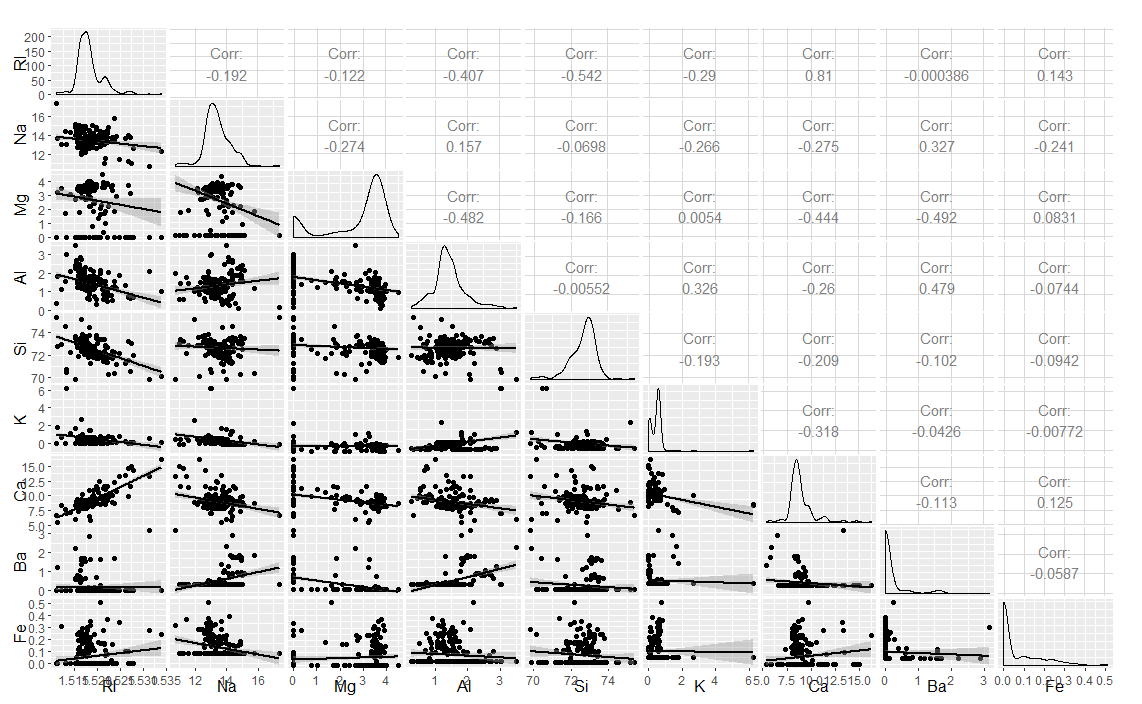
Oct 10, 2016

**Problem1: Glass Identification**

**1(a): Visualizations:**

A number of methods (scatter plot, box plots, adjusted box plots , histograms, etc.) were explored and are included in the R script. One of the methods, ggpairs provides a comprehensive view of the data and shown below.



**1(b): Skewness:**

Skewness was calculated using 2 methods.

Method 1: *apply(Glass[,1:9], 2, skewness)*

Method 2:

*cor(Glass[,1:9],method = "kendall")*

*cor(Glass[,1:9],method = "pearson")*

*cor(Glass[,1:9],method = "spearman")*

Based on the results, Mg, K and Ba were selected for skew transformation

Symbox function was used to calculate appropriate value of λ was determined and symbox function was run again. An example for Mg is shown below.

boxcox(Glass$Mg+0.00001)

symbox(Glass$Mg+0.00001, data= Glass, powers=c(3,2,1,0,-0.5,-1,-2))

symbox(Glass$Mg+0.00001, data= Glass, powers=c(2))

**1(c): PCA**

Summary of PCA is shown below:

Importance of components:

