Quality Control for Data Channels

Pavan Kumar, Devarsh Dani

March 3rd 2017

1 Abstract

This project aims at plotting both physiological signals (Ex.Heart Rate, Breathing Rate, Palm EDA etc..) and driving performance signals (example: Speed, Acceleration etc.) Only the pp file of physiological signal does not require cleaning of data as it is pre-cleaned, whereas the other signals (BR,HR etc..) require pre processing and cleaning of outliers not in range and plotting of cleaned signals thereafter.

2 Introduction

The following tasks mentioned in the abstract cannot be hand picked and hand drawn as we have an enormous amount of dataset.This dataset includes thousands of signal values which have few outliers that may fall out of range. This data that has outliers in it along with in-range values is considered to be what is called Raw-Data. Our primary task is to collect the raw-data, analyze it for any ambiguities and if any, those outliers have to be marked 'suspect' and hence should be eliminated from the entire signal set and new signal set should be plotted thereafter and this set of plots obtained thus gives us the cleaned outlier-free signals. Obviously doing these manually would take forever because of the enormity of the dataset we have. Thereby we have to take the help of a convenient programming language to program our requirements in a specific format to ease task of manual calculations. R can come to the rescue as R programming language is easy and far more efficient to deal with when it comes to plotting of graphs with hundreds to thousands of values on both the axes. Also we have the liberty in R to use graphs of our choice that would best suit the need. For example Histograms would be great in comparison of different value ranges. Line plotting would be helpful in determining the growth or decline of a pattern in the dataset. Point plotting would be helpful in establishing and analyzing specific x,y value results at any given instant. This multimodal dataset was acquired in a controlled experiment on a driving simulator . The set includes data for n=68 volunteers that drove the same highway under four different conditions: No distraction, cognitive distraction, emotional distraction,

and sensorimotor distraction. The experiment closed with a special driving session, where all subjects experienced a startle stimulus in the form of unintended acceleration - half of them under a mixed distraction, and the other half in the absence of a distraction. During the experimental drives key response variables and several explanatory variables were continuously recorded. The response variables included speed, acceleration, brake force, steering, and lane position signals, while the explanatory variables included perinasal EDA, palm EDA, heart rate, breathing rate, and facial expression signals; biographical and psychometric covariates were also obtained. This dataset enables multidimensional research into driving behaviors under neatly abstracted distracting stressors, which account for an increasing number of car crashes. The set can also be used in physiological channel benchmarking and multispectral face recognition.

3 Data Significance

We are provided with 5 signal sets namely Heart Rate, Breathing rate, Palm EDA, pp and Res(performance) respectively. These signal sets are present as excel sheets in various sessions like PD,RD,MD,FD(FDN or FDL),CD,BL,ND etc.. which are present in T001 to T088 folders. Our preliminary task is to understand the data in the .xlsx sheet that has all Code Sessions and Signal sets. We have values -1,0,1, NA in the excel sheet that give us a fair idea as which signal set in what code session has outliers and in which files are the signal sets missing and in which files all the signals are in range and other information. So, essentially the .xlsx sheet serves as the final output cross verification sheet to cross check the values obtained by plotting graphs with both uncleaned data and cleaned data respectively. In the present repository, an Excel spreadsheet named 'Dataset-Table-Index.xlsx' gives an exhaustive enumeration of the dataset's files. Files expected to be present and are present indeed, are denoted by 1; there are 4,960 such files. Files expected to be present, which are not present for technical and other reasons, are denoted by 0; there are 236 such cases. Files that are not supposed to be present due to the experimental design are denoted by NA; there are 544 such cases. Files that have been redacted due to IRB restrictions, are denoted by IRB; there are 40 such cases. Files associated with derivative thermal variables, such as perinasal EDA, which could not be extracted due to the presence of facial hair, are denoted by N; there are 144 such cases. Files that are present, but found during the technical validation process to be marred by noise, are denoted by -1; there are 60 such files and we do not recommend using their data. To identify these signals, we considered [4, 70] bpm as the legitimate range for the breathing rate variable, [40, 120] bpm for the heart rate variable, and [10, 4,700] k for the palm EDA variable. Signals that had at least one value outside the corresponding legitimate ranges, were the signals marked by -1 in the 'Dataset-Table-Index.xlsx'.

