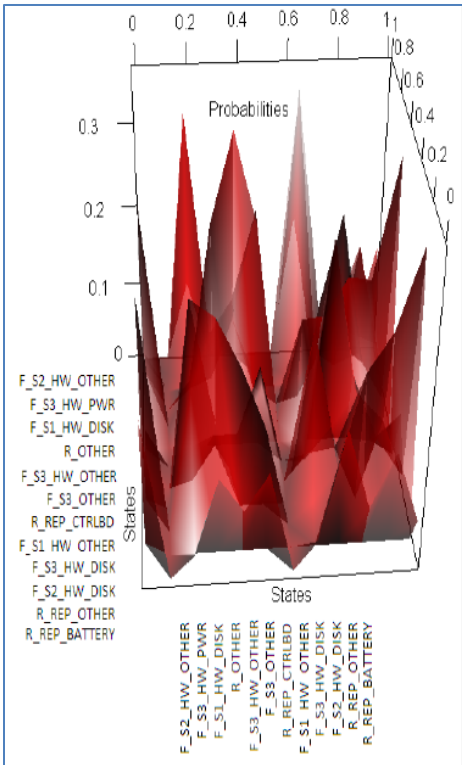
Problem 1: Device Failure Prediction

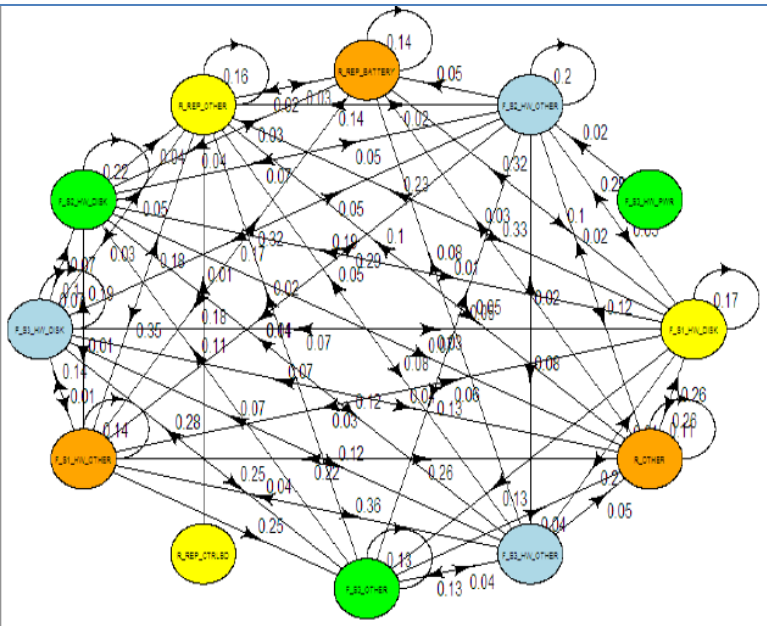
- Input: Event logs from devices
- Predict when / what / how / why the next failure is going to take place, predict probability of failure for a device, suggest fix for a failure, find frequent patterns.
- Event Sequence Clustering / Frequent Subsequence Mining / Markov Chain / Correspondence Analysis / n-gram based Classification techniques.
- Time-to-failure Analysis / Fit distribution (Gamma / Weibul / Exponential) / Reliability / Survival Analysis.

Some Results



Input
1. E_t: Temporal (time-ordered) event sequence.
2. k: Order of Markov Chain.

Output
T: Stochastic Markov Transition Probability matrix $T = [p_{ij}]$ (normalized).

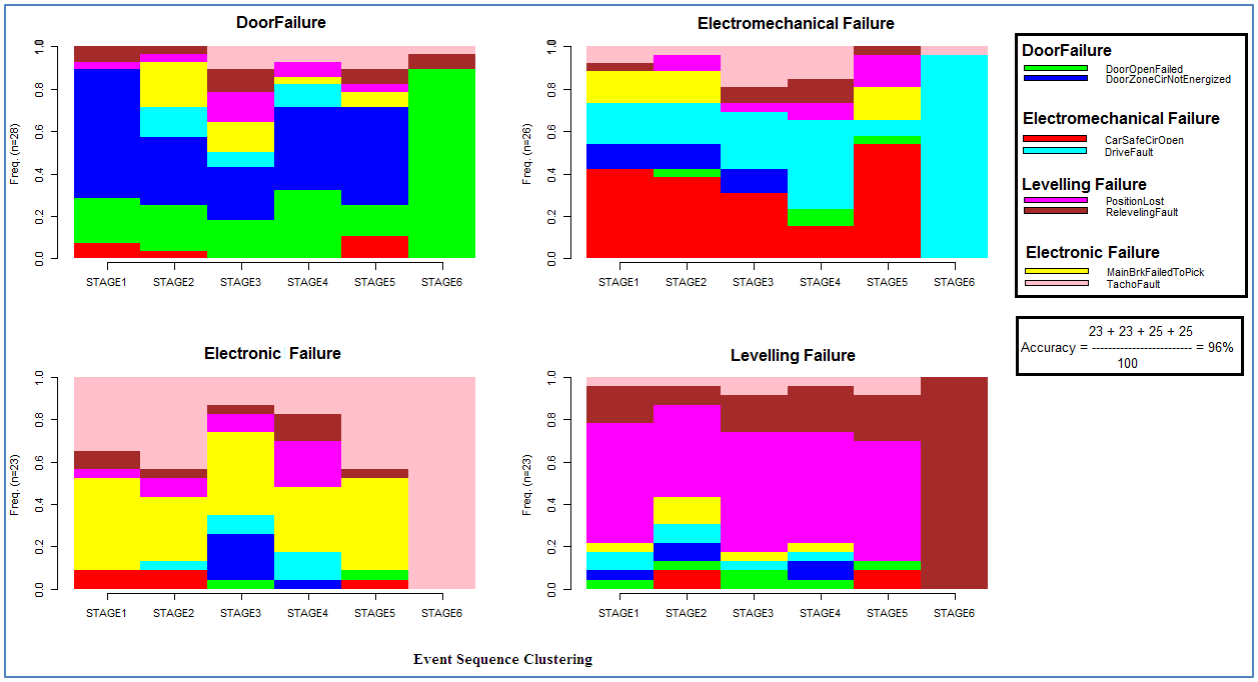


State Digram for Markov Process of Order 1

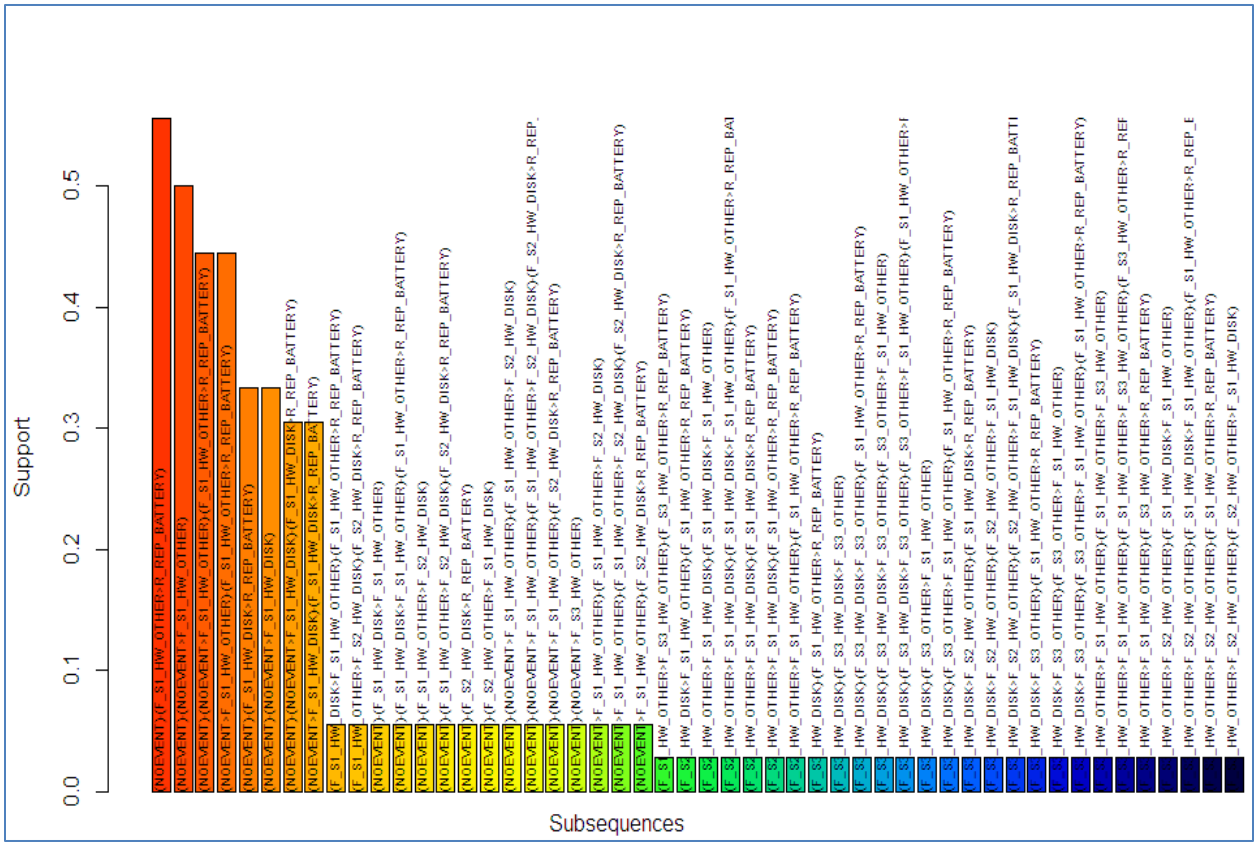
Algorithm ComputeTran: Compute the Markov Transition Matrix for Markov Chain of order k