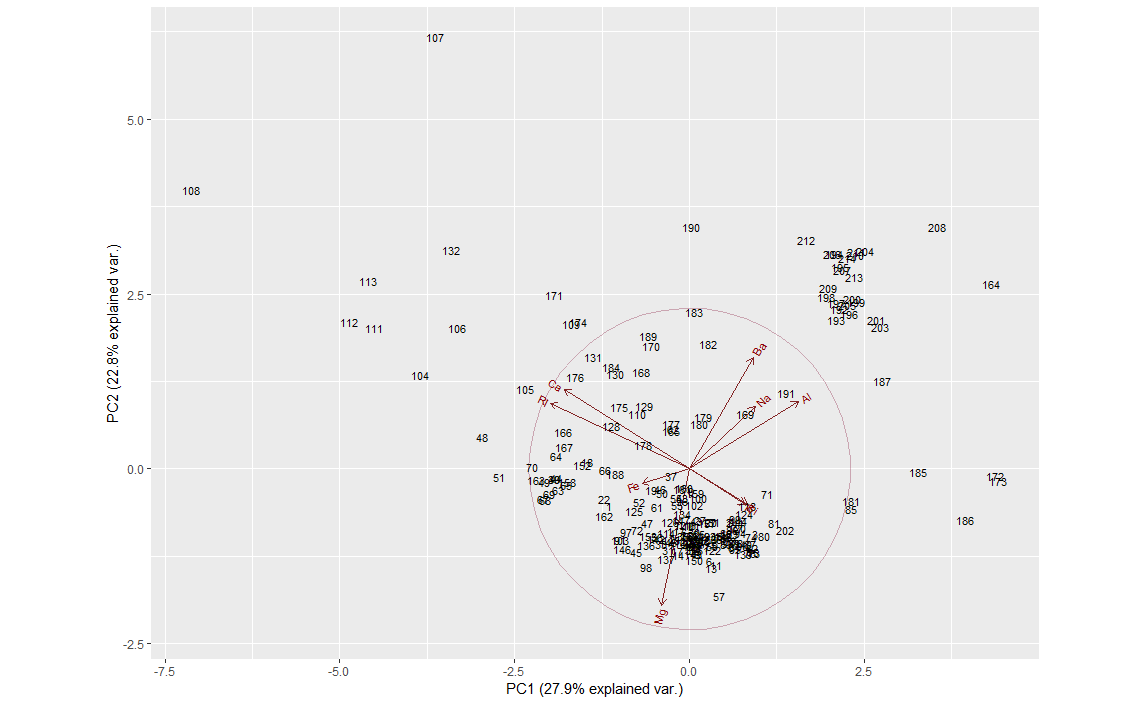
PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8 PC9

Standard deviation 1.585 1.4318 1.1853 1.0760 0.9560 0.72639 0.6074 0.25269 0.04011

Proportion of Variance 0.279 0.2278 0.1561 0.1286 0.1016 0.05863 0.0410 0.00709 0.00018

Cumulative Proportion 0.279 0.5068 0.6629 0.7915 0.8931 0.95173 0.9927 0.999821 1.00000

As shown above, PC1-PC5 account for approx. 90% of variance. This indicates that PC6-PC9 can be dropped. Also, as indicated by ggplots, Ca and RI in the same direction as well as magnitude and can be combined in one.



**1(d): LDA:**

Results of the LDA are shown below:

fit.predict 1 2 3 5 6 7

1 52 17 11 0 1 1

2 15 54 6 5 2 2

3 3 0 0 0 0 0

5 0 3 0 7 0 1

6 0 2 0 0 6 0

7 0 0 0 1 0 25

PCA is an unsupervised learning technique (all variables are treated as independent of others) whereas LDA is supervised technique that takes into account class information. For dataset where variables are independent, PCA is sufficient. However, if variables are related (i.e. one affects the other) it is better to use LDA.

**Problem 2: Missing Data**

**2(a): Regression using listwise deletion:**

Script for this is shown on lines 211-220 of the script file.

Co-efficient: 6.29

**2(b) Regression using mean imputation:**

Script for this is shown on lines 223-236 of the script file.

Coefficient: 10.89

**2(c): Regression using multiple imputation.**

Four different method were tried to ascertain propriety of method for this dataset. There were mean, rf, cart and sample. For some reason, none of these methods were able to genenerate the co-efficient for me.