In the signals of certain performance variables we recommend noise cleaning interventions, as follows: For speed signals, we suggest replacing values -0.1 ¡ X ¡ +0.1 kph with X = 0 kph, while substituting values X ¡ -0.1 kph with

missing values. For accelerometer signals, we suggest replacing negative values with missing ones. For brake force signals, we suggest replacing values Y ¿ 300 N with Y = 300 N. The repository's data is organized per subject under two major directories: (a) Raw Thermal Data - 1.54 TB in size (b) Other Study Data - 57.5 GB in size. In these directories, the subject folders are named Txxx, where xxx stands for the subject number. In the Raw Thermal Data directory, each subject's folder contains the facial thermal sequences for all experimental sessions.

4 Approach

Files Breathing rate, Heart Rate, Acceleration , Palm Eda (NR Perinasal) Speed, Braking of Res files require data pre processing. For BR,HR,Peda we need to plot signals of both uncleaned and cleaned side by side.

4.1 Heart Rate

Heart Rate has been measured by a sensor that operates close to the heart. So the outliers recorded can be attributed to sensor not properly recording the values which can thus be attributed to sensor not touching the body properly to record those values and hence those errors. Hence to avoid these outliers we try to eliminate the entire signal set that contains outliers to obtain cleaned data free of outliers and error prone values and rates. Essentially, the heart rate signals in range [40,140] beats per minute have been included in the cleaned section. The R script has been designed in such a way that any deviation from these values, the entire heart rate signal has been dropped and this can be observed in the graphs obtained with the help of 'n' count mentioned on each and every cleaned and uncleaned graphs plotted.

4.2 Breathing Rate

Similarly for Breathing Rate signals we have developed somewhat similar R script as Heart rate that checks for particular range of values and if any outlier in any signal set, would drop the entire signal set and hence would return us the cleaned signal sets that have all the values in range. The range for this breathing rate is [4,70]..Any deviation from the given range, the entire signal has been dropped of the list.

4.3 Perinasal Perspiration PP

pp signals have time , frame, NR perinasal, perinasal columns in it. We are interested in capturing time and NR perinasal values to plot time on X axis and NR perinasal on Y axis. pp signals do not require any preprocessing or cleaning and hence only a set of signals (cleaned) are plotted for all the sessions that contain pp signal set.

4.4 Palm EDA peda

This Excel file contains three synced columns: Frame,Time, Palm EDA signal. Hence, scanning each row from left to right we find the chronological rank order of the instantaneous measurement, the time [in s] the measurement was taken with respect to the beginning of the session, and the value of the palm EDA signal at that time [in k]. No palm EDA measurements were recorded during the Baseline (BL) sessions. The Palm EDA sensor values have been recorded using Shimer 3 GSR sensor which is powered with Lithium Rechargeable ion battery and it's measurement range being from 10 kilo ohm to 4,700 kilo Ohm. So, any value outside this range is considered as an outlier and hence that particular entire signal set has been omitted.

4.5 Performance

The performance measurements included in the res file are acceleration, speed, Lane offset, Lane Position, Braking, Steering. Lane Position is of no interest to us in this task set.

4.5.1 Braking

In the cleaning of Braking , any value greater than 300 has to be reduced to 300 and then plotted. Practical significance of this is not known. Hence any signal value greater than 300 is to be reduced to 300.

4.5.2 Acceleration

In the acceleration signal set, the physical acceleration pedal is connected to a simulated throttle valve that can move from [0,90) which implies 0 included and 90 not included. If acceleration value in any of the signal set is greater than 90 degrees, we omit the entire signal set and if the acceleration value is less than 0 we replace those values less than 0 with 'NA' and also to the corresponding time value to avoid difference in lengths while plotting the graph. All the values in range [0,90) are considered clean and thus appear in all the graph sets plotted for acceleration.

