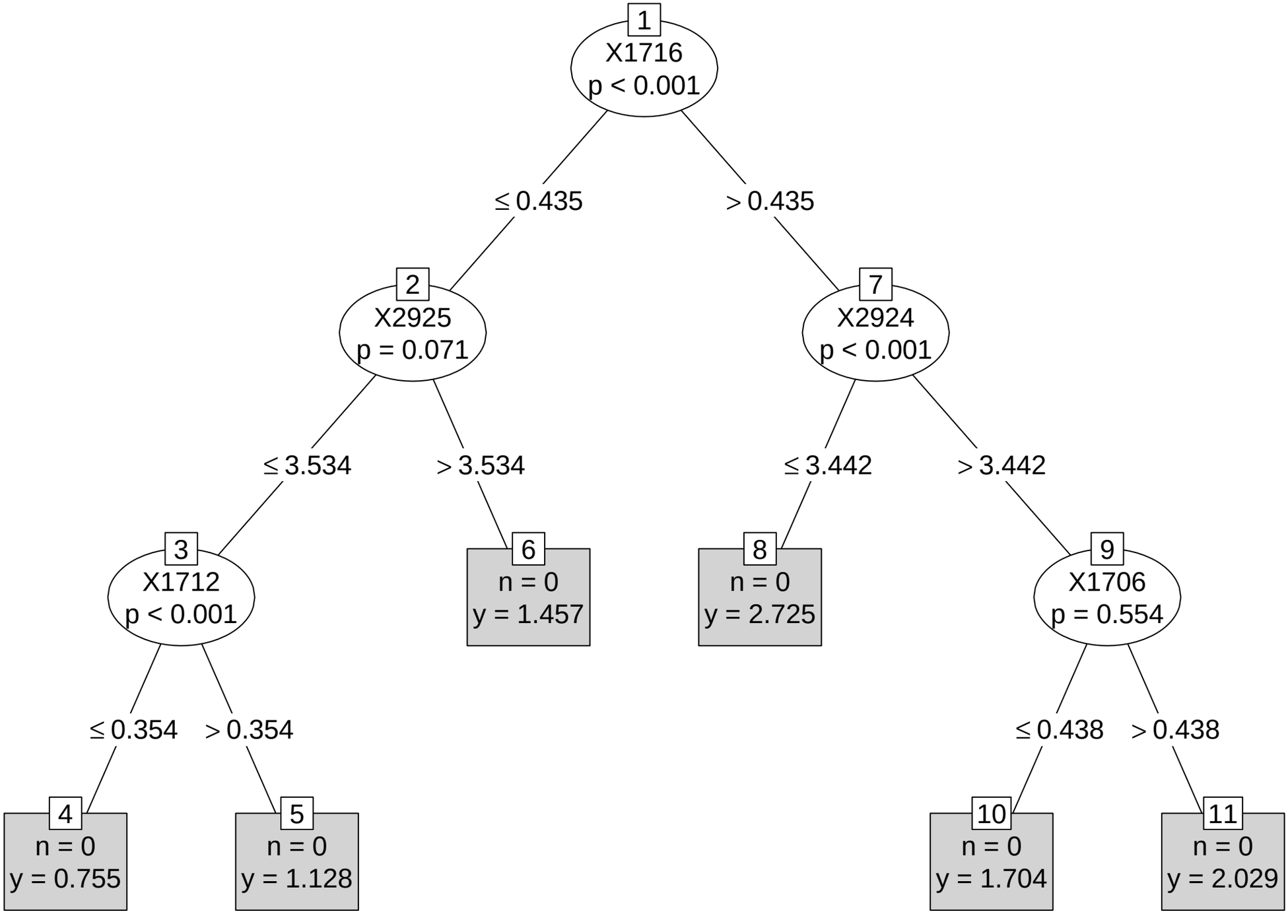


Figure 6a. RMSE error estimation for validation dataset

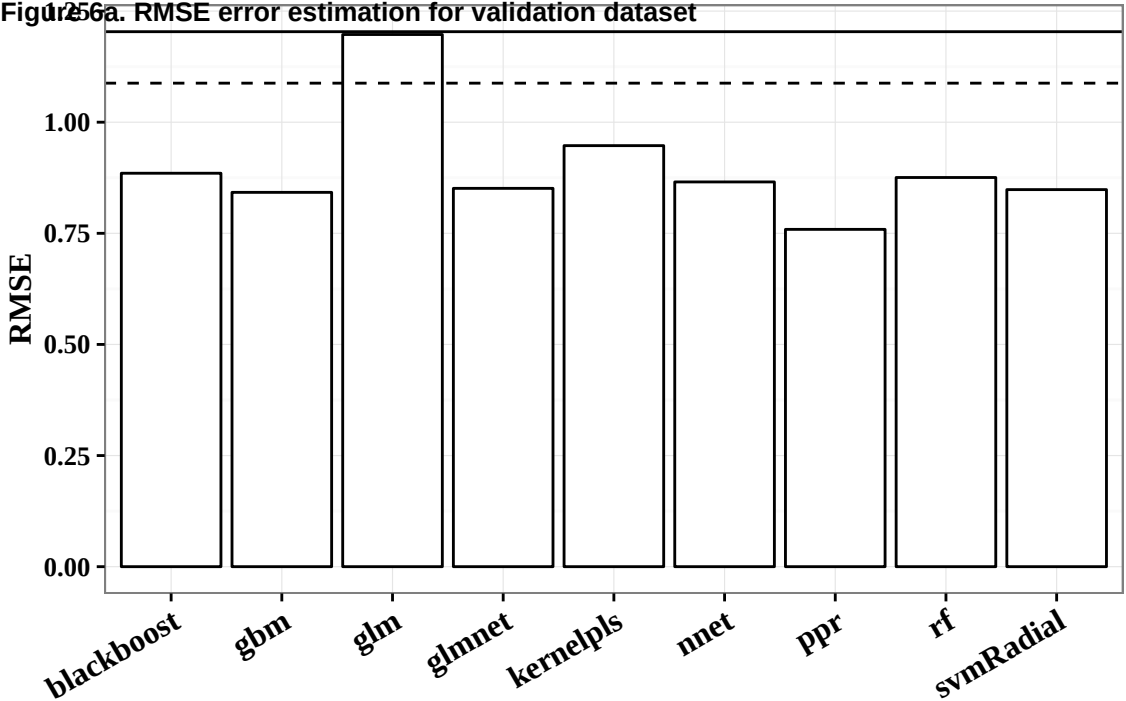


Figure 6b. MAE error estimation for validation dataset

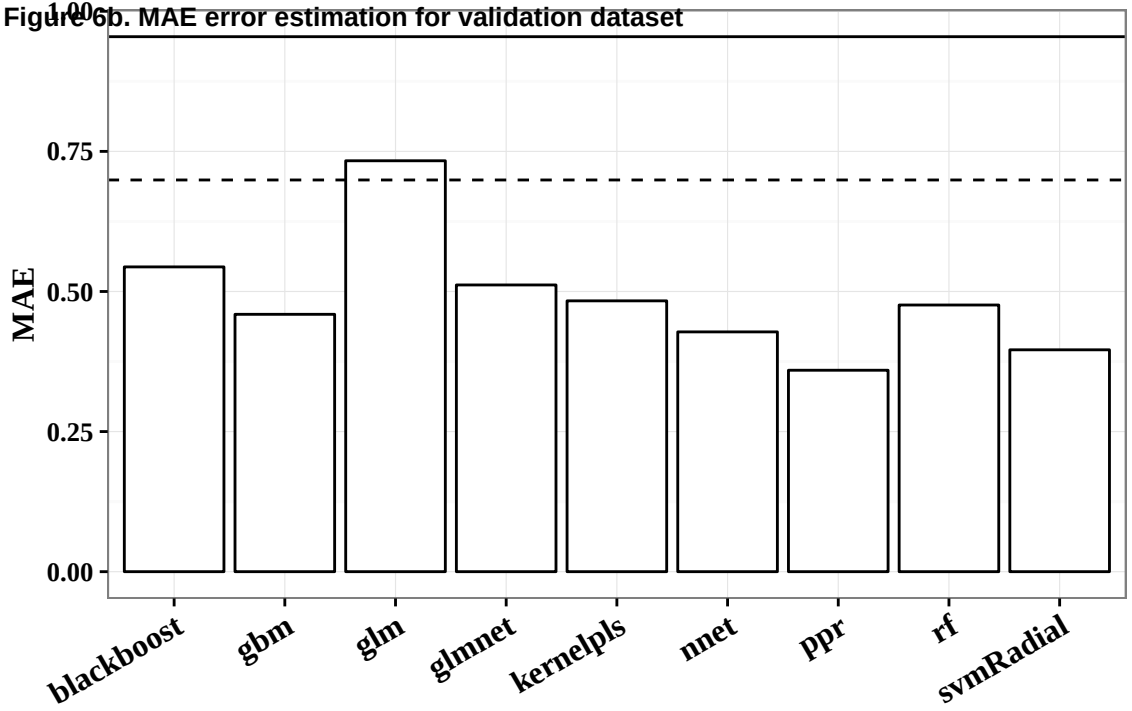
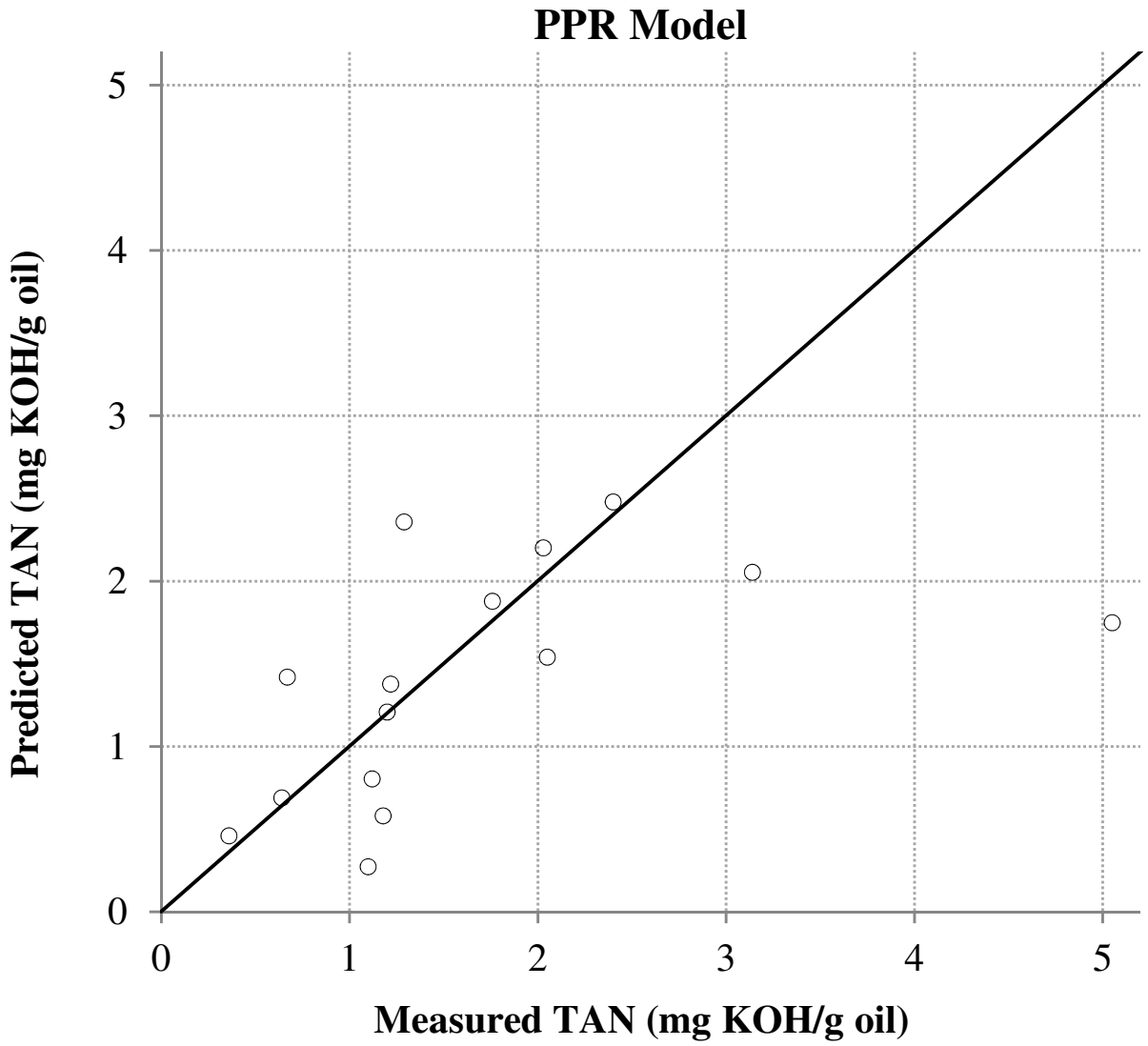


Figure 7 Forecasting of PPR model



Determination of the total acid number (TAN) of used mineral oils in aviation engines by FTIR using regression models

Beatriz Leal De Rivas, José-Luis Vivancos, ,Joaquín Ordieres-Meré, Salvador F. Capuz-Rizo

SUPPORTING INFORMATION

A. Regression Models

As the main goal was to predict TAN values, regression was considered the proper problem type to be dealt with. Then several regression techniques were tested to determine which was the most successful to be used hereinafter.

