New Vectors for Machine Data

Nikhil Muthukrishnan

25 October 2018

List of Vector for Machine Data

Objective:This document seeks to explore the various vectors that can created from PI tag data for use in predictive modelling.
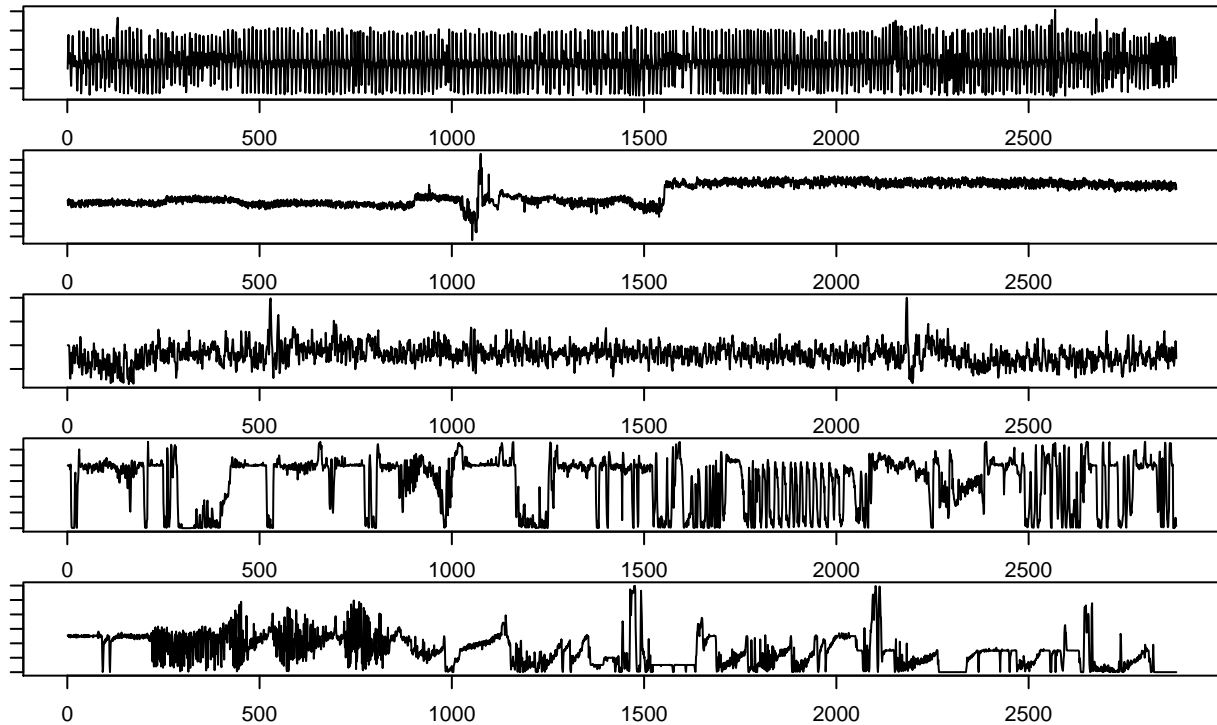
Data:The data used can be found in drive xxx. The data has been cleaned of white spaces and text with only selected coloumns viz. .PV sensors pertaining to FIC.

Sprint:PO_SPRINT_12_OCT_2018.The machine data has been loaded into pipeline and formatted for staging table ready for streaming visualisation.

However the large number of features makes for difficult computation and hence needs to be selected after which appropriate transformation will also be done for final modelling.

Basic plot of the cleaned machine data

To begin please observe the plots provided for each of the FIC PV sensor from
01-01-2018 00:00 to 30-06-2018 23:55.

Several statistical transformations will be applied to the PI tags to explore new vectors for machine data. As can be seen from the plots above the chosen PI tags are diverse and somewhat representative of the entire line.