4.5.3 Steering

In the res dataset, steering values are in the 6th column, hence we try to extract 2nd column ie time and 6th ie steering on X and Y axis respectively. No limits or ranges have been specified for steering as steering is considered a physical attribute and it depends solely on driver which differs from one driver to another.

4.5.4 Speed

In the res dataset , the speed is in the 3rd column and time in 2nd column. Hence we try to extract 2nd column ie time and 3rd ie steering on X and Y

axis respectively. Following cleaning has to be done on speed dataset to obtain an error free speed-time graph. To clean all the speed signals if -0.1¡Speed¡0.1 and to replace all such values with speed =0 and if there is any value such that speed¡-0.1, then those values have to be replaces with 'NA' or missing values. Doing this cleaning would return us an error free speed time graph.

4.5.5 LanePosition

In the res dataset , the lane position is in the 8th column and time in 2nd column. Hence we try to extract 2nd column ie time and 8th ie steering on X and Y axis respectively. There was no range specified on Lane position vs time graph. Hence Lane Position graph has been plotted along wih time without any cleaning or pre processing.

5 Conclusion

This project gave us many insights into usage of R programming to prune data and perform cleaning of data on an initially given unclean data which contains both psychological and driver physical data such as steering, braking, acceleration, speed etc.. The understanding of data set has played a key role in the development of the project.

6 Drawbacks

There were certain drawbacks encountered during the development of the project. In the braking graphs of all sessions, there is one/two signals crossing the given limit of 300.Also, we have missed out on plotting Lane position of MD session which is sorted out and the error was in reading the excel sheet of all the signals with columns. Had it been with names/headers of the columns, we could have avoided the problem beforehand.