**Problem 3: Sensor data**

**3(a)**

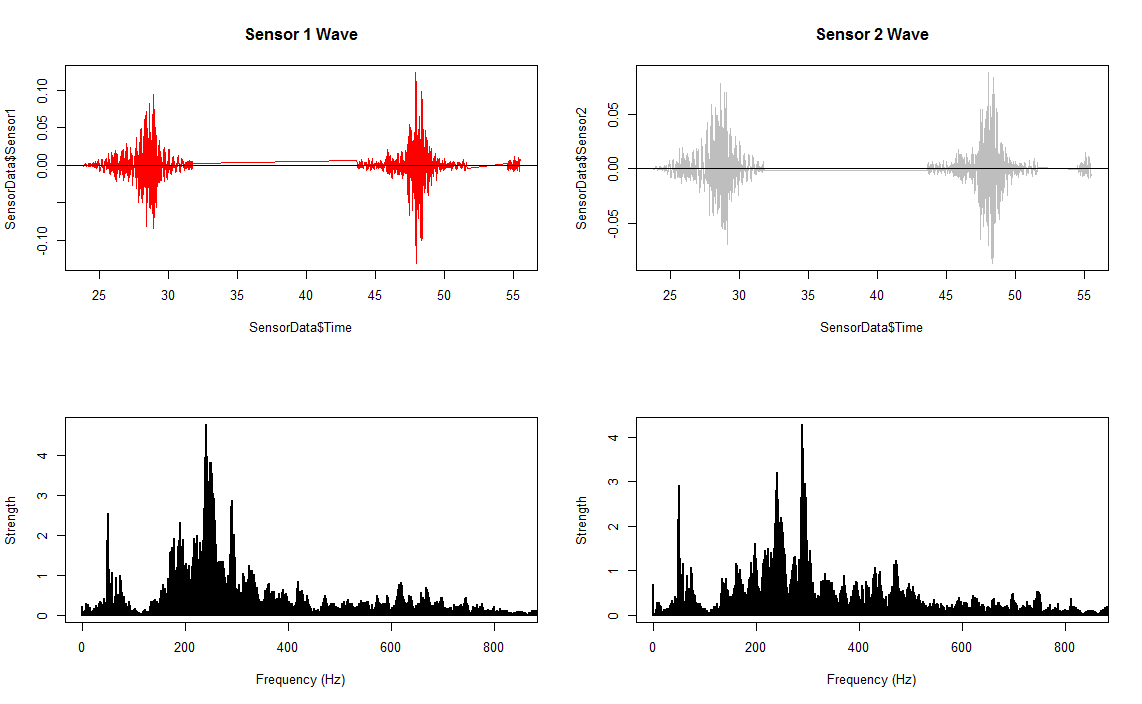
Project Goal: Identification of vehicles based on frequency data.

The frequency data captures the vibrations (frequency, duration and amplitude). Frequency may depend on weight and speed of the vehicale, duration is directly related to the speed of the vehicle and amplitude may depend on the weight (GVW and no. axles). If dataset includes this information it is quite likely that each vehicle crossing the bridge can be reasonably identified.

Features that can be obtained from this limited set of data are frequency histograms (fft), Sensor ratios (may indicate speed – increasing, decreasing stable), sensor duration of vibrations (may also depend on the distance from closest pier), etc.