From our point of view, Linear regression and Projection Pursuit Regression (PPR) can be regarded as two extremes; one is totally rigid in its adherence to an assumed structure, while the other is completely flexible and allows a linear piecewise approach. Projections of the observed data to their latent structure by PLS were developed by H. Wold (S1). Therefore, both have been used, keeping in mind the ridge expression for the linear model in order to reduce the multicollinearity problems.

Partial Least Squares (PLS) have been used as well, even though they account for a wide range of methods that describe the relations between sets of observed variables by means of latent variables. The underlying assumption of all the PLS methods is that the observed data are generated by a system or process driven by a small number of latent (not directly observed or measured) variables. PLS-regression (PLSR) is the PLS approach in its simplest form, and the two-block predictive PLS is the most widely used in chemistry and technology. PLSR has the desirable property that the precision of the model parameters improves with an increasing number of relevant variables and observations (S2). The PLSR implemented herein was the multipredictive and single explained variable (TAN), which is frequently called the PLS1 model. The best number of latent variables was found in one according to the root mean squared error (RMSE) criterion.

Support vector machines (SVMs) have been widely applied in regression analysis, therefore, they were taken into consideration as well. SVMs are a very specific class of algorithms characterized by allowing the usage of kernels, absence of local minima, sparseness of solution and capacity control achieved by acting on the margin, or on a number of support vectors, etc. The system’s capacity is controlled by parameters that do not depend on the dimensionality of the feature space. In our case, the selected approach was the non-linear SVM with a radial kernel. The parameters selected were sigma=0.0816 and RMSE was used to select the C parameter, 0.25 in our case. We fixed the epsilon parameter to the classical value of 0.1. Our design used a bootstrap strategy and there were 72 support vectors.

Multivariate Adaptive Regression Splines (MARS) is a non-parametric non-linear regression procedure that makes no assumption about the underlying functional relationship between dependent and independent variables. The method is based on the "divide and conquer" strategy, which partitions the input space into regions, each with its own regression equation.

For this example, the number or regions was fixed to six. A boosted regression is a recent data mining technique that has been proven considerably successful in predictive accuracy and it was selected to compare against other more complex nonaggregated strategies. It is based on the idea that it is easier to find and average many rough rules of thumb than a single, highly accurate prediction rule (S3). In this case, the boosting method was used in combination with a gradient technique to provide the relevance of all the variables shown in Table S1.

Random forests (S4), which are a combination of tree predictors using both the bagging and randomization approaches (ensemble of models), have received considerable attention in recent years and they combine many trees to form a forest for analyses. An individual tree represents a model describing the characteristics of an input feature present in a subset of the whole dataset. The accuracy of RF has proven competitive with many other data-mining techniques (S5, S6). Biau and Devroye (S7) discussed the links between the layered nearest neighbour estimate and RF estimates, and proved the universal consistency of the bagged nearest neighbour method for regression and classification. The results indicate that besides RF's comparable accuracy and mathematical simplicity, they are computationally fast and robust to noise. In our case, the forest was designed to use 500 trees and each tree randomly involved three input variables.

Variable	Relevance
X454	17.2
X450	11.5
X2925	10.9
X954	6.8
X2864	6.6
X451	6.5
X1454	5.9
X1475	4.4
X452	4.1
X2863	3.7
X1710	3.6
X2924	3.2
X1718	2.6
X1706	2.6
X1711	2.5
X1715	2.4
X1717	2.4
X1709	1.2
X1712	1.1
X1708	0.3
X1713	0.2
X1716	0.1
X1705	0.1
X1707	0.1
X1714	0.0

Table S1. Importance of relevant variables.

Figure S1 shows the averaged relative error obtained after applying different types of models. In Figure S1 results on training data and on cross-validation test data are shown As it can be seen from the figure S1, training errors are higher than the test ones because the value reported for the test is based on the best model found but, represented training error is the averaged errors of all the models considered during the cross-validation. The lower the error over the test set, the

better the behaviour displayed by the model. Therefore, RF is a useful tool for regression studies and has the potential to model linear and non-linear multivariate calibration as it offers good behaviour if compared to many other techniques. Our analysis also supports the same conclusion, even when accuracy is significantly higher, as exhibited in Figure S1.