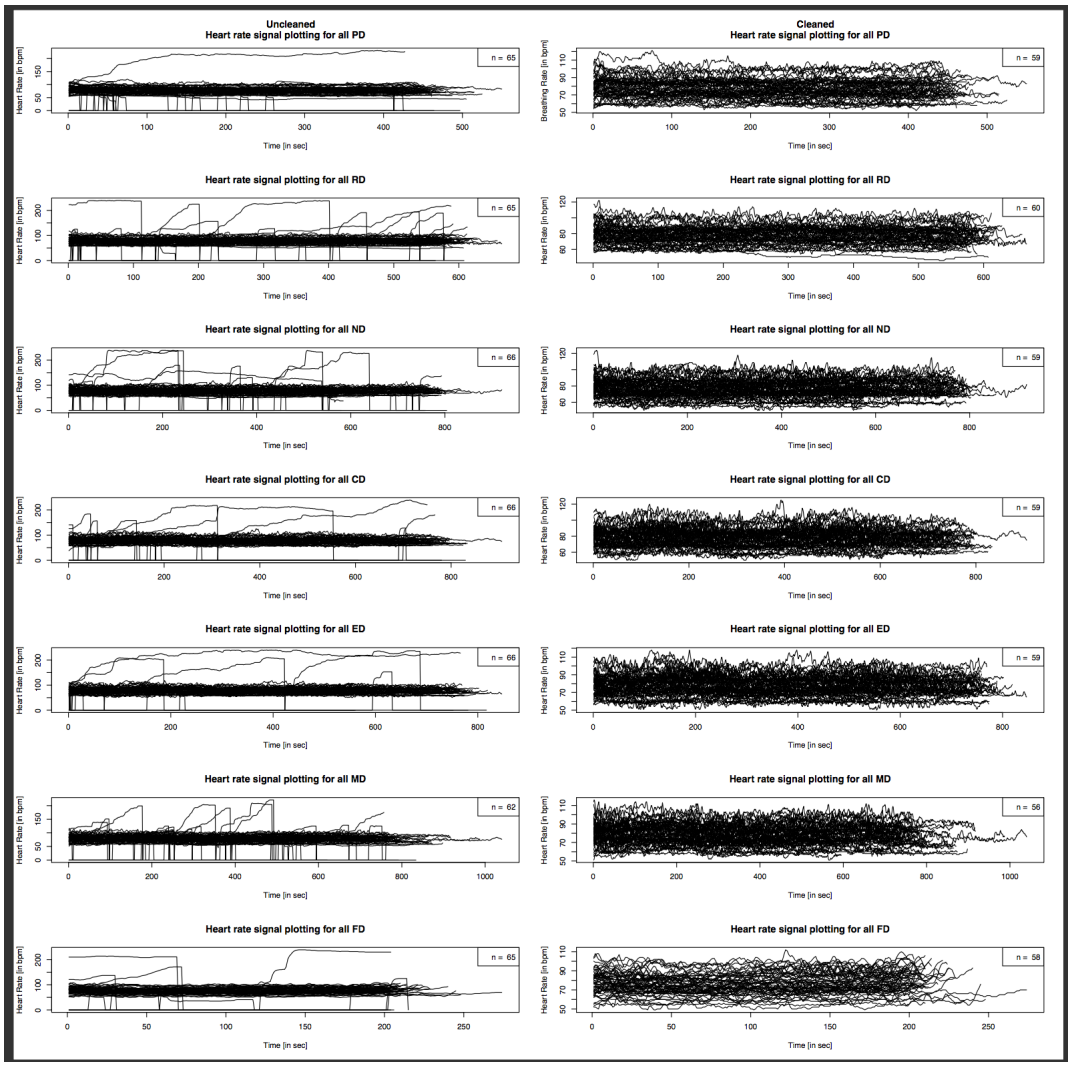

Figure 1: HR signals uncleaned and cleaned

