**1. Introduction**

Aromatic compounds are the hydrocarbons derived from petroleum. The most common are benzene, toluene and xylene. Near-infrared spectroscopy illuminates a substance with a broad spectrum of near-infrared light (**NIR**). NIR can identify chemical properties based on the infrared colours detected and their intensity.

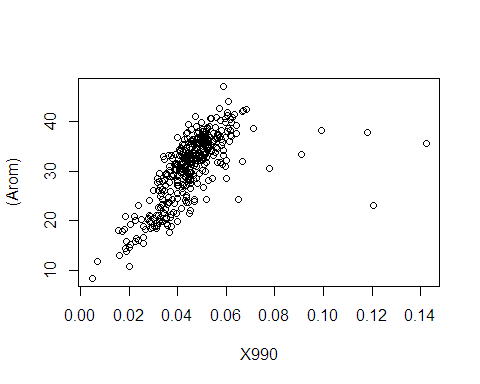
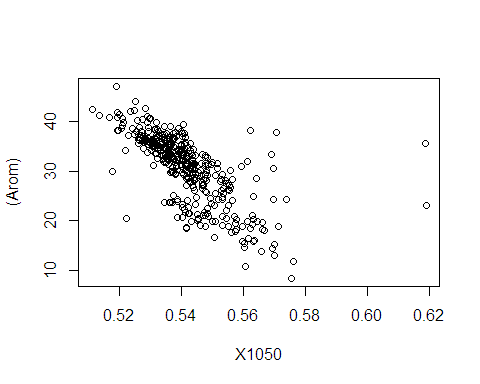
The aim of this paper is to predict the mass percentage of total aromatic compounds in diesel fuels using NIR spectroscopy in the supplied data set. The wavelength range of the spectra is from 950-1398nm, increasing at 2nm intervals. The data will be explored in search of trends, correlations and potential outliers. It will be separated into training and testing data sets where single and multiple linear regression will be performed, along with analysis of the goodness of fit and predictive value of each model.

An accurate model will allow for onsite calibration and real time prediction of the aromatic compounds of diesel fuel.

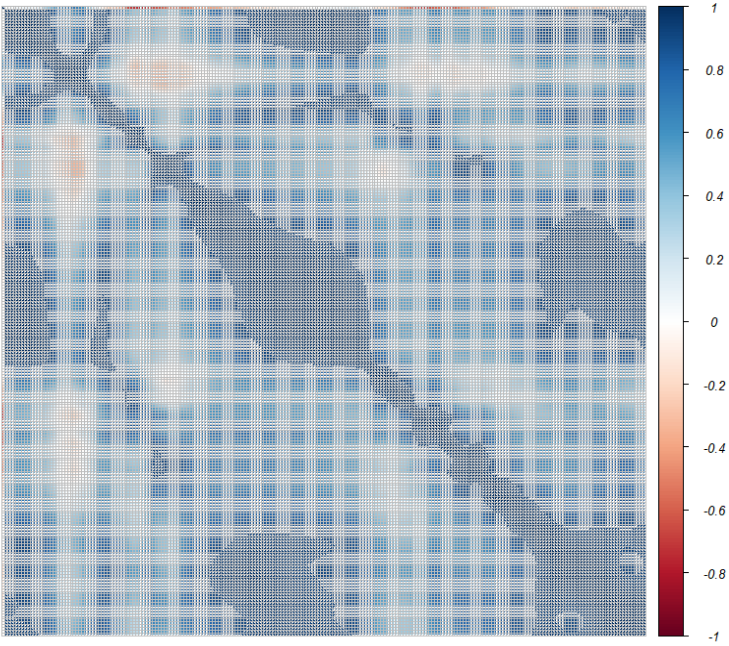
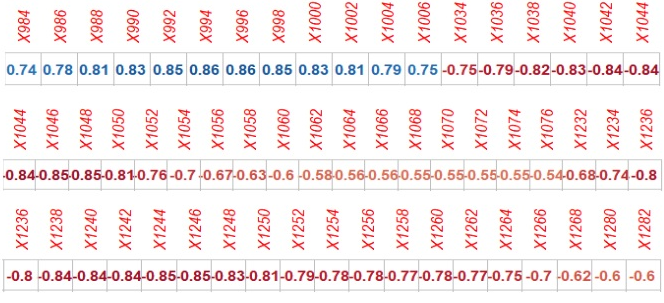
**2. Exploratory Analysis**