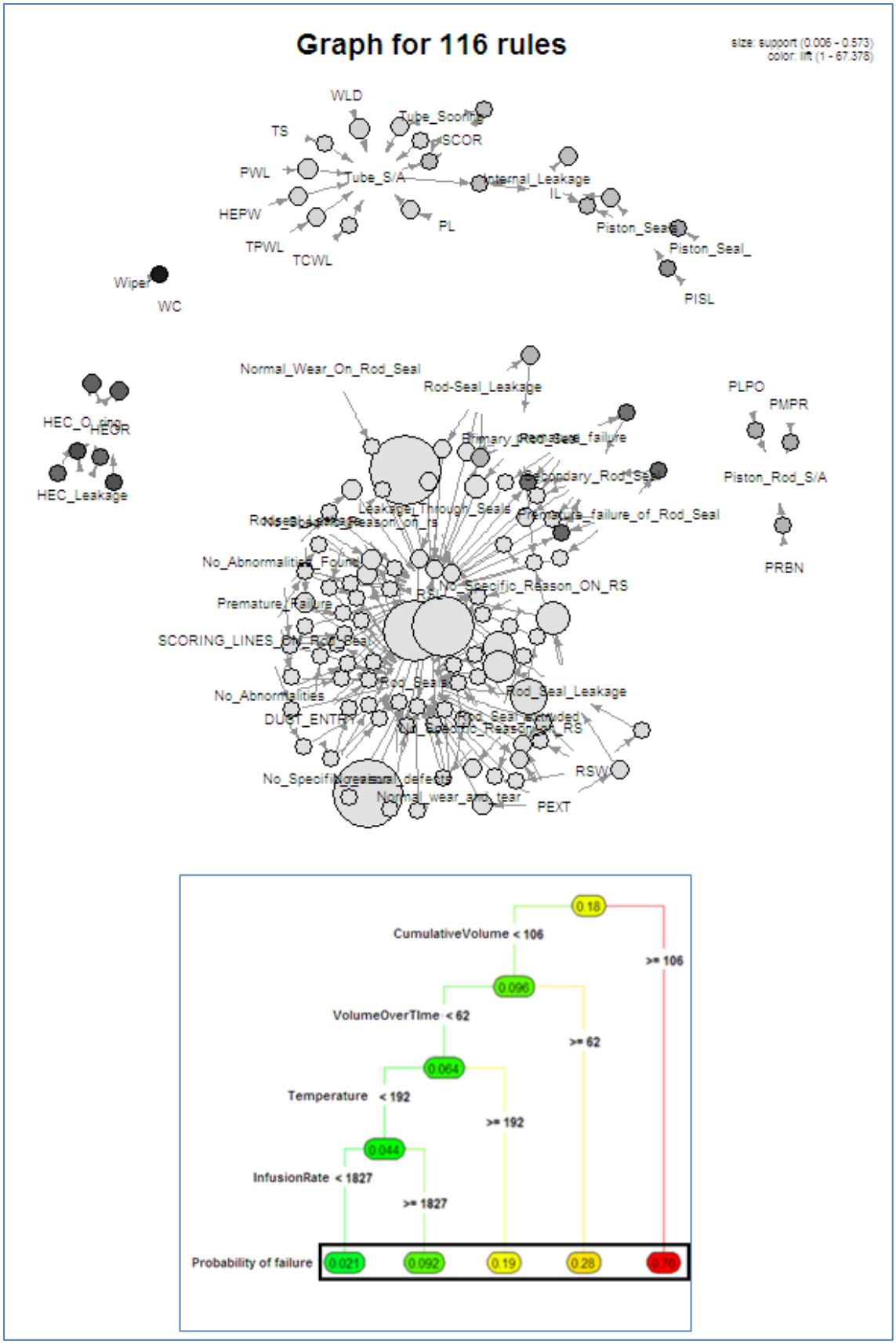
1: For $k = 1$, Compute the stochastic transition probability matrix in the following manner:
$p_{ij} = P(E_i = e_i | E_j = e_j) = \#(E_i = e_i | E_j = e_j)/Z$,
where Z is the partition (normalization) function, with $Z = \sum_e \#(E_i = e_i | E_j = e)$,
here $\#$ represents the emperical count (frequency). If $Z = 0$, define the conditional probability to be 0. Similarly, compute for other 'k's.

2: Sort the matrix with decreasing probabilities and use this transition matrix to predict the next possible event given the current ones.



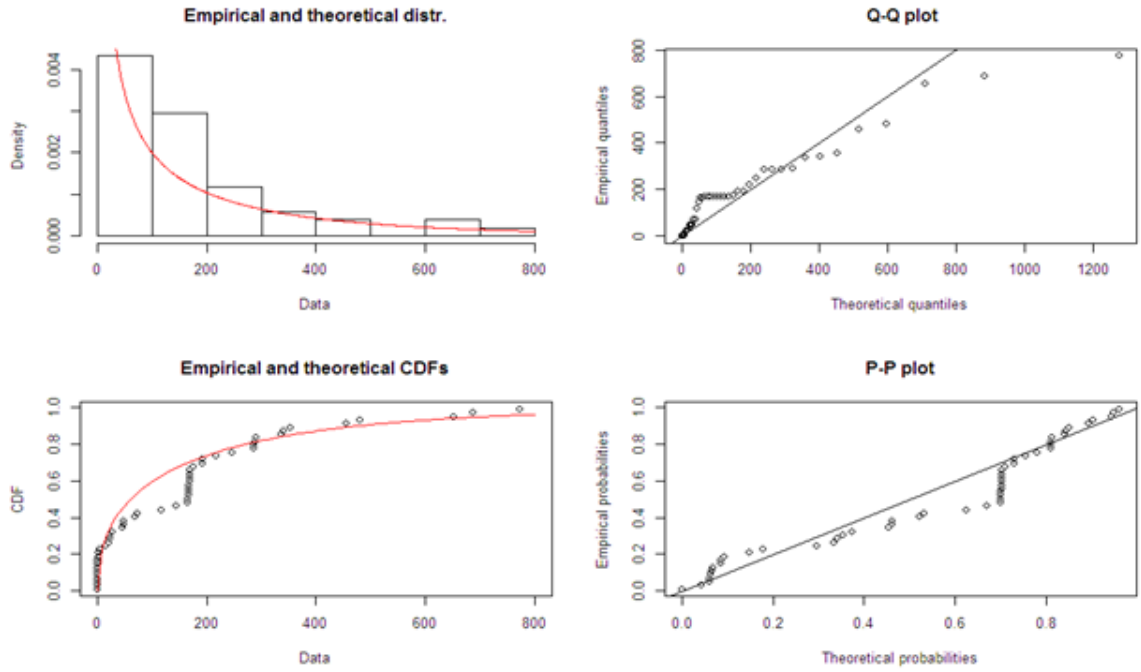
Event Sequence Clustering




Graph for 116 rules
size: support (0.006 - 0.573)
color: lift (1 - 67.378)
WLD
Tube_Scoring
TS
SCOR
PWL
Tube_S/A
Internal_Leakage
IL
HEPW
PL
Piston_Seal
Piston_Seal_
TPWL
TCWL
Wiper
WC
PISL
Normal_Wear_On_Rod_Seal
Rod-Seal_Leakage
PLPO
PMPR
HEC_O_ring
HEGR
Piston_Rod_S/A
HEC_Leakage
Leakage_Through_Seals
Premature_failure_of_Rod_Seal
No_Abnormalities_Found
NO_Specific_Reason_ON_RS
PRBN
Premature_Failure
SCORING_LINES_ON_Rod_Seal
Rod_Seals
No_Abnormalities
Rod_Seal_Leakage
DUST_ENTRY
RSW
PEXT
Normal_wear_and_tear
0.18
CumulativeVolume < 106
>= 106
0.096
VolumeOverTIme < 62
>= 62
0.064
Temperature < 192
>= 192
0.044
InfusionRate < 1827
>= 1827
Probability of failure
0.021
0.092
0.19
0.28
0.76

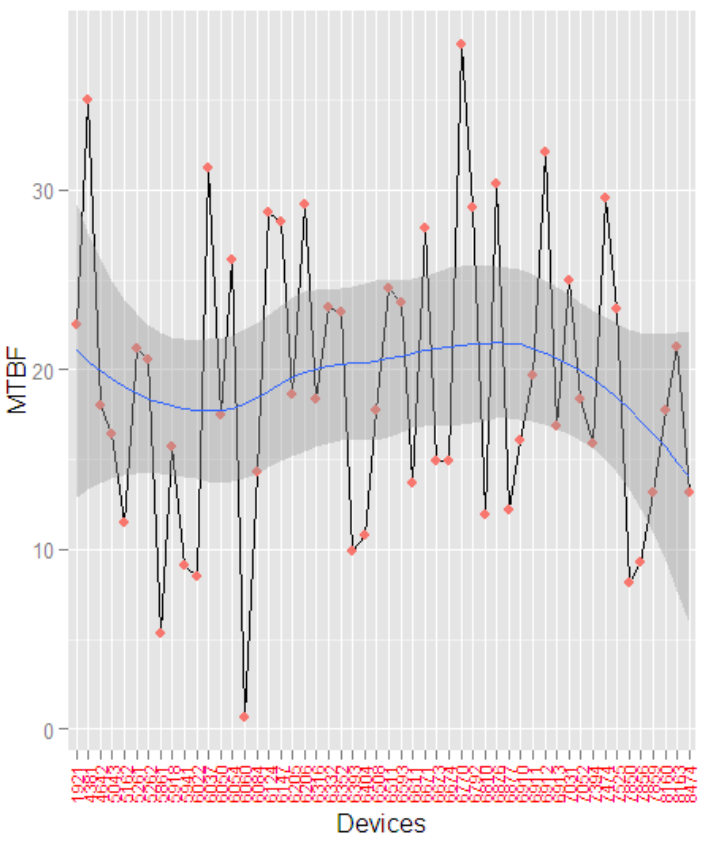
Distribution to fit	Method	Estimated Parameters	Estimated mean time between failures	KS Test Output D (dist. ECDF)	KS Test Output p-value
exponential	MLE	rate (λ) 0.005937159	168.4307 $\left(\frac{1}{\lambda}\right)$	0.1765	0.4051
gamma	MLE	shape (α) 0.396826781 rate (β) 0.002355428	168.4733 $\left(\frac{\alpha}{\beta}\right)$	0.2353	0.1188
weibull	MLE	shape (k) 0.5296621 scale (λ) 115.9829857	209.715 $\lambda\Gamma(1+1/k)$	0.2745	0.04285
exponential	MME	rate (λ) 0.005937159	168.4307 $\left(\frac{1}{\lambda}\right)$	0.1961	0.2807
gamma	MME	shape (α) 0.841830023 rate (β) 0.004998079	168.4307 $\left(\frac{\alpha}{\beta}\right)$	0.2157	0.1863