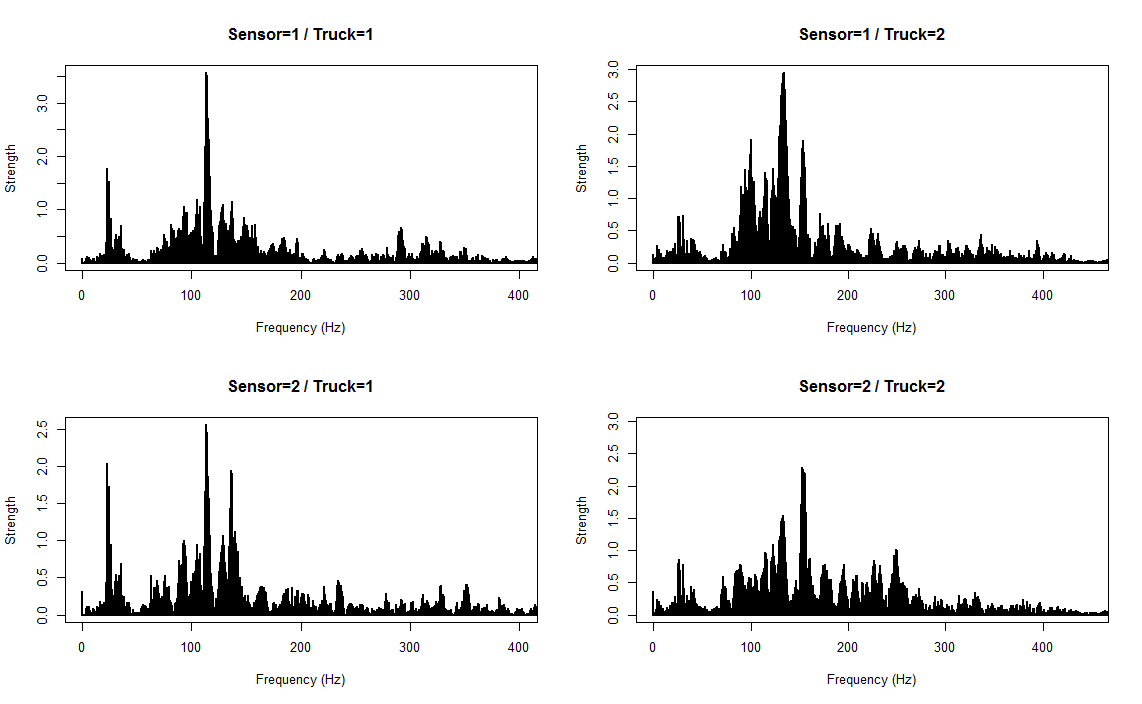
**3(b):**

**Feature 1:** FFT. This transforms vibration data into frequency ~ amplitude histograms. (code for this feature is provided in lines 274-294 of the script)



**Feature 2:**

Establish a relationship between 2 sensors for each truck. (code for this is provided in lines 300-325 of the script). Given a larger set of data, we may be able to develop a meaningful relationship with the 2 sensor data.



**3(c): Difficulties:**

The biggest difficulty in analyzing this data is we only have data for 2 events (i.e. 2 trucks). There are many heavy duty trucks on the road and they all carry varying type of cargo and travel at different speeds. In order to take into account all the variations in type of vehicle and type of load (including amount of load) we would need a much bigger sample size. Another challenge is to find data on a singular truck on a busy interstate. Also, we do not have any information about sensor data when there are more than one vehicle on the bridge. Another challenge would be to be able to identify a truck when it is not the only vehicle on the bridge. It might be achievable with enough data so that we can devise algorithms for subtracting vibrations induced by other vehicles.

In short, we need a lot more data and lot more information about the vehicles in order to accurately discern individual vehicles traveling on the bridge.

**Problem 4:**

**4(a)**

After looking at the current and previous competitions, I selected CDC zika virus epidemic dataset. The dataset includes approx. 107,000 lines of data. It tabulates worldwide cases of zika virus in 6 different categories in different regions (municipality/state or province). It can be accessed on Kaggle.com at <https://www.kaggle.com/cdc/zika-virus-epidemic>.

4(b): No. of rows: 107619

No. of variables: 9

Data Summary:

summary(zikaData)

report\_date location location\_type

Min. :2015-11-28 Length:107619 municipality:87557

1st Qu.:2016-02-20 Class :character state : 8462

Median :2016-04-09 Mode :character province : 7583

Mean :2016-04-04 county : 1613

3rd Qu.:2016-05-14 country : 866

Max. :2016-07-02 territory : 777

NA's :247 (Other) : 761

data\_field data\_field\_code time\_period

zika\_confirmed\_laboratory :28963 CO0001 :28963 Mode:logical

zika\_suspected :28963 CO0003 :28963 NA's:107619

zika\_suspected\_clinic :16170 CO0004 :16170

zika\_confirmed\_clinic :12793 CO0002 :12793

yearly\_reported\_travel\_cases: 1035 US0003 : 1035

weekly\_zika\_confirmed : 960 MX0001 : 960

(Other) :18735 (Other):18735

time\_period\_type value unit country

Mode:logical 0 :70205 : 7 Length:107619

NA's:107619 1 : 8423 cases :106519 Class :character

2 : 4325 municipalities: 1093 Mode :character

3 : 2518

4 : 1772

(Other):20238

NA's : 138