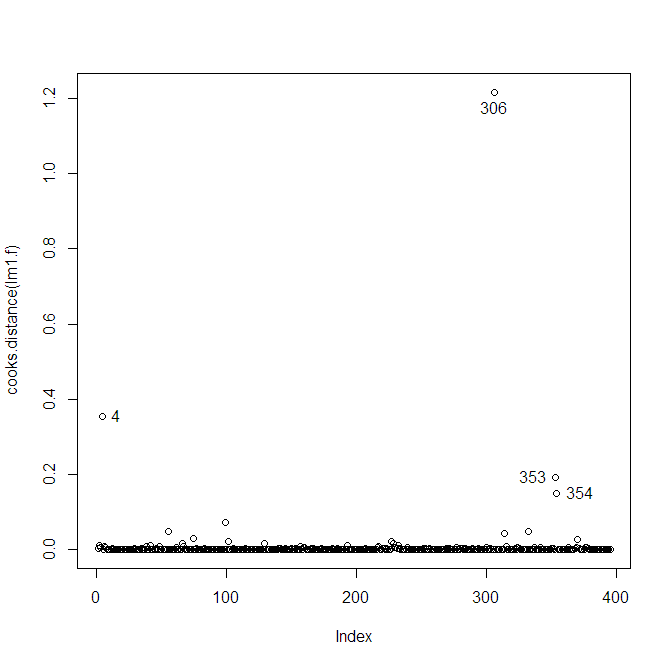
The aromatic data started as a vector of the aromatic response variable, along with a matrix of the spectra results. These were combined to make a single data set, allowing for simpler manipulation and visualisation. To begin the exploratory data analysis, scatterplots were produced of all individual wavelengths against the response variable, Arom.

The graphs were then reviewed manually in search of apparent positive and negative correlations. The wavelengths with the strongest linear relationship appeared in the range of X990 to X1090.



For further clarification, Pearson’s Correlation Coefficient was produced to identify strong linear relationships between the spectra and aromatic levels, those which returned values between ± 0.50 and ±1. All wavelengths outside these values were removed from further exploration and analysis.

As expected due to the nature of the wavelengths being sequential, there is multicollinearity amongst the spectra. This will not reduce the predictive power of models generated as a whole but may alter the impact of calculations specific to individual predictors. Aside from this, there is evidence of strong linear relationships between Arom and the spectra.

A linear model with Arom against all wavelengths was generated. An analysis using Q-Q plots, Residuals vs Fitted, Scale-Location, and Residuals vs Leverage were undertaken. Given the fit of the QQ line, the spread of variability and the leverage, there were some outliers. To validate those identified, plots of Cook’s Distance were made for comparison. This process revealed six observations that were deemed outliers. From this a new dataset was created which removed the 4th, 306th, 218th, 332nd, 353rd and 354th observations removed.

The linear model was refitted on the new dataset for further analysis. There was clear improvement with a more normalised fit and the residuals having less leverage. Although there were other observations that may be outliers in the data, and removing them would return a more optimal dataset, a cautious route was chosen, as removing them might affect the other spectra and actually damage their linear relationship to Arom. For this reason, those points were kept in and the number of observations removed was limited to avoid biasing our data.

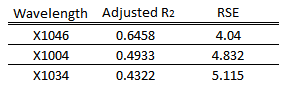
Further tests involving various transformation methods, such as log, square root, cube root, squared and cubed, were conducted in search of underlying trends. Using the boxcox methodologies, an optimal transformation of approximately 0.45 was recommended. When future linear models were tested, the square root transformation yielded lower R2 and overall suboptimal results, so it was decided to proceed with the data untransformed.

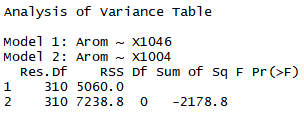
Finally, an integral part of predictive modelling is separating the data into two subsets; train and test. The goal is to use the model to predict future values. As there is no means of accessing ‘future data’, the creation of a test set acts in its place. This allows for tangible measurements on the accuracy of the model, while attempting to lower overfitting and bias.