Failure Probability

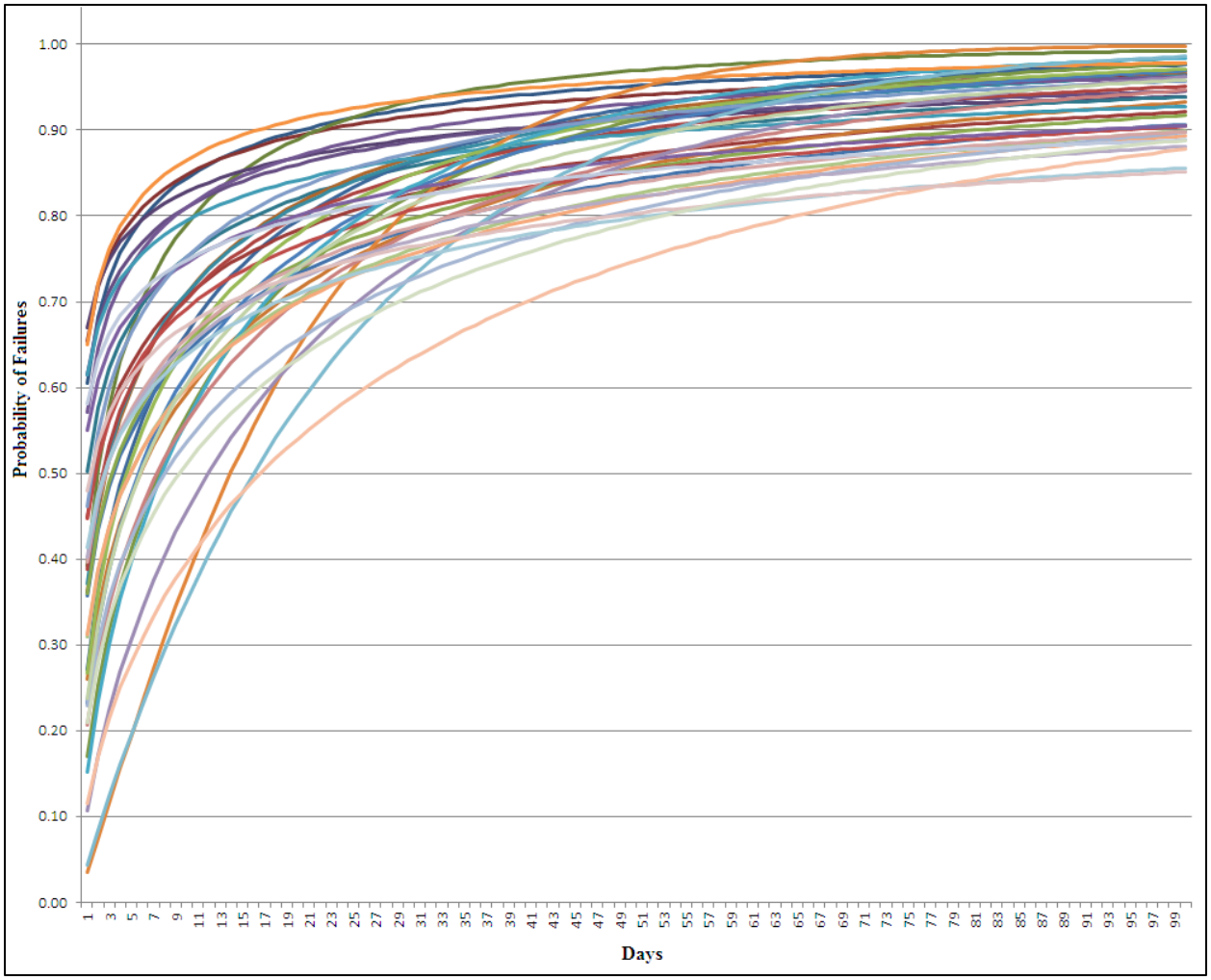


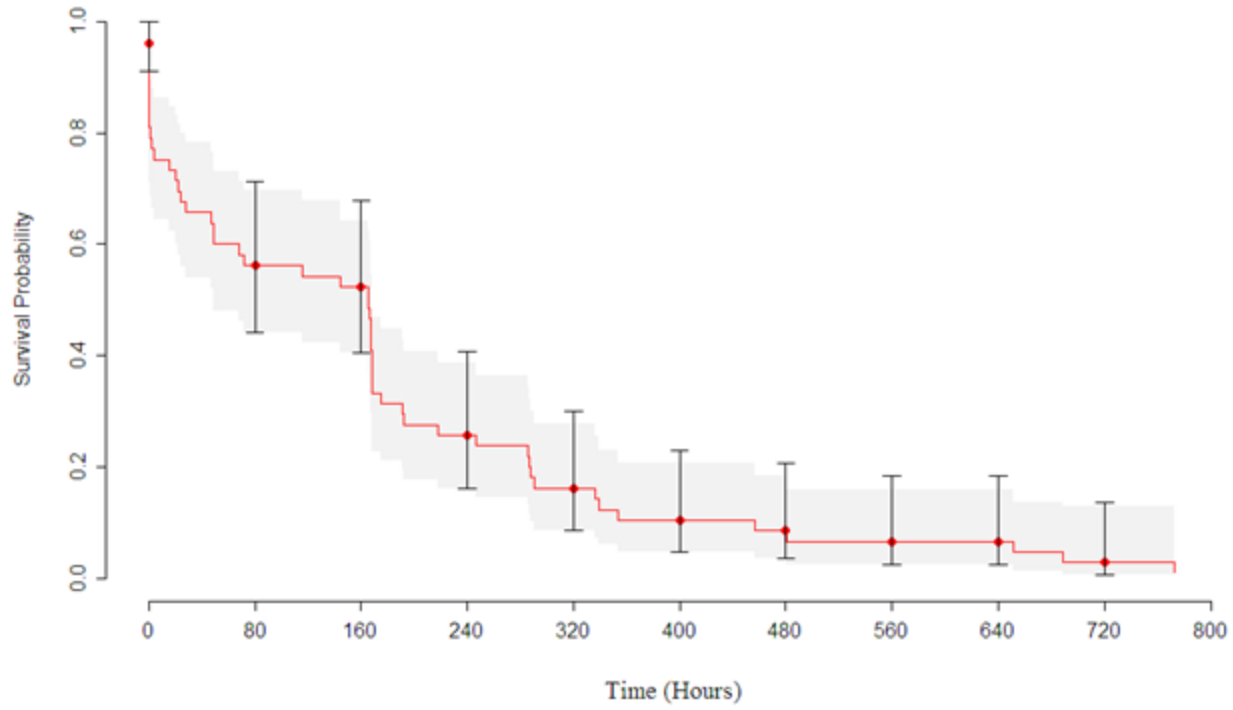
Fitting Weibull Distribution

Devices	Laplace Scores	Device	MTBF
6404	3.6701707	6060	0.693503997
6671	3.5307887	5861	5.353747745
6674	2.8822418	7850	8.167718333
6084	1.7911798	6022	8.529457751
5941	1.1972018	5941	9.099262137
1921	0.9905926	7856	9.258320932
5043	0.9578725	6393	9.867834074
6147	0.9573976	6404	10.76585967
8163	0.8022508	5162	11.49651538
7856	0.3466974	6810	11.96021531
6876	0.3103483	6877	12.20824413
7850	0.298096	7859	13.13386417
7394	0.1149123	8474	13.14538037
4642	0.1075487	6611	13.71489889
6810	-0.3862886	6084	14.33227273
7859	-0.4084318	6674	14.92131682
8474	-0.4443324	6673	14.9343354
8160	-0.4684823	5918	15.70993063
4381	-0.5091515	7394	15.94167667
6054	-0.5765511	6910	16.04151327
5262	-0.6105039	5043	16.44730532
6511	-0.6277981	6913	16.89979632
7474	-0.6342535	6050	17.48257173
6911	-0.6807914	8160	17.70690484
6205	-0.7091824	6508	17.78698917
6912	-0.7932293	4642	18.05237111
6913	-0.8417798	6316	18.33616741
7052	-0.9037764	7052	18.40514889
6910	-0.9475168	6205	18.62501
6792	-0.9847287	6911	19.68695444
6037	-1.0158466	5262	20.5234909
6673	-1.1727057	5261	21.15348222
6770	-1.2362374	8163	21.25896833
6393	-1.2653536	1921	22.51190815
6352	-1.3211369	6352	23.17865583
6593	-1.3745835	7525	23.3890973
6508	-1.3857121	6332	23.43403778
6050	-1.388042	6593	23.74011417
7031	-1.3912138	6511	24.56424611
6316	-1.4838644	7031	24.9403412
5162	-1.6193257	6054	26.1321147
5918	-1.6284914	6671	27.91488269
6332	-1.7512199	6147	28.20047296
6206	-2.1949578	6124	28.79091278
5861	-2.2668381	6792	28.99711648
6877	-2.3120153	6206	29.15874843
5261	-3.2753816	7474	29.52724765
6124	-3.3474478	6876	30.33507944
6611	-3.5522309	6037	31.23283828
7525	-3.7679667	6912	32.10160722
6022	-4.4017819	4381	35.0033279
6060	-14.6818115	6770	38.10001

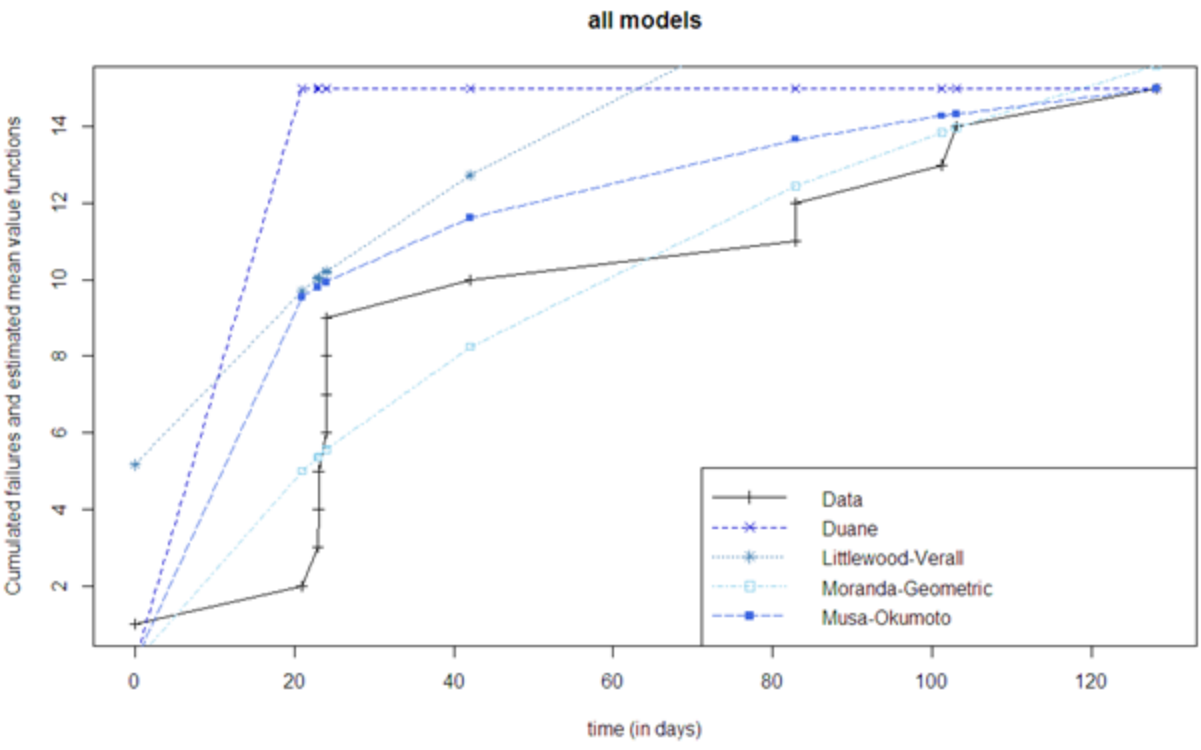


Mean MTBF	19.28642	Days
SD	8.049306	
95% CI	17.04548	21.52736
90% CI	17.4164	21.15644