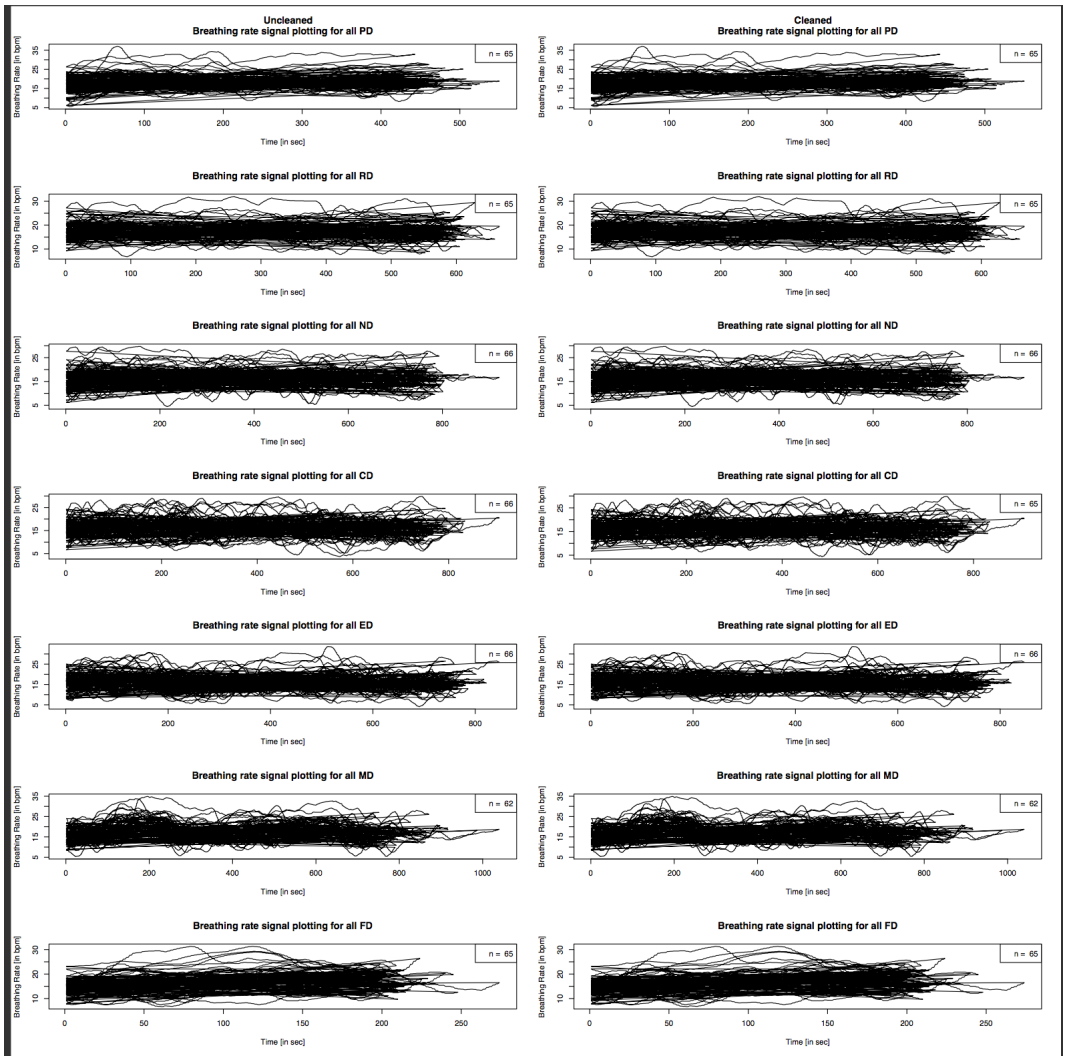

Figure 2: BR signals uncleaned and cleaned

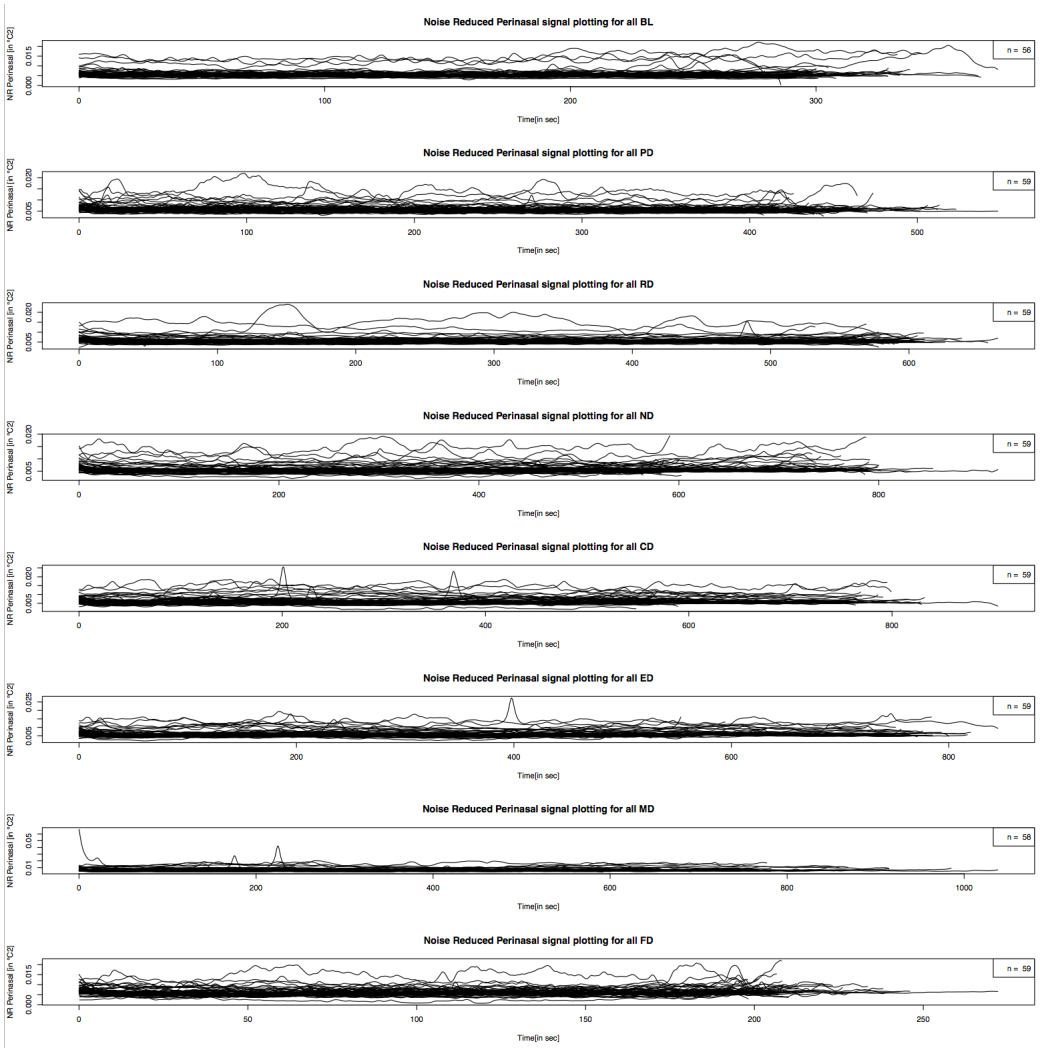