**3. Simple Linear Regression**

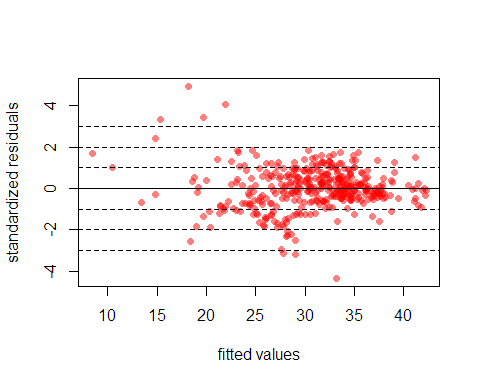
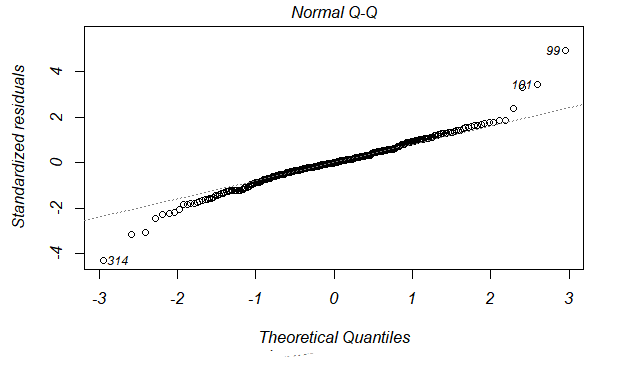
To find the best explanatory variable along with the intercept in a simple linear regression (**SLR**), the regression subset selection method was used in R. X1046 was returned as the best wavelength for a model with Arom, followed by X1034 and X1004.

Each of the listed wavelengths were fitted with a linear model and produced the following results:

There is a significant difference in the R2 of X1046 compared with the next best of X1004. X1046 is able to explain approximately 64.6% of the variance aromatic compound mass percentage in diesel fuel, while X1004 only allows for 49.3%. A lower Residual Standard Error (RSE) is observed for X1046, meaning the model will have clustered residuals closer to the actual points of the predicted data.

To view whether the reduction of the RSS between models is significant, ANOVA was reviewed on X1046 and X1004, which as expected showed a significant reduction of RSS by approximately 30%.  
  
ANOVA also automatically sets up a hypothesis for interpretation, the null stating that β1 = 0, and the alternative hypothesis being β1 ≠ 0. An individual ANOVA on X1046 reveals an F-test value of 568.08, from which gives us strong evidence to reject the null in favour for the alternative, that β1 ≠ 0. Hence, there is a confirmed and validated linear relationship between X1046 and Arom.  
  


The next step was to perform diagnostics on the standardised residuals. It remained clear that within these plots there are still some observations of concern; points that are greater than 3 standard deviations away from the mean. The Q-Q plot shows that most points fit the linear line, reflecting a normal distribution of the data, however both tails do deviate, indicating more data is located at the extremes of the distribution. The standardised residual plot shows that the points form an almost even cloud shape, implying an even spread of variance between the variable X1046 and the response variable Arom.

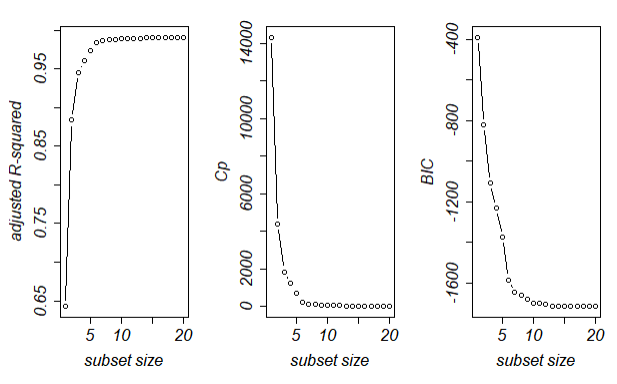


As previously mentioned, removing these outliers would provide a better fit but risk damaging the integrity of the data as a whole. Given this data, the best simple linear model is that which was created with X1046 and is deemed as acceptable.