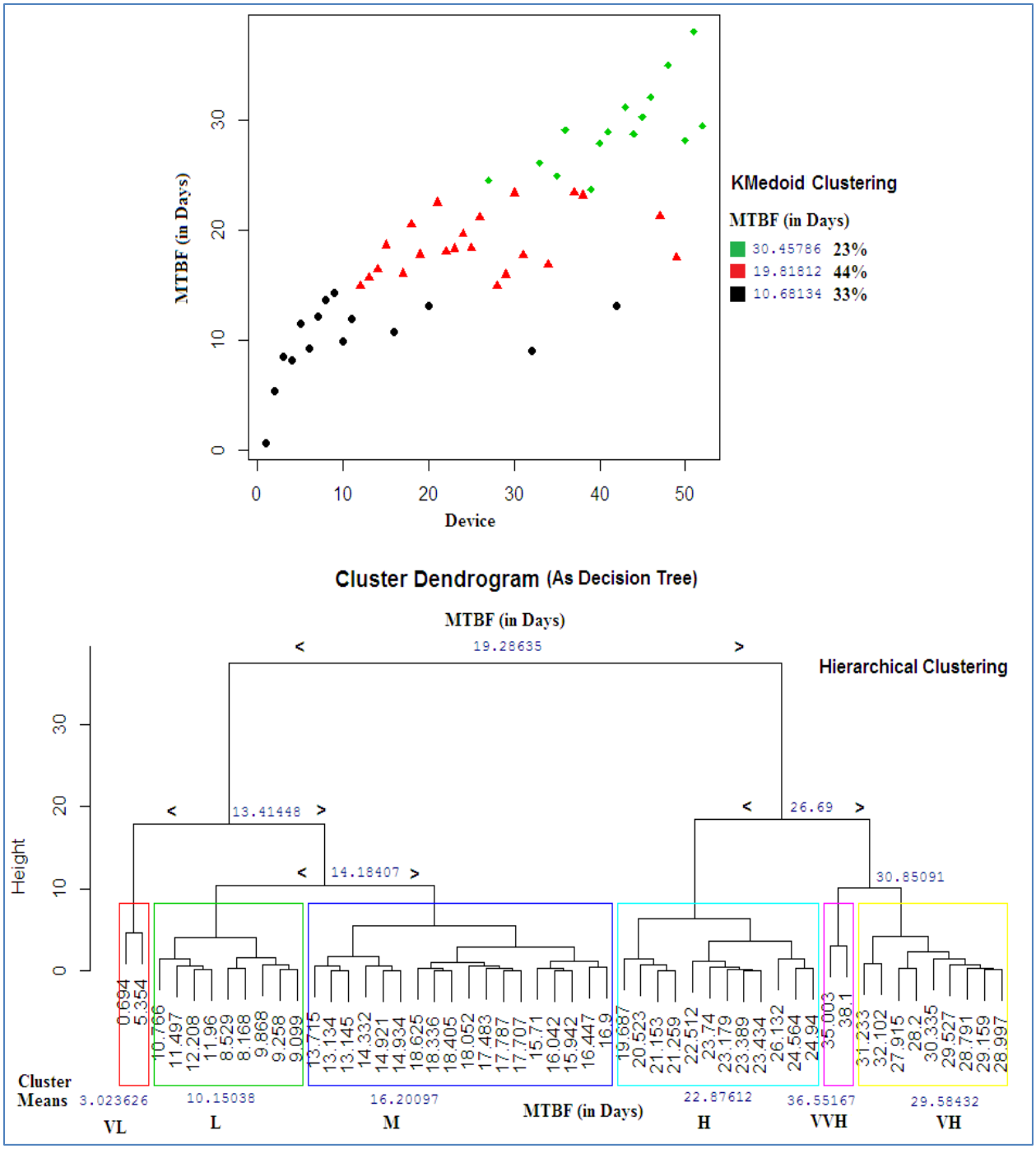

Probability of Failures
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
61
63
65
67
69
71
73
75
77
79
81
83
85
87
89
91
93
95
97
99
Days

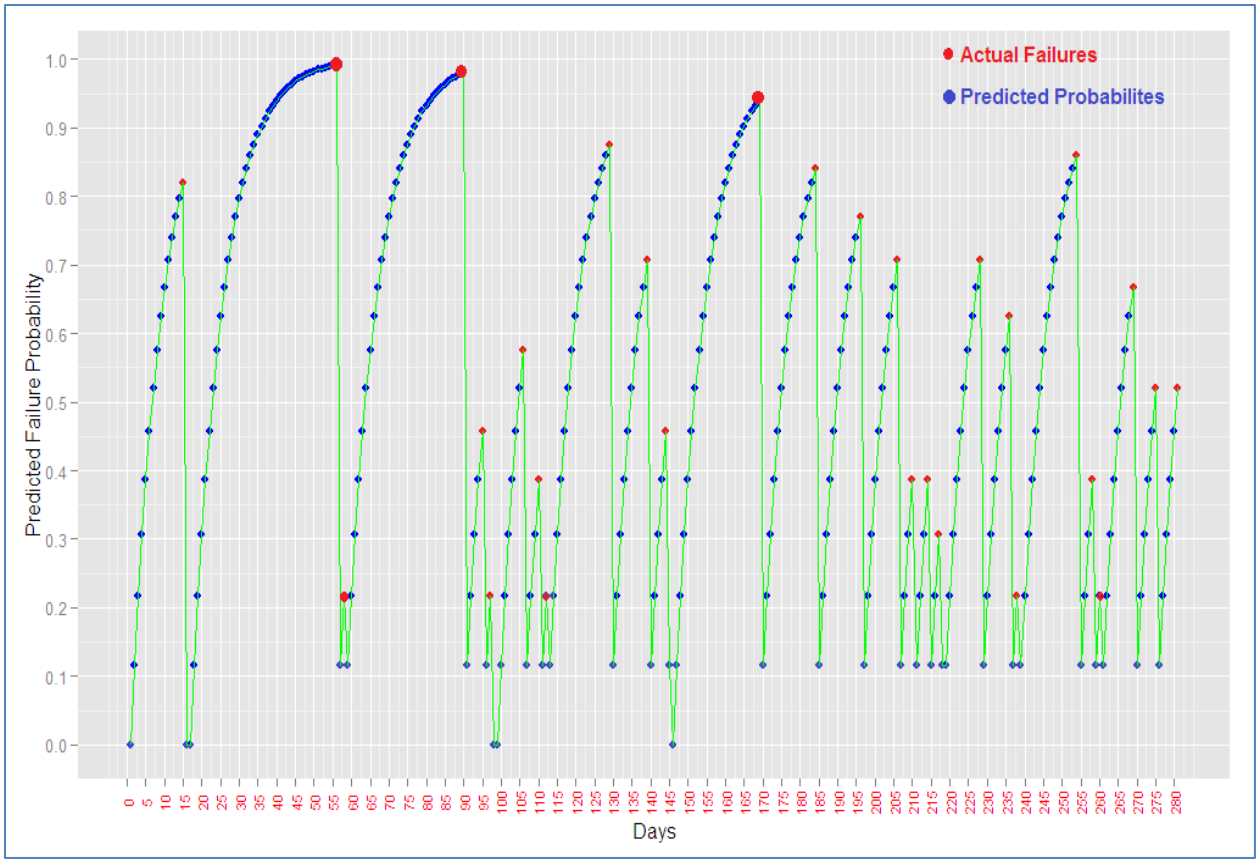


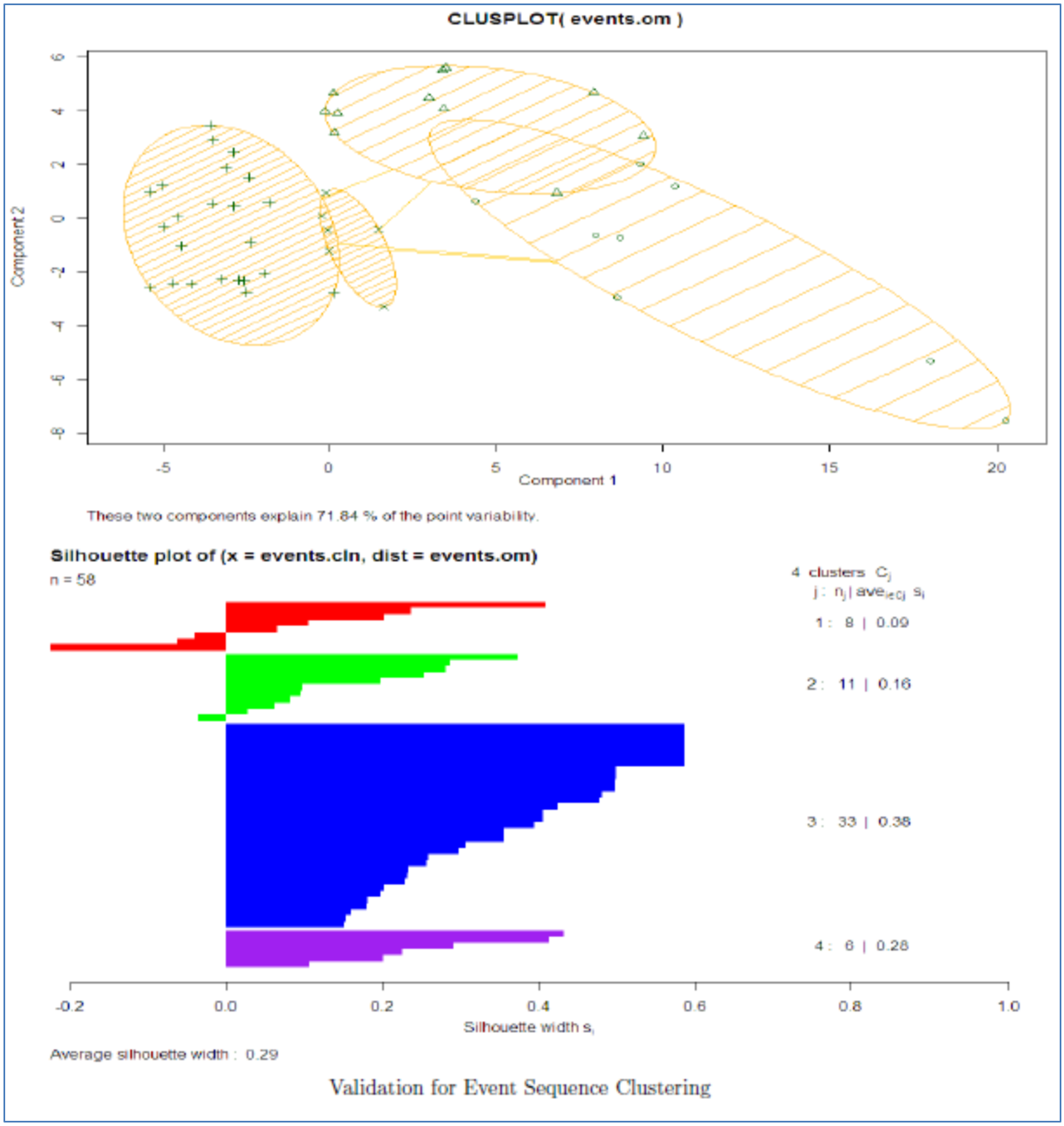
Survival Analysis (with **Cox Prop Hazard Model**)



Reliability Models


MTBF (in Days)
Device
KMedoid Clustering
MTBF (in Days)
30.45786 23%
19.81812 44%
10.68134 33%
Cluster Dendrogram (As Decision Tree)
MTBF (in Days)
19.28635
Hierarchical Clustering
13.41448
14.18407
26.69
30.85091
Height
Cluster Means
3.023626
10.15038
16.20097
22.87612
36.55167
29.58432
VL
L
M
MTBF (in Days)
H
VVH
VH