Vectors based on Central Limits

```
##                        mean      median          SD         IQR          Min
## SC3_FIC20234.pv   26.630701   26.769920    4.184984    3.372448 17.966856000
## SC3_FIC20267.pv   16.765898   16.292307    1.641093    3.010804  9.416485786
## SC3_FIC20295.pv    8.054341    8.008667    2.030057    2.487564  1.799371362
## SC3_FIC20443.pv   29.270306   37.796867   16.122501   26.411901  0.001678518
## SC3_FIC22250.pv  146.931938  144.767601  106.789182  181.856747  0.018311106
##                         Max
## SC3_FIC20234.pv    40.39430
## SC3_FIC20267.pv    22.95480
## SC3_FIC20295.pv    19.99817
## SC3_FIC20443.pv    54.99832
## SC3_FIC22250.pv   598.59003
```
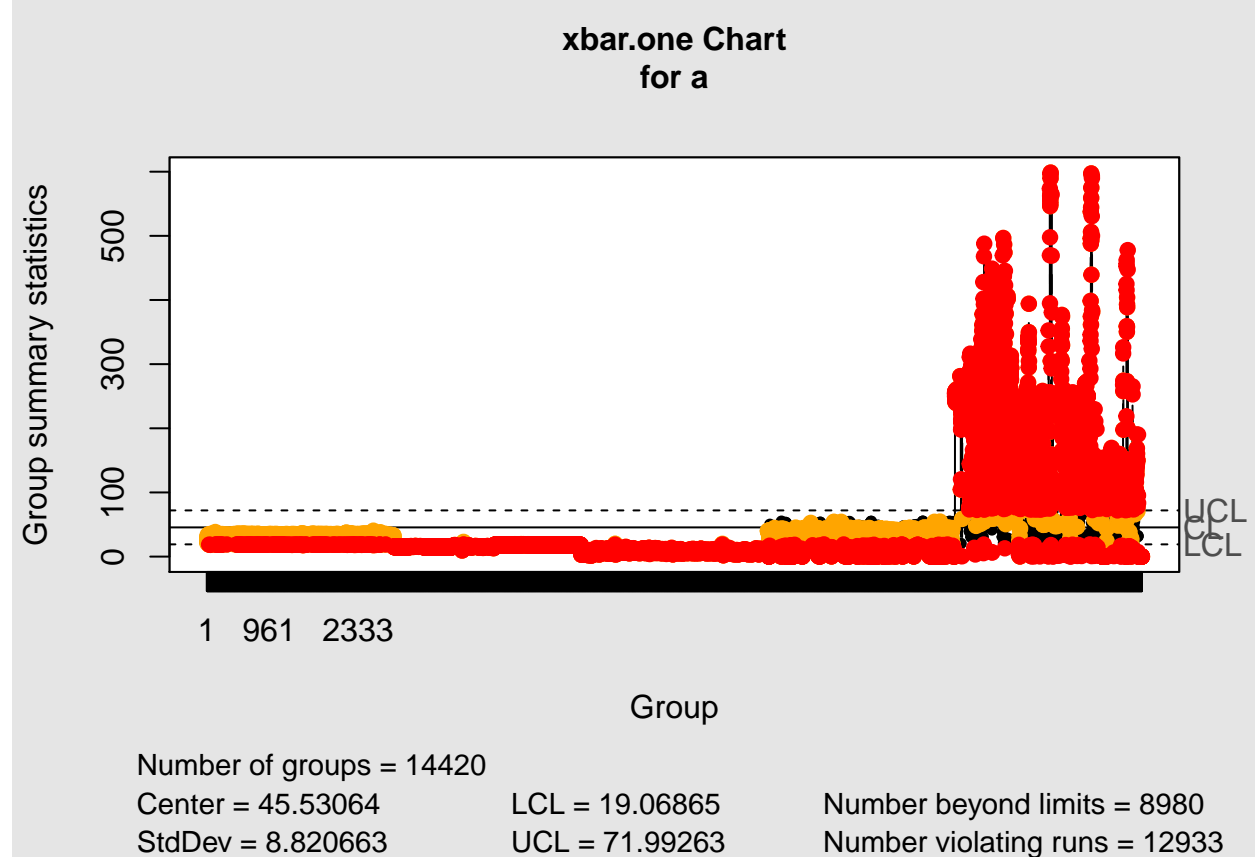
The above table contains 5 select PI tags treated with 4 statistical methods namely range,mean,median,standard deviation and variance.
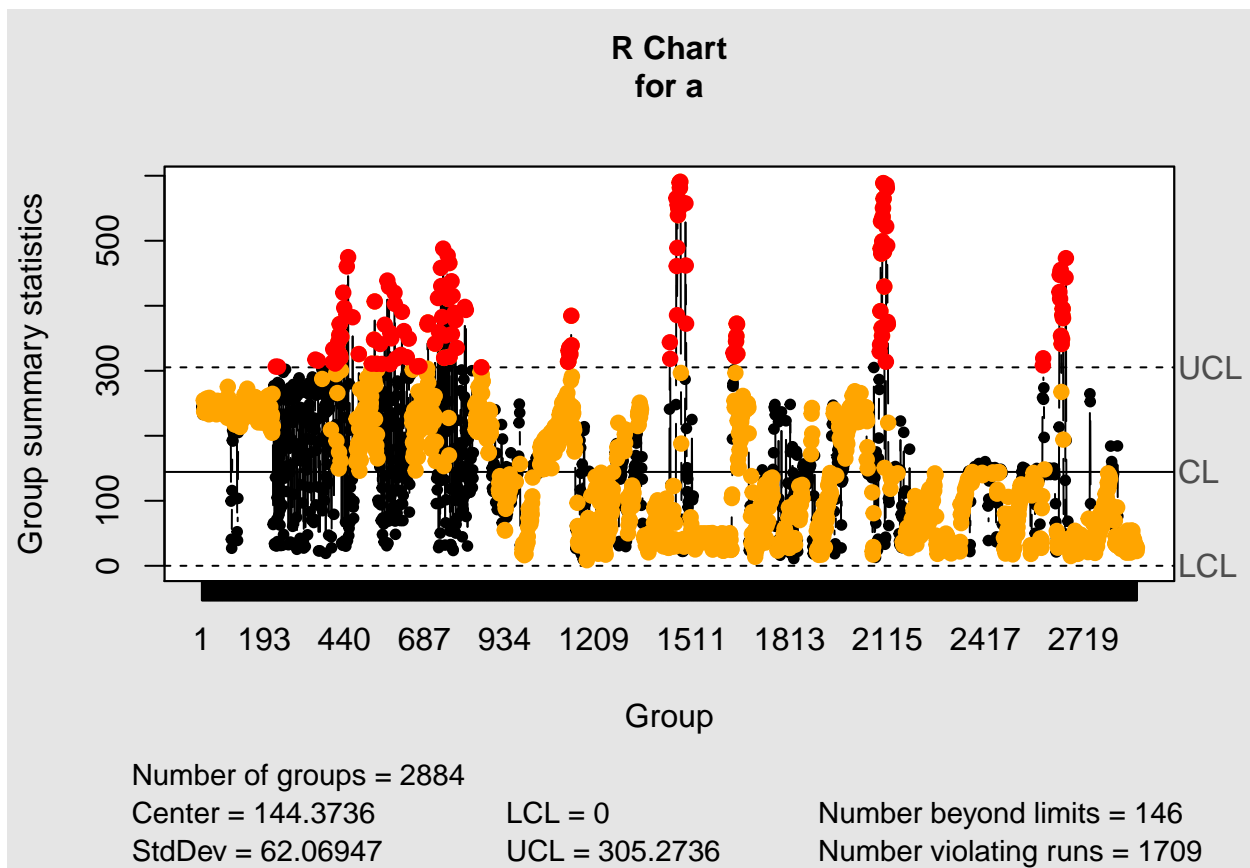
As can be seen mean and median are almost the same for 4/5 PI tags but fail to describe the volatility of the time series data. Hence SD and variance can be used. However some sensors may generally have high variance and may get unnecessarily hightlighted leading to the popular QC charts.
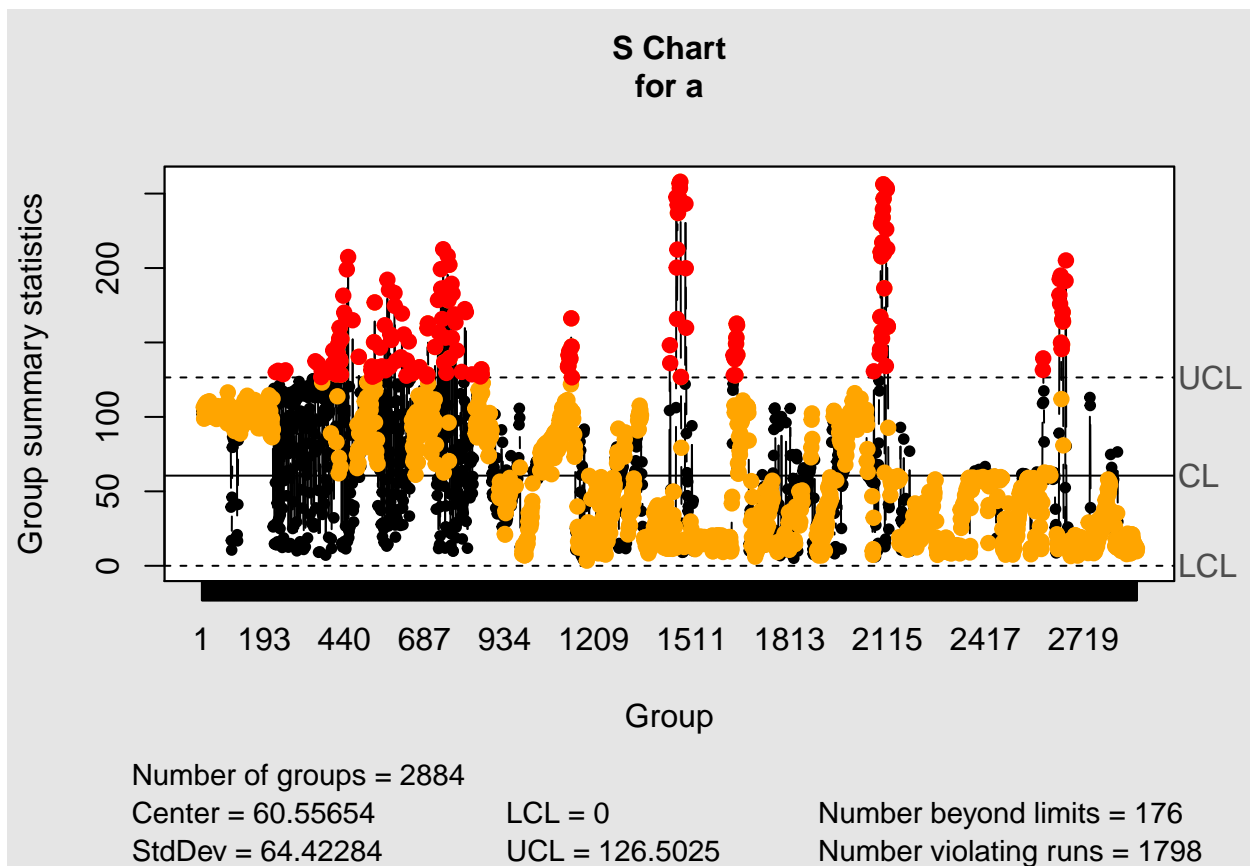
Vectors Based on Quality Control Charts

Normally Indsutrial Statistics makes use of control charts. Below are 3 control charts of process variables.

"xbar" is the means of a continuous process variable

R Chart
for a

Number of groups = 2884
Center = 144.3736   LCL = 0   Number beyond limits = 146
StdDev = 62.06947   UCL = 305.2736   Number violating runs = 1709

"R" is the ranges of a continuous process variable

S Chart
for a

Number of groups = 2884
Center = 60.55654   LCL = 0   Number beyond limits = 176
StdDev = 64.42284   UCL = 126.5025   Number violating runs = 1798

"S" is the standard deviations of a continuous variable

Upon the central limits of the sensor data further insights can be garnered by introducing Upper and Lower Control Limits as can be seen above.