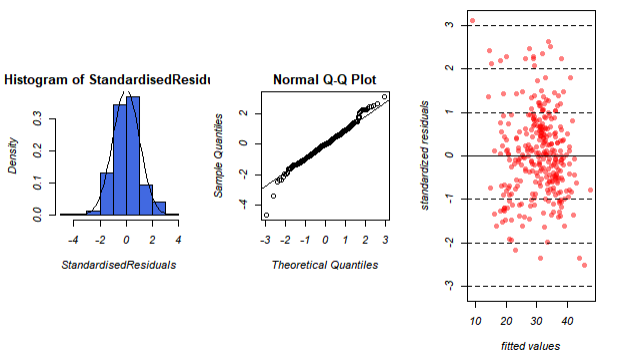
**4. Multiple linear regression model**

The recent work of [Hastie et al. (2017)](https://arxiv.org/pdf/1707.08692.pdf) has emphasised that forward stepwise selection and best subset selection perform similarly, and least absolute shrinkage and selection operator (**LASSO**) out-performs both. Given this information, forward stepwise selection and LASSO will be performed as the multiple linear regression on this dataset, followed by a comparison of results.

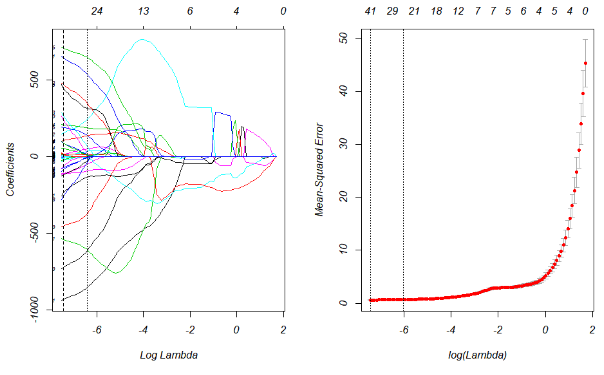
The forward stepwise method created a model utilising 30 of the available 56 explanatory variables, an adjusted R2 of 0.9912 and an RSE 0.6363. At first glance these results look significantly better than those of the SLR, though this may be a case of our model being over fitted to our data and tailoring too much to the noise of the data, meaning our model would not be as accurate as the true model of the population. The summary output has multiple variables with a small t-value, implying their coefficients may only be coincidentally significant.

To further investigate this concern the R2, Mallow’s CP and BIC (a more penalised version of the AIC), were reviewed against varying predictor subset sizes. There are visible diminishing returns once the model reaches seven predictors, with no apparent benefit in increasing the model size to more than eight variables. This enforces the idea that the forward stepwise model has too many superfluous predictors assigned, negating the value of the results. Reviewing the ANOVA table highlights five variables with insignificant probability, but reinforces our initial claim of X1046 being the most influential explanatory variable with an F-value of 22901. This is followed by X1006 with 3266, X1242 with 882 and X990 at 295. The RSS is a comparatively small 113.8.

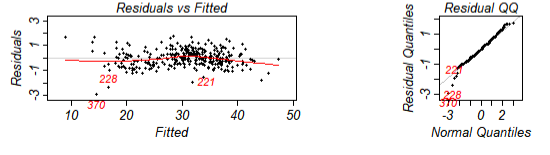
Additional refinement could be performed, such as retesting the model each time the single variable with the lowest t-scores is dropped and comparing the results. This would act as a manual backwards stepwise model with the goal of making the model parsimonious.

The diagnostic plots of the standardised residuals reveal an approximate normal distribution with an even spread of variance. There appear to still be some potential outliers, specifically on the low extreme of the values giving the Q-Q Plot a ‘fat tail’. This is not significant enough to heavily impact the reliability of the data.

Next, the LASSO regression analysis method was applied to the data. LASSO uses variables selection, shrinkage and a penalising system on the absolute size of the regression coefficient. Through LASSO imposing a constraint on the sum of the absolute values with a specified value as an upper bound, this causes the regression coefficients of some variables to shrink towards 0, potentially being removed. This identifies the explanatory variables with the strongest relationship to the response variable, in this case Arom. The tuning parameter λ is applied to the model. As λ increases, more coefficients are reduced to 0, and bias increases. While λ decreases, variance increases. This is especially useful on data sets containing predictors with strong correlation.