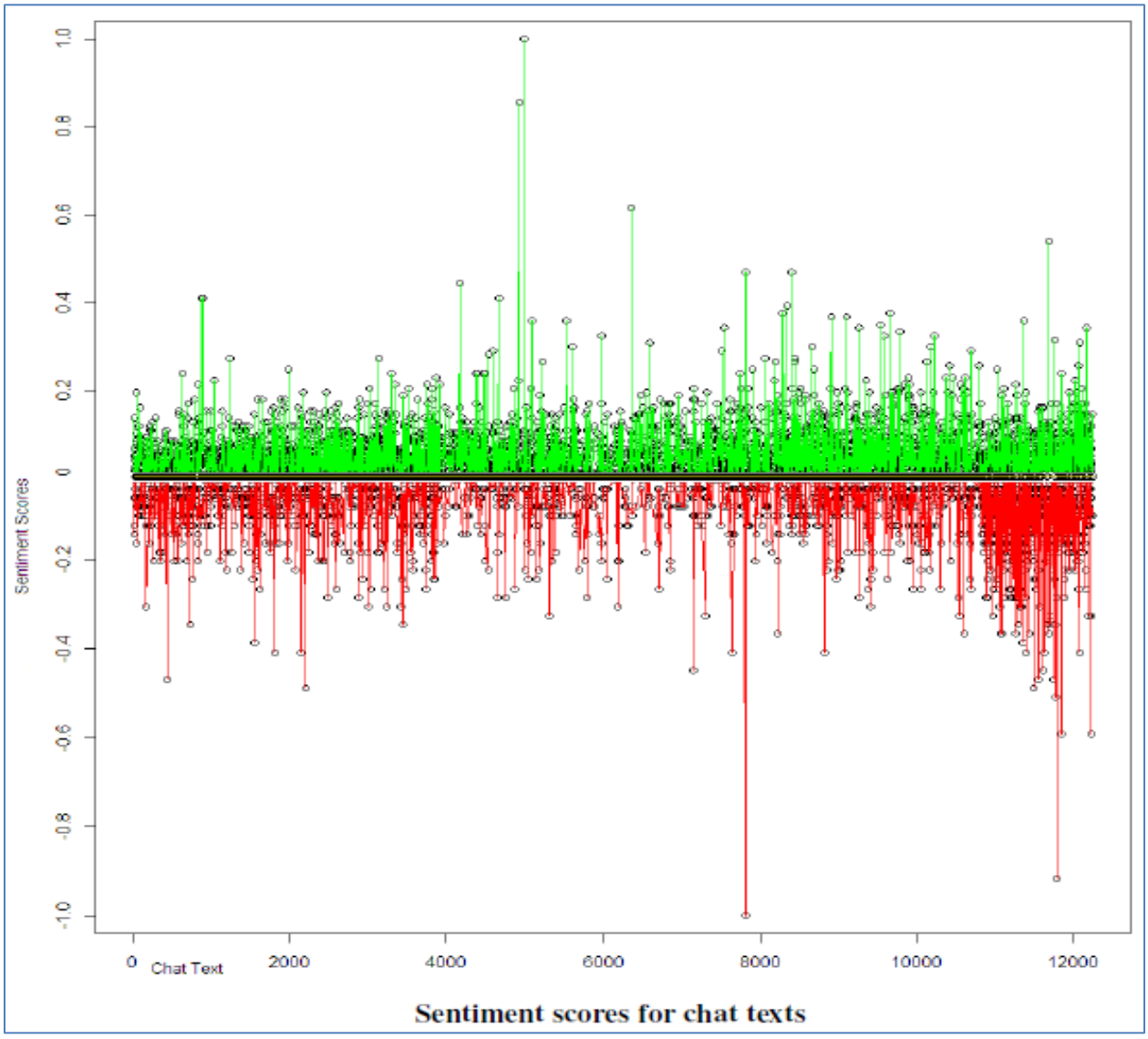

Actual Failures
Predicted Probabilites
Predicted Failure Probability
Days

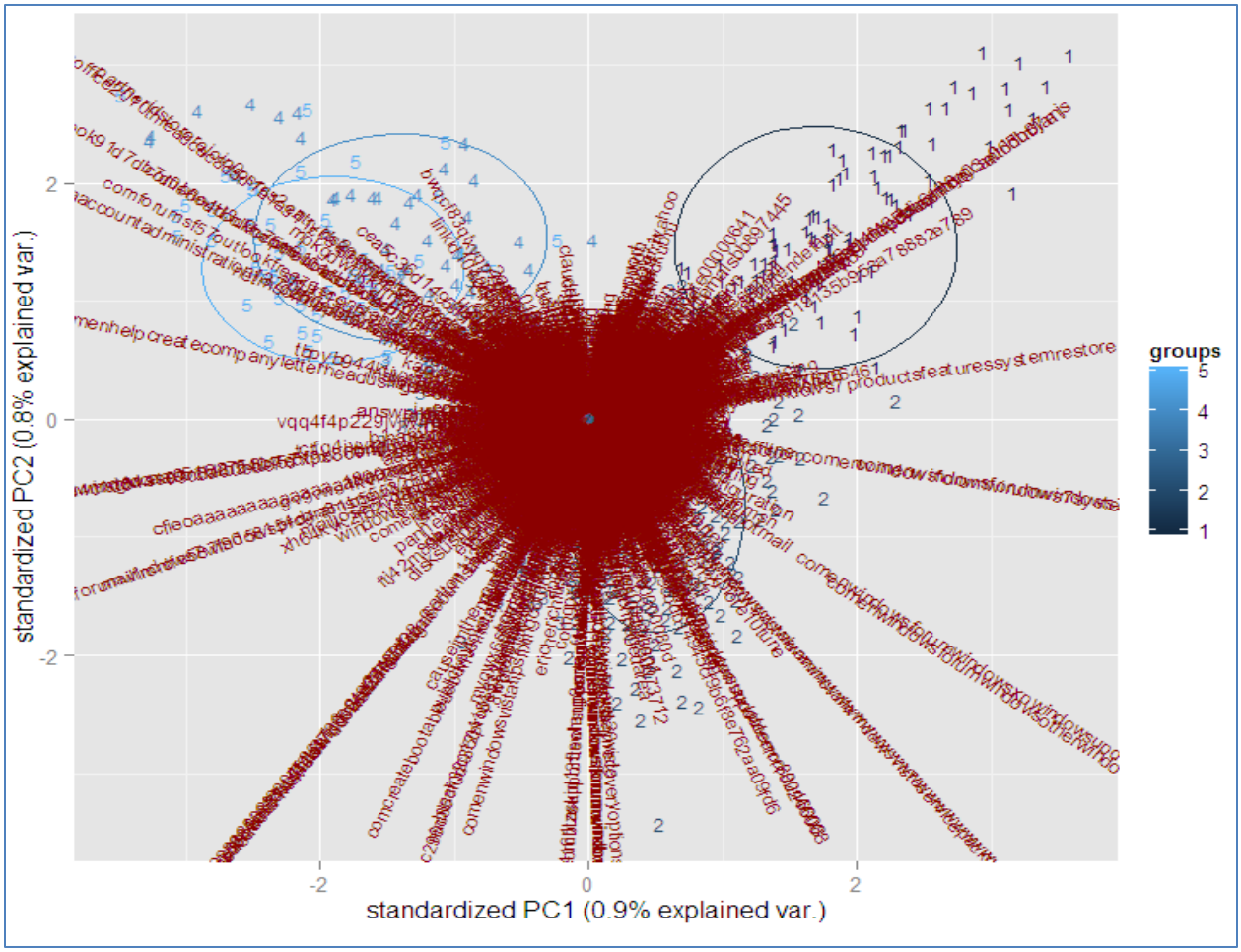


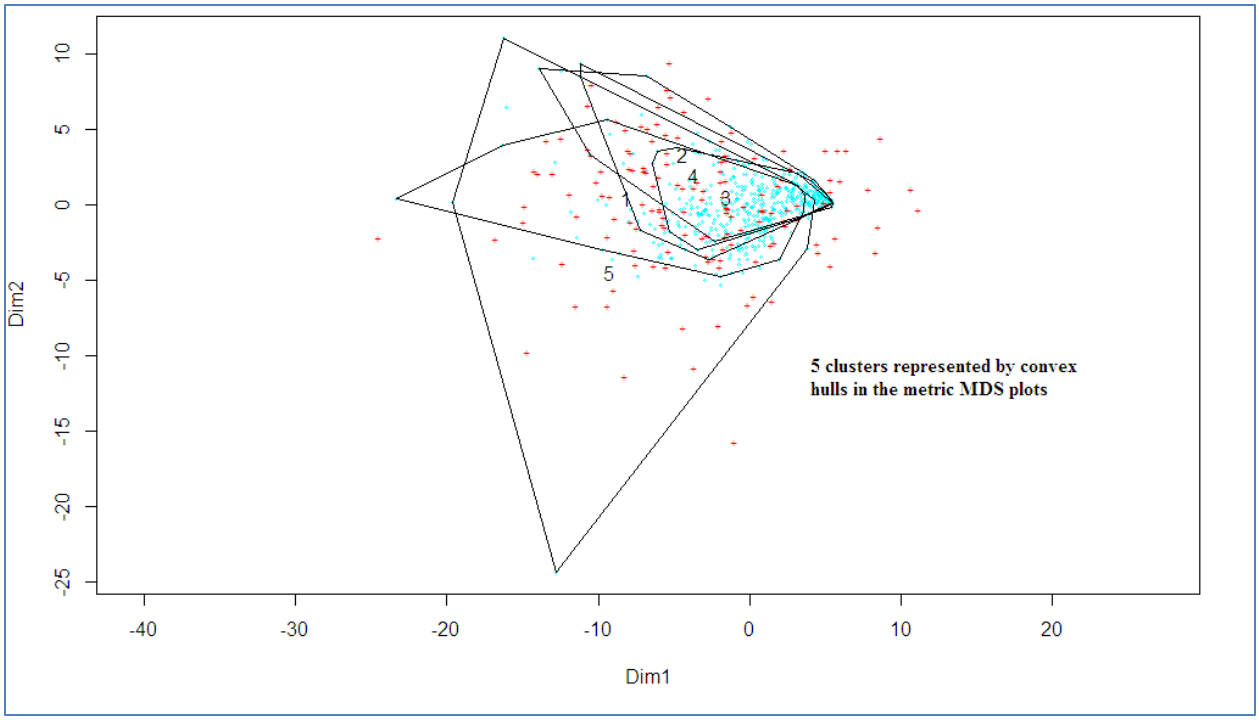
Validation for Event Sequence Clustering

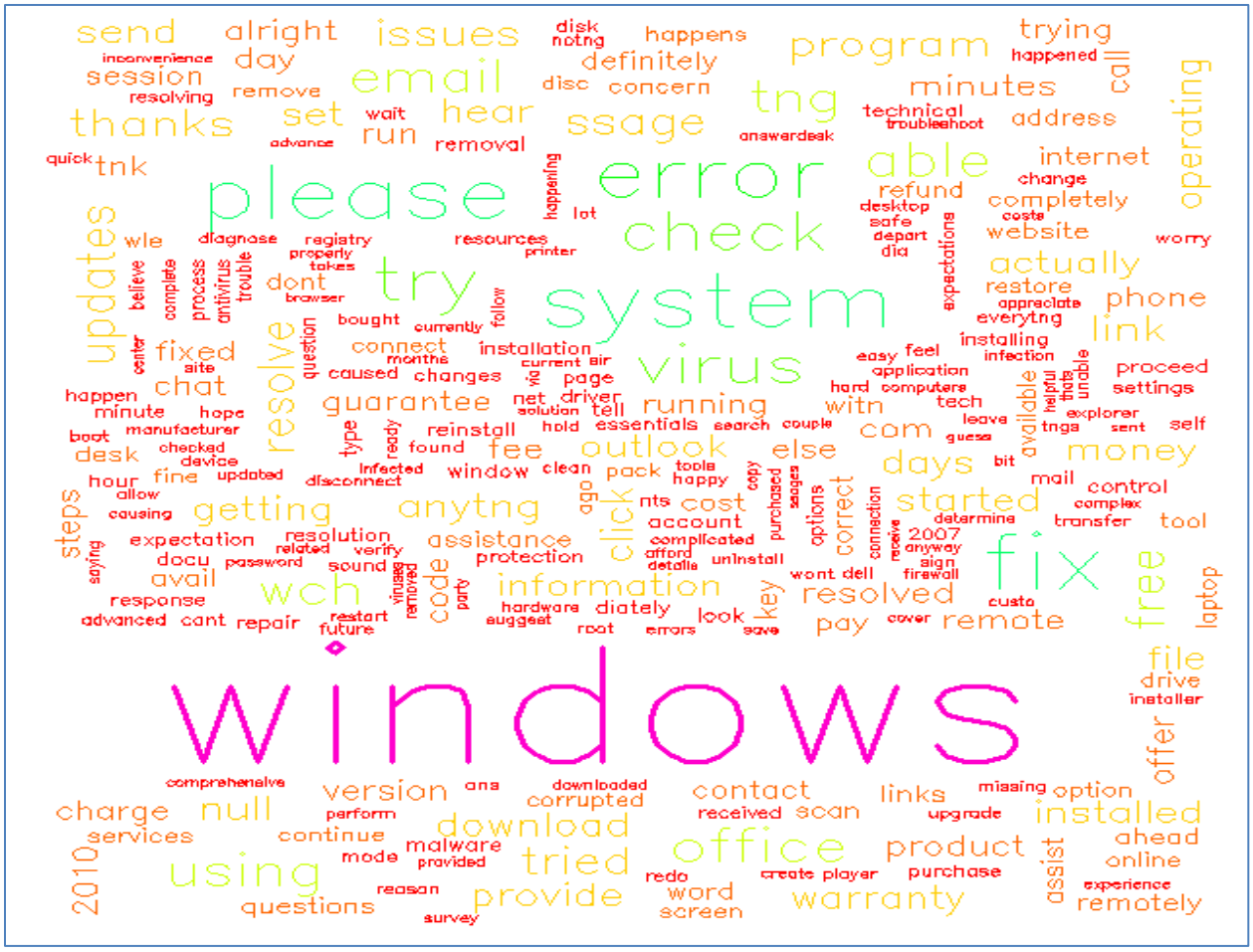
Problem 2: Text Clustering and Sentiment Analysis (from chat texts)