Figure 3: PP signals cleaned

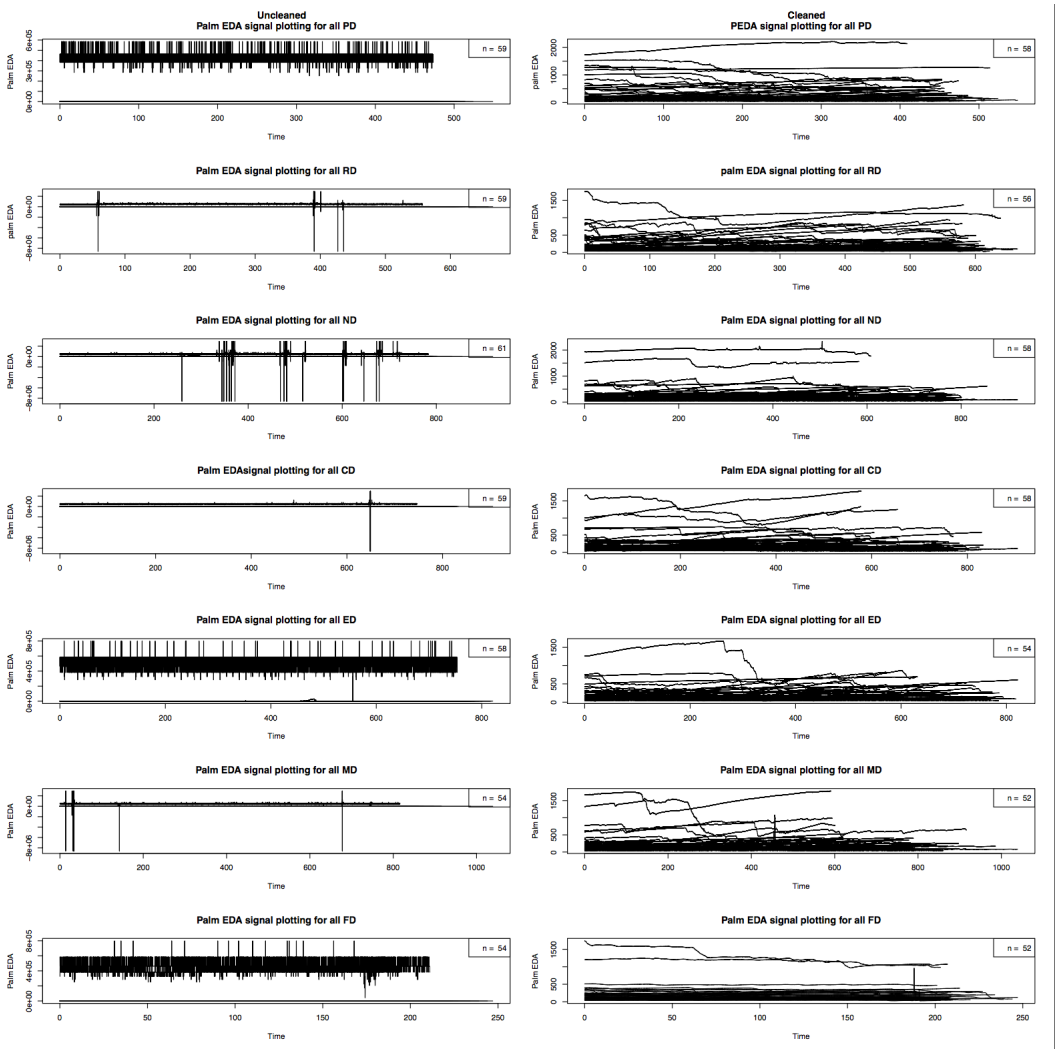

Figure 4: Palm EDA signals