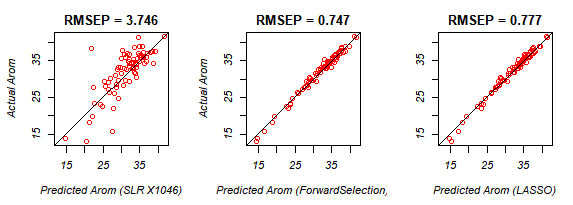
Using a cross-validation method and based on a log λ of -6, the resulting outputs state a model with approximately 41 of the 55 predictors as the maximum benefit, while recommending approximately 25 variables for a more parsimonious fit. This aligns with the ANOVA review of the forward stepwise model, but contradicts the best subset results. The variable with the largest coefficient in the LASSO model was X984, -795.52.

Diagnostics of the LASSO regression show an even spread of variance and an approximate normal distribution, with the same potential issue of outliers as seen in the low end of the Q-Q tail.



**5. Conclusion**

To validate and select the most appropriate model, predictions must be generated and tested on an independent data set. Using the previously generated test data set, each model predicted Arom values, which were then graphed for visual validation.



It is clear that the single linear model using X1046 performed the worst, returning a root mean square error of prediction (**RMSEP**) of 3.746. RMSEP is often referred to as being the simplest and most efficient measure of the predictive levels [when using NIR technology](http://www.aunir.co.uk/light-bites/aunir-glossary-rmsep/). The RMSEP can be used to express the average variation that can be expected when predicting new samples, therefore the lower relative number implies a better potential predictor. Both visually and through calculating the RMSEP, the forward selection model and LASSO performed very similarly. LASSO produced equivalent adjusted R2 of approximately 0.99, while forward selection had the superior 0.9912.

When comparing the AIC of LASSO at -160.65, and forward selection of -252.74, there was a difference of approximately 36%, meaning forward selection is the model with the least probability of information lost and the best fit of the two. Though LASSO was expected to generate the optimal predictive model automatically, the model selected is the forward selection due to having the lowest RMSEP and AIC.

It is a cautious recommendation as forward selection does retain the highest number of predictive variables out of all the models which, incidentally, may be overfitting or registering trends not true to the entire population. In the [statistical industry](https://link.springer.com/article/10.1057/jt.2009.26), stepwise has a reputation of producing biased models with overly optimistic confidence intervals. For a more accurate and complete model, further analysis is required on the data, including but not limited to least-angle regression (**LARS**), partial least squares (**PLS**), and a statistician individually assigning and predicting variables deemed most reliable.

**6. Appendix**

**Analysing the Models**

When trying to compute which model has the best fit, it is important to determine the model that explains the most variability within the dataset while simultaneously using the least number of variables as possible. This is referred to as the parsimonious model.

The baseline test for goodness of fit is to look at the coefficient of determination, also known as the R2. It is a statistical measurement of how well the regression line fits to the data points, and how much of the variability within the data set is explained by the model, with 1 being a perfect fit, and 0 being the poorest fit. Although, it is unlikely to see a perfect fit in a real-world scenario.

A problem with R2 is that every time a predictor is added the R2 will only ever increase, not decrease. This may be misleading and give the impression that the model fits better than it does, purely due to the number of terms. Given the volume of explanatory variables tested within our data set it would be more prudent to review the adjusted R2.

The adjusted R2 is a modified version of the R2 that takes the number of explanatory variables into account. The adjusted R2 will only increase if the number of terms added explain more variability than would be expected by chance alone. Similarly, it decreases if an explanatory variable improves the model less than expected from chance.

Another estimator of the quality of model fit is the Akaike Information Criterion (**AIC**). It does not reflect the absolute quality of a model, but rather estimates the fit of model relative to other models. The AIC should be a used as a supplementary test where comparing multiple models, and will be used in the multiple linear regression part of this paper. Comparatively, the model with the lowest AIC is deemed most appropriate.

Finally, analysis of variance (ANOVA) table will be used to compare the residual variance left by the models in relation to the residual sum of squares (RSS), and the probability of this being in account of chance through the use of an F-test.