ID	Sentiment Scores	Sentiment Scores (normalized)	ByOutcome	Words with positive sentiments
78295341183350936	3	0.029411765	Consult	support good right thank well thank great well
84943501183499156	0	0	Consult	recovery work great
88643461183649791	3	0.029411765	Consult	work fine work
91020391183725062	-1	-0.009803922	Consult	worked like like support work support work work smooth thank good good well available great
99903891184040823	-5	-0.049019608	Premium	fine right right glad support like available like like good support enough good
112202311184458858	2	0.019607843	Consult	best welcome
113527741184511543	3	0.029411765	Consult	welcome fairly top
122210591184781985	2	0.019607843	Consult	best like excel support advanced support support helpful work prompt like improve great
148394761185612978	0	0	Consult	welcome important fine positive improving
214438041186888741	-2	-0.019607843	Premium	correctly support support thank
217346731186965584	12	0.117647059	Consult	win win well convenience available win win clean like pleasure thank good
219158921187013896	8	0.078431373	Consult	lean lean lean like thank thank correct worked productive advantage lean thank good successfu
235665771186171554	-6	-0.058823529	Premium	welcome work properly refund refund thank support thank sincerely support free free refund r
265658401187868619	0	0	Consult	win support like free important
270765671188002032	3	0.029411765	Consult	right support ready good like work well like improved defender safe
282740041188258284	12	0.117647059	Consult	support complimentary clean thank welcome work advanced support support support compreh
306012521188535324	3	0.029411765	Premium	free thank great love work

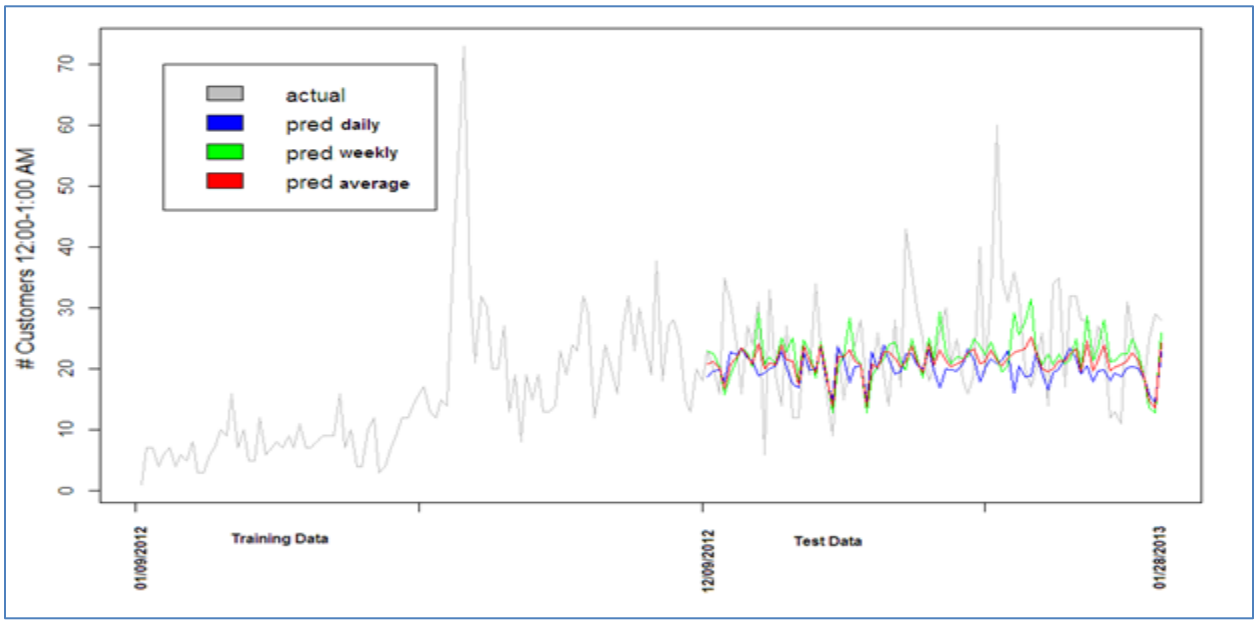


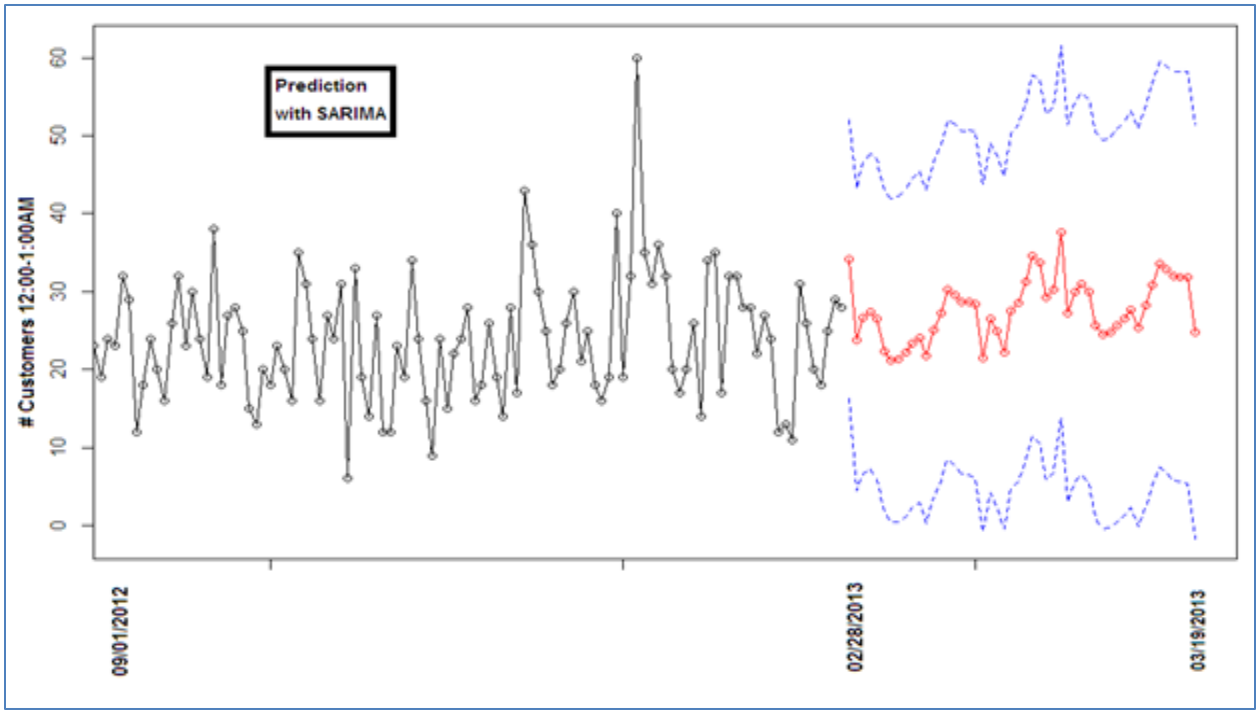

standardized PC2 (0.8% explained var.)
standardized PC1 (0.9% explained var.)
groups
5
4
3
2
1

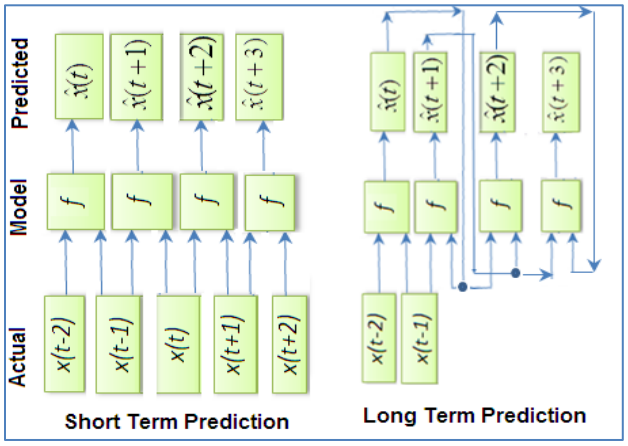

Dim2
Dim1
5 clusters represented by convex hulls in the metric MDS plots



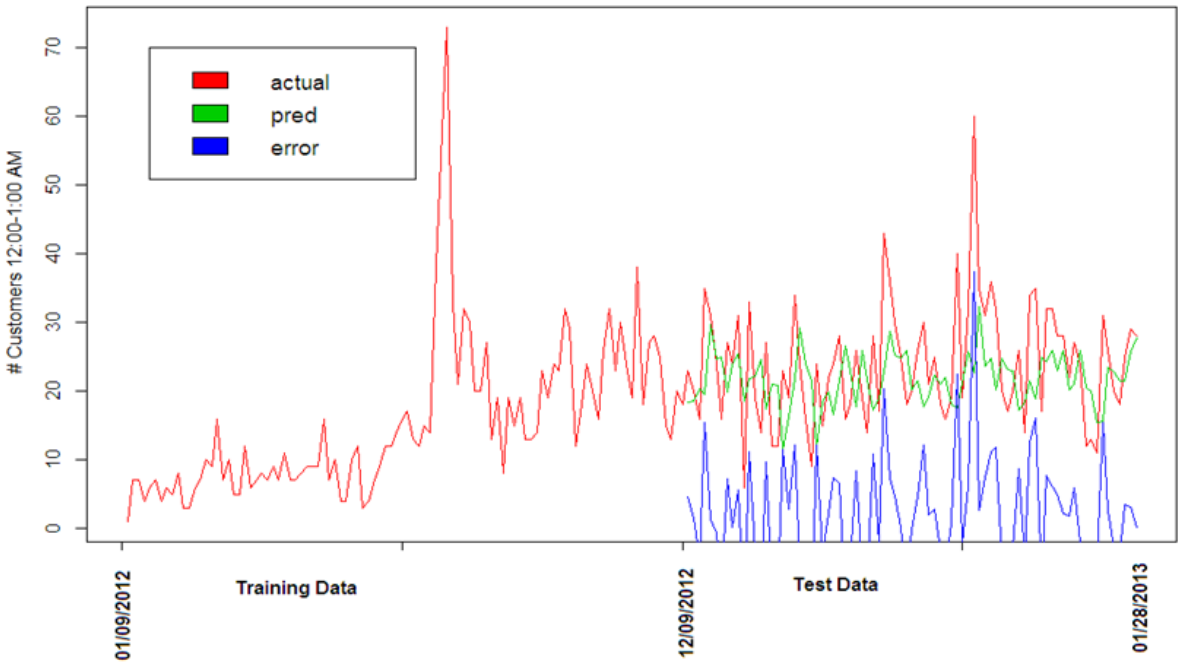
Problem 3: Predict Customer Arrival (# Customers to arrive in future)