Uncleaned and cleaned

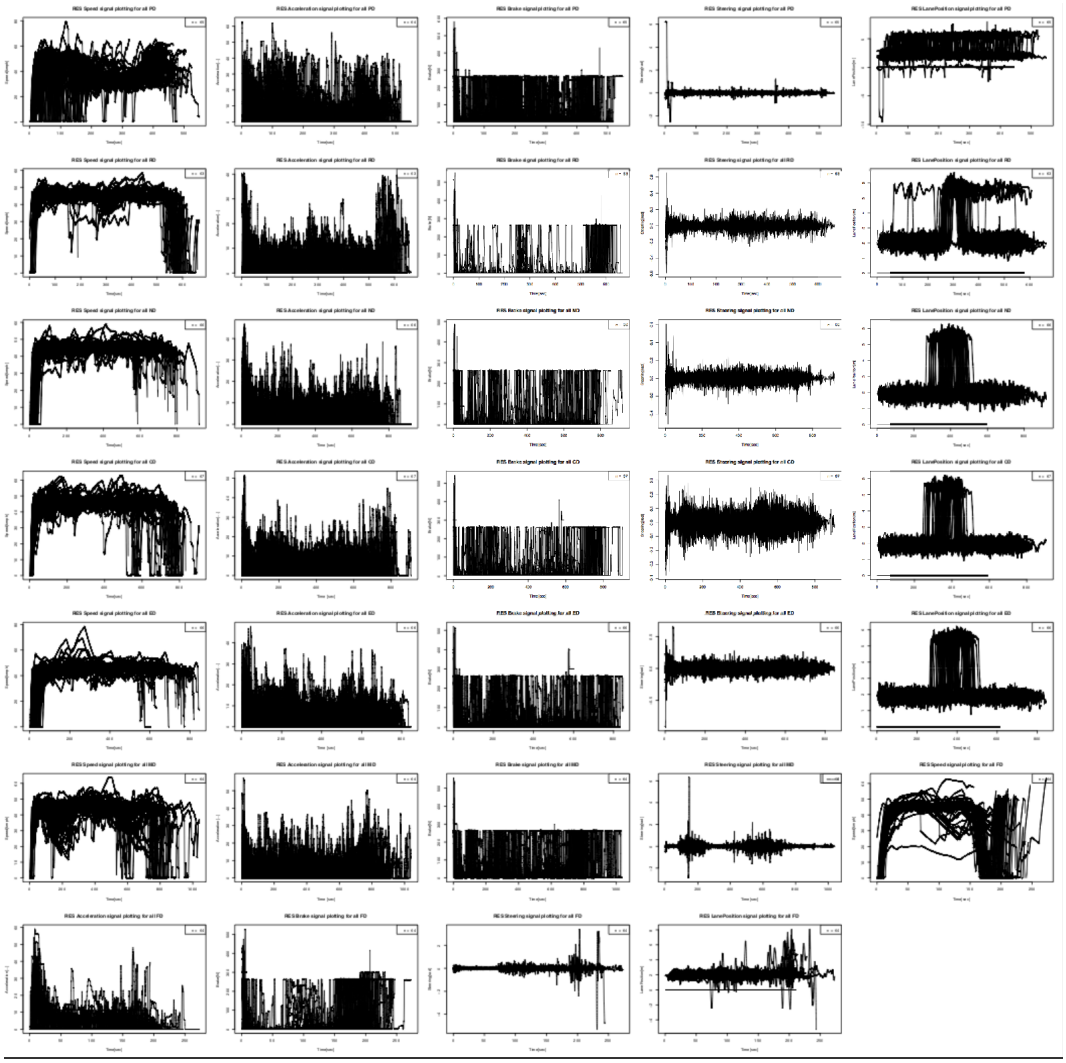

Figure 5: Res signals Uncleaned and cleaned