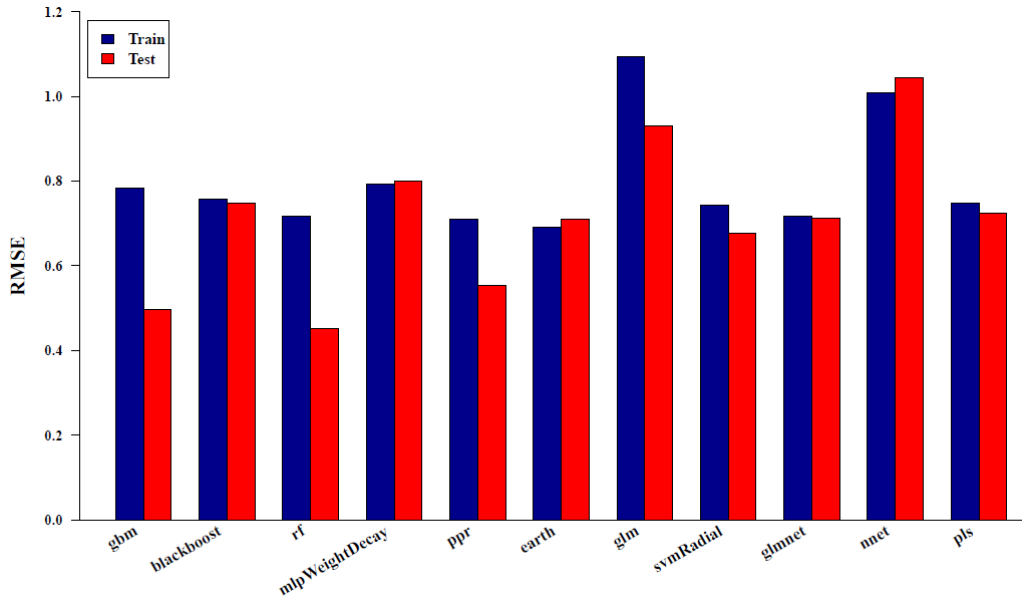


Figure S1. Error for different types of models.

For this particular technique, model accuracy is represented in Figure 4: by way of comparison, the PLS type model was also evaluated and showed a lower correlation (R-square was 0.453), probably due to a poorer fit between the TAN and spectrum changes measured. Finally, the linear model with Tikhonov regularization (Ridge) was also evaluated (R-square was 0.197). Thus, it can be concluded that Random Forest provides acceptable modelling technology with grouped dataset predictions to allow the performance of transformations to fit the measured values. The standard method indicates that for the oils used in Potentiometric Titration, reproducibility can reach an error of 44%, and the developed model can predict within these limits for almost all the samples, except for a few samples (less than 10%) in which they were higher.

B. Band-based featured regression models.

	RMSE	MAE
gbm	0.842	0.459
blackboost	0.885	0.544
rf	0.876	0.476
ppr	0.759	0.359
glm	1.197	0.733
svmRadial	0.848	0.396
glmnet	0.851	0.512
nnet	0.866	0.428
kernelpls	0.947	0.483

Table S2. **Performance of models for validation dataset. RMSE error estimation and MAE error estimation.**

References

[S1] H. Wold. Soft modeling: the basic design and some extensions, in J.-K. Jöreskog, H. Wold, (Eds.), Systems Under Indirect Observation, North Holland, Amsterdam, 1982, vol 2, pp. 1–53.

[S2] H. Wold. Partial least squares, in: S. Kotz, N.L. Johnson, (Eds.), Encyclopedia of the Statistical Sciences, John Wiley & Sons, 1985, vol 6, pp 581–591.

[S3] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab. 58 (2001) 109–130

[S4] R. Schapire, The boosting approach to machine learning – an overview. MSRI Workshop on Nonlinear Estimation and Classification, (eds D.D. Denison, M. H. Hansen, C. Holmes, B. Mallick & B. Yu). Springer, New York, 2002

[S5] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[S6] M.R. Segal, Machine learning benchmarks and random forest regression. (2004) Available at http://repositories.cdlib.org/cbmb/bench.

[S7] G. Biau, L. Devroye, On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification, J. Multivariate Anal. 101 (2010) 2499–2518.