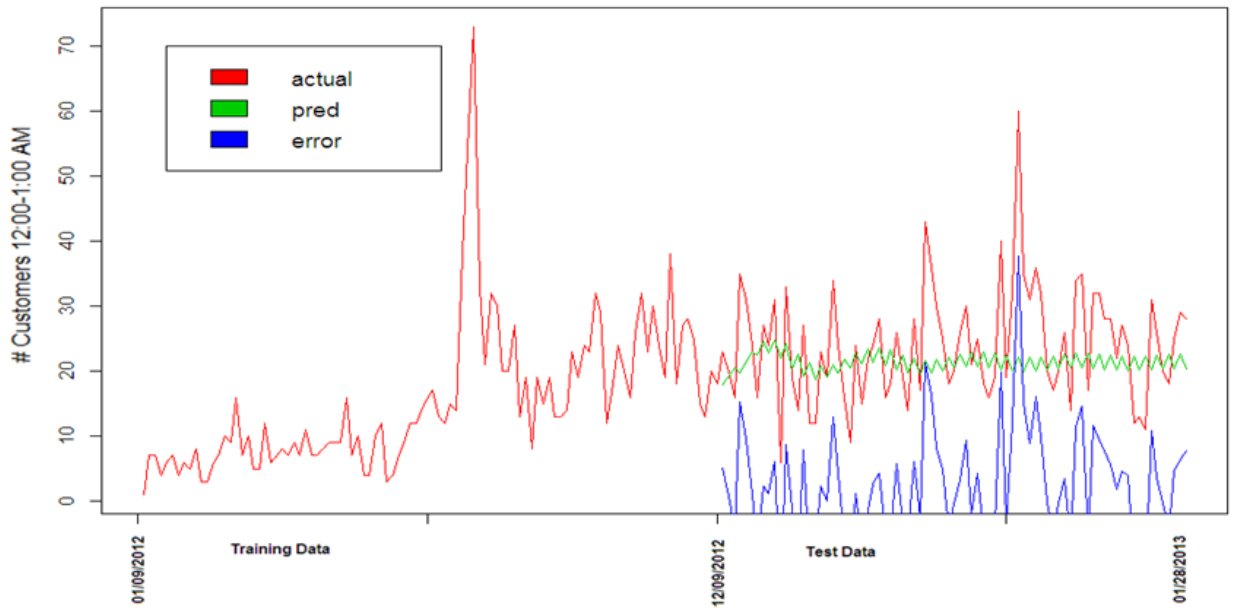

Prediction
with SARIMA
Customers 12:00-1:00AM
0
10
20
30
40
50
60
09/01/2012
02/28/2013
03/19/2013


Predicted
Model
Actual
f
x(t-2)
x(t-1)
x(t)
x(t+1)
x(t+2)
Short Term Prediction
Long Term Prediction

Short term prediction with SVM (RMSE=9.0436, MAD=7.0240) with $\hat{x}(t) = f_{SVM}(x(t-1), x(t-2), x(t-7))$

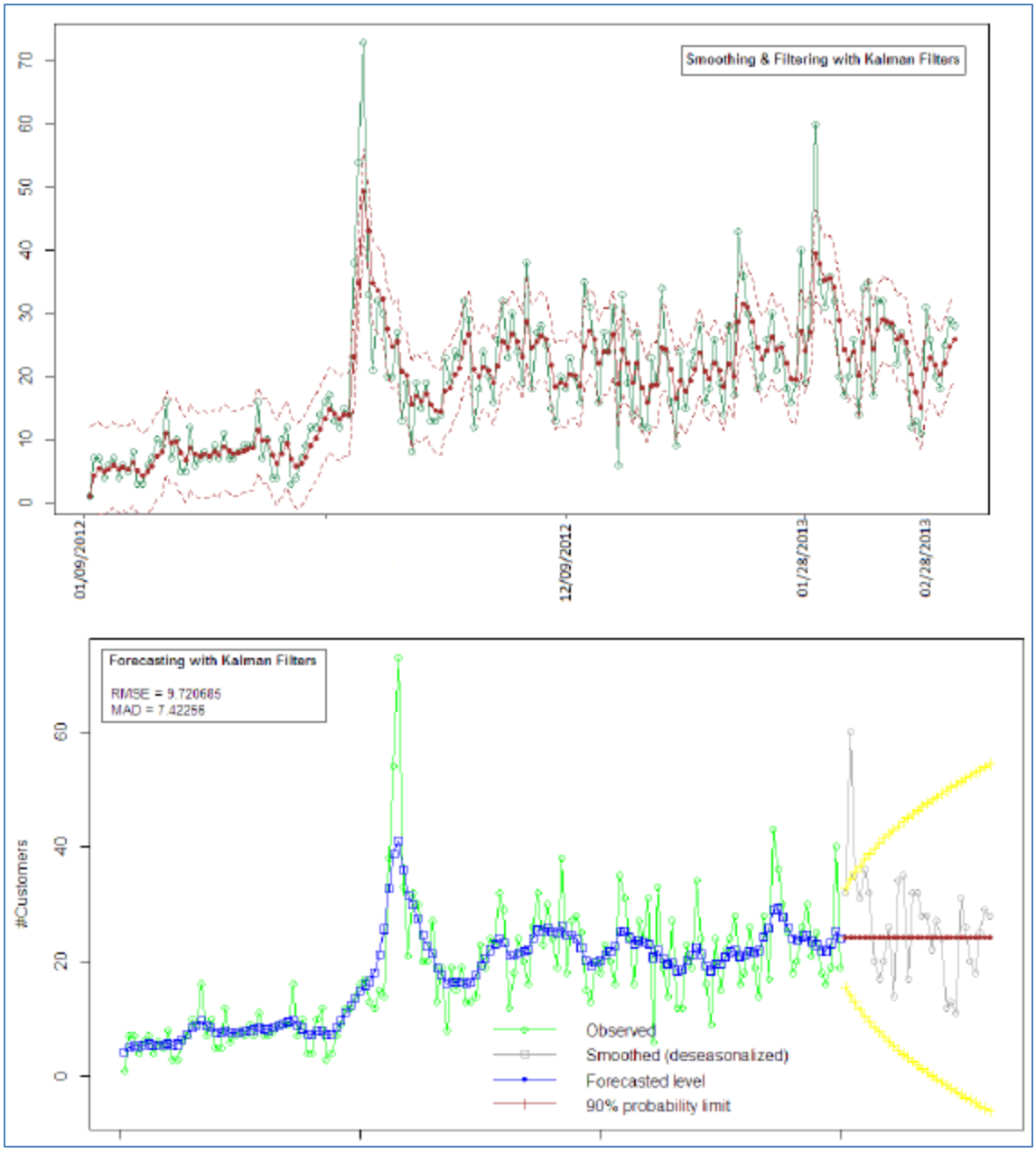


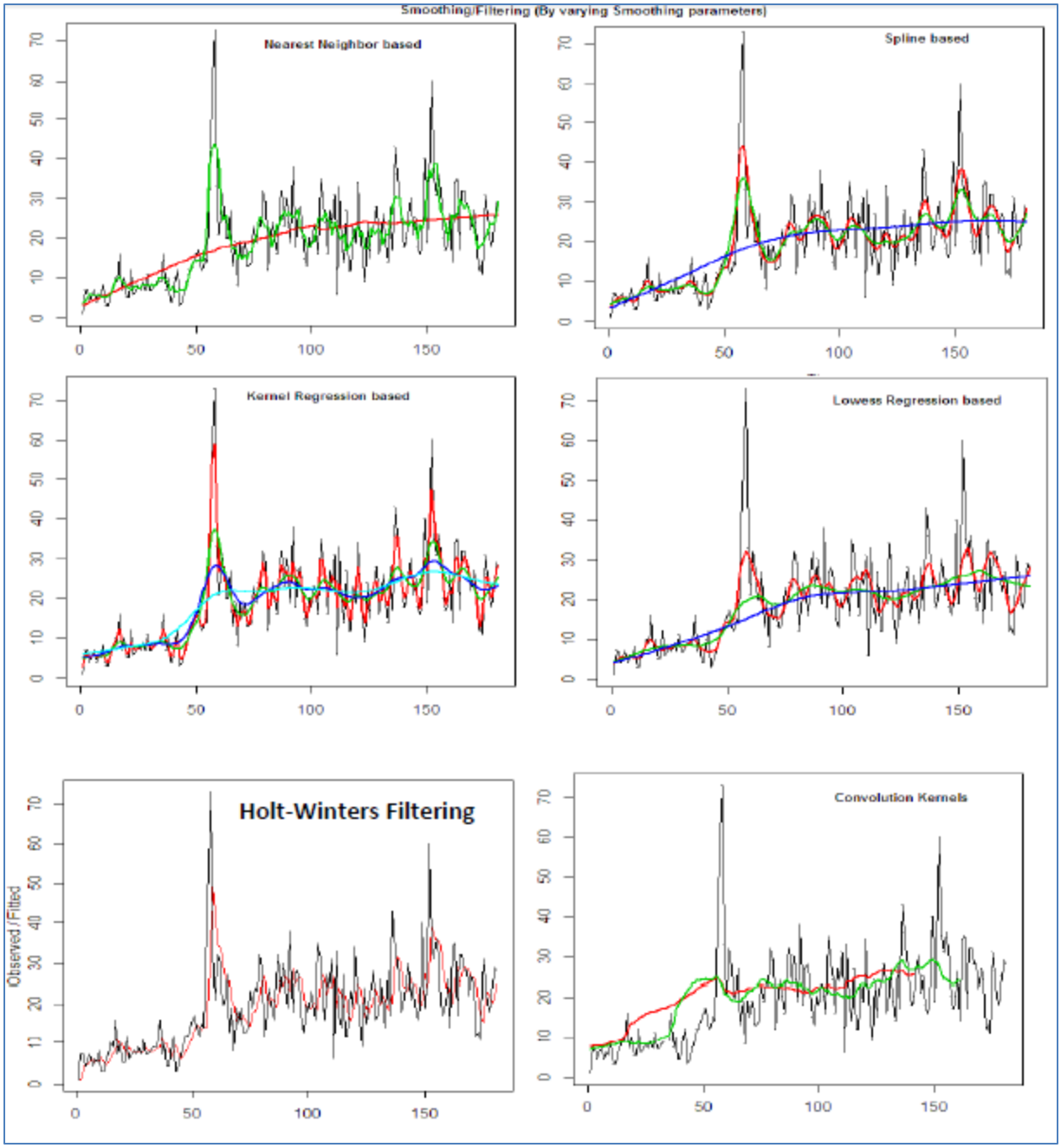
Long term prediction with 2 SVM models (RMSE= 9.075, MAD= 6.998)



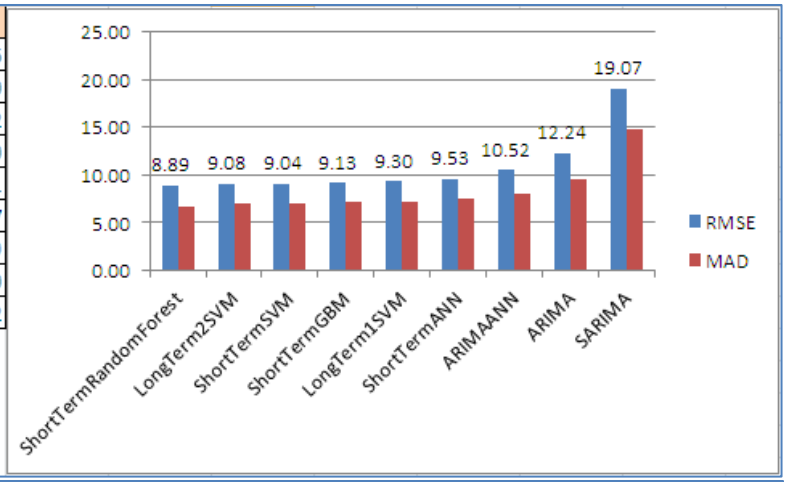
1st SVM predicts the #customers for **odd** t, with $\hat{x}(t) = f_1(x(t-1), x(t-2), x(t-5), x(t-6), x(t-7)), \; t \in Z^+_{odd}$

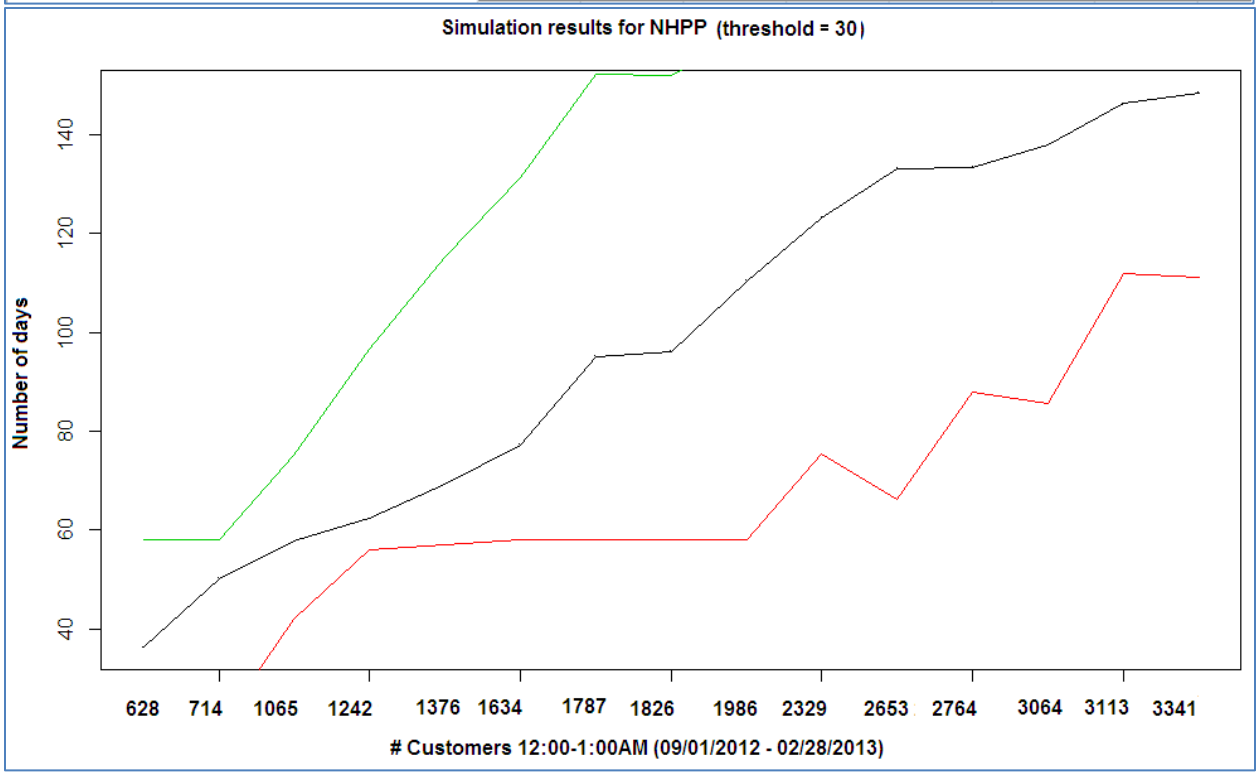
2nd SVM predicts the #customers for **even** t, with $\hat{x}(t+1) = f_2(x(t-1), x(t-2), x(t-5), x(t-6), x(t-7))$


Smoothing & Filtering with Kalman Filters
0
10
20
30
40
50
60
70
01/09/2012
12/09/2012
01/28/2013
02/28/2013
Forecasting with Kalman Filters
RMSE = 9.720685
MAD = 7.42256
#Customers
0
20
40
60
Observed
Smoothed (deseasonalized)
Forecasted level
90% probability limit

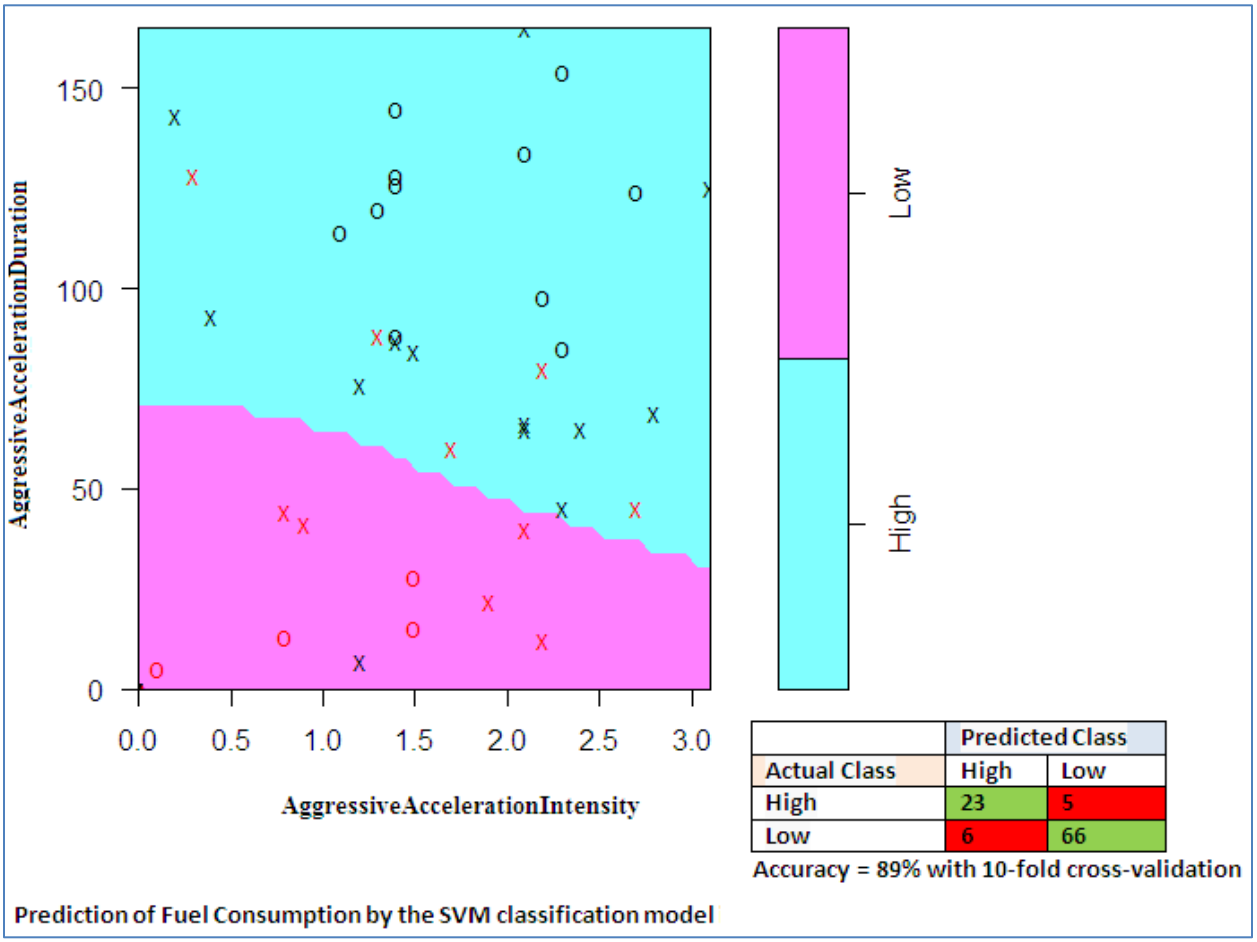

Smoothing/Filtering (By varying Smoothing parameters)
Nearest Neighbor based
Spline based
Kernel Regression based
Lowess Regression based
Holt-Winters Filtering
Convolution Kernels
Observed / Fitted

Model	RMSE	MAD
ShortTermRandomForest	8.89	6.66
LongTerm2SVM	9.08	7.00
ShortTermSVM	9.04	7.02
ShortTermGBM	9.13	7.10
LongTerm1SVM	9.30	7.11
ShortTermANN	9.53	7.47
ARIMAANN	10.52	7.99
ARIMA	12.24	9.60
SARIMA	19.07	14.82





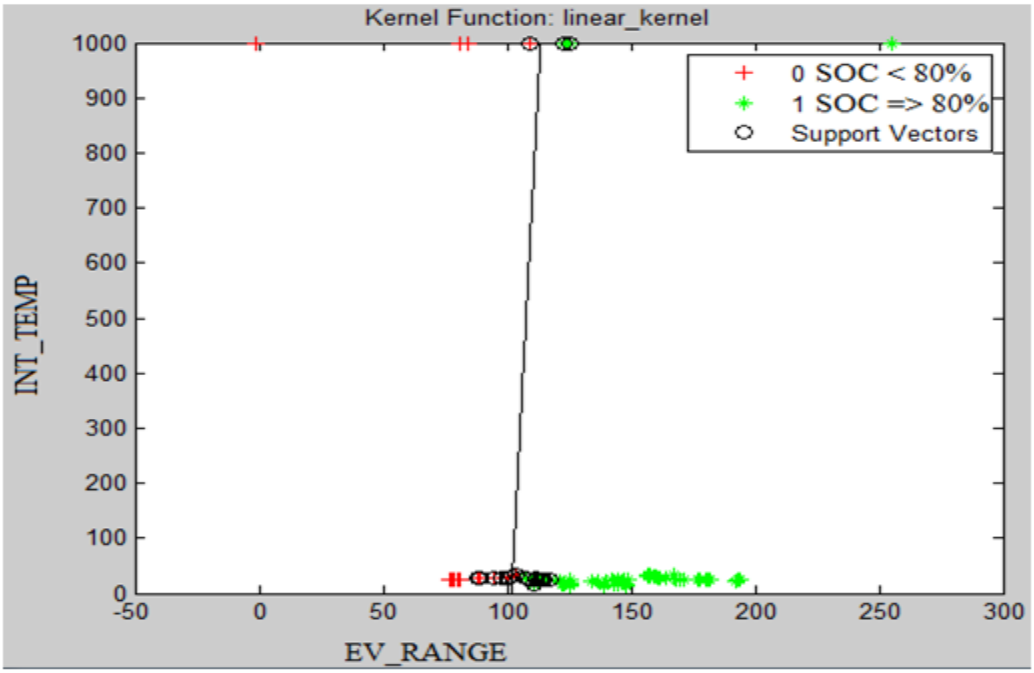
Problem 4: Supervised Models (Predict Total Cost of Estimation / State of Charge)



	Predicted Class	
Actual Class	High	Low
High	23	5
Low	6	66

Accuracy = 89% with 10-fold cross-validation

Prediction of Fuel Consumption by the SVM classification model



SVM (Training Phase)