Improving Analog Product knowledge using Principal Components Variable Clustering in JMP on test data.

Yves Chandon, Master BlackBelt at Freescale Semiconductor

Feb 27 .2015

freescale™

External Use

We Touch Millions of Lives Every Day

At Home | At Work | In your Car | Online & On the Go

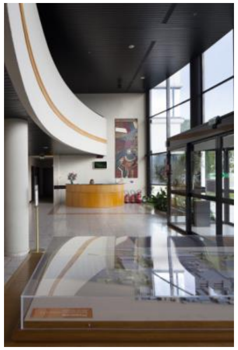
We Touch Your Life **Every Day**

Automotive | Applications & Products

	Driver Assistance	Powertrain	Safety	Body	Cockpit
					
Applications	• Active Cruise Control • Blind-spot Detection • Collision Warning & Prevention • Emergency Braking • Night Vision • Surround View Park Assist	• Engine Management • Transmission Control • Alternator Regulator • Hybrid Electric Inverter Controller • Battery Management	• Braking • Chassis • Airbags • Electronic Stability Control • Electronic Power Steering • Tire Pressure Monitoring System	• Body Control Modules • Secure -Vehicle Networking • Doors, Window Lifts, Seat Control • Security, Lighting • Heating, Ventilation, Air Conditioning	• Instrument Cluster • Infotainment • Navigation • Internet of Things connectivity
Products	• Millimeter Wave Radar Transceivers • Microcontrollers • Microprocessors • Sensors • Power Supply / Management	• Microcontrollers • Power Management • Sensors • System Basis Chips • Injector Drivers	• Microcontrollers • Power Management • Drivers • Network Transceivers • RF Transmitters • Sensors	• System Basis Chips • Microcontrollers • Network Transceivers • Drivers & Switches • Sensors • Solution Integration	• Applications Processors • Microcontrollers • System Basis Chips • Power Management • Audio Codecs • Integrated Graphics Processors

Freescale Semiconducteurs France – Toulouse

- Largest Freescale facility in EMEA
- Broad range of skills & competencies
- Global or European leadership for programs in
    - Analog
    - RF high power
    - Radar & RF connectivity
    - Sensors
- Leveraging on a rich R&D ecosystem
- *Location of a Freescale Discovery Lab*



Headcount : 500 employees
Turn over (2013) : 73 millions €
Operation started : 1967
Certifications : Quality : ISO9001, QS9000, ISO/TS 16949,
Environment : ISO14001

Introduction

Analog Integrated Circuits for automotive are used in many applications like Breaking systems, Airbags, Lighting, Injection Driving etc… Some of those application are very critical for safety.

The parts are manufactured on silicon wafers. It takes about 3 months of process for a 25 wafers lot, each wafer having 1000 dies. A first set of 1000 tests is performed on each die of the wafers. (Typical values)

A final product is obtained after sawing the wafers and assembling silicon dies in a package. Cars must work in Siberian winter as well as Sahara's summer. For this reason many products are tested at -40°C and 125°C on 2000 tests.

Principal components

Test files are tables of 100 to 5000 columns containing from a few lines to 300000 lines.

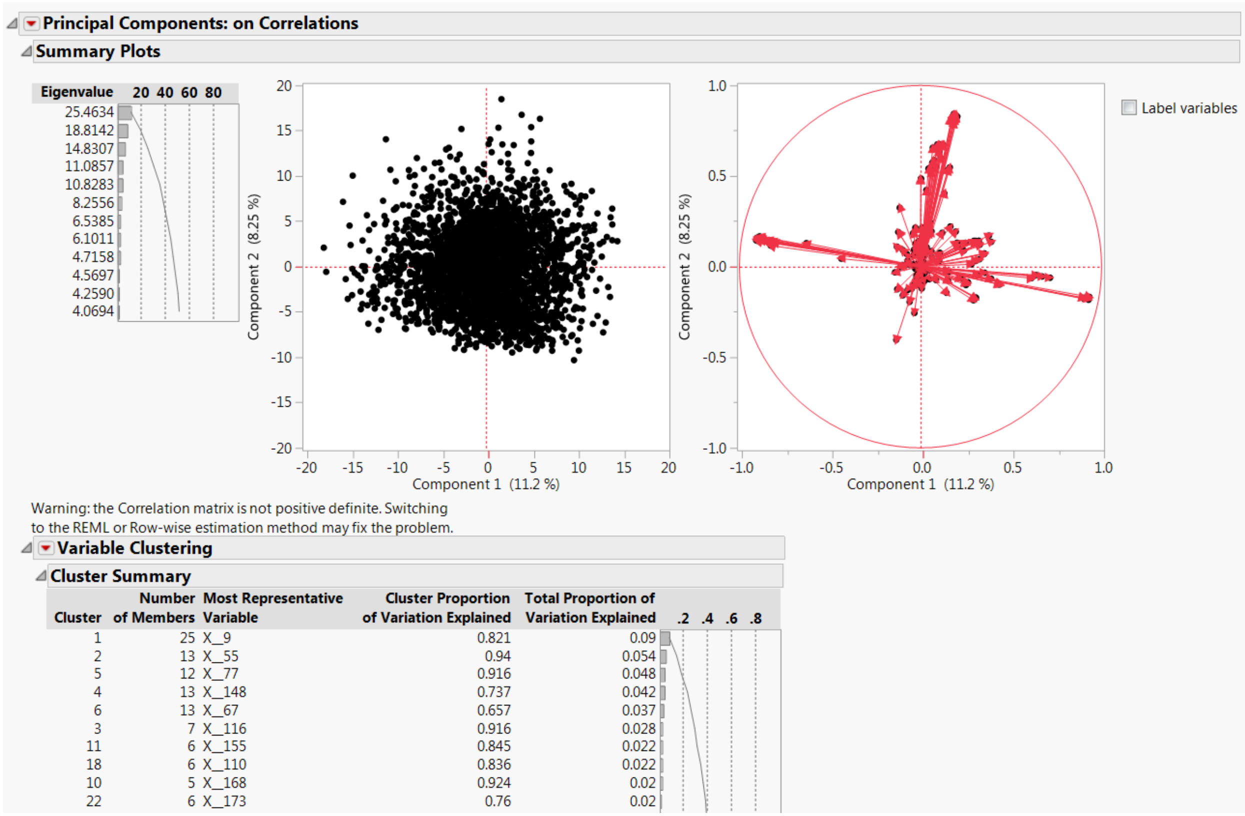
We know the Pareto principle ” for many events, roughly 80% of the effects come from 20% of the causes”. *With so many tests to handle, could we find the minority of tests that would represent the rest?* Those would be the “vital few”

Principal Components Analysis has been known for long as a useful technique for variable reduction. However from a practical point of view, principal components are not always convenient for the subject matter expert.

JMP now includes a « Cluster variables » command within the Principal Components procedure that is very helpful. The clusters are groups of variables that are somehow similar. The clusters are not orthogonal to each other like principal components

The file used is a reduced set for demonstration purpose. It has 228 variables and 3000 rows

Clustering example



Cluster	Number of Members	Most Representative Variable	Cluster Proportion of Variation Explained	Total Proportion of Variation Explained
1	25	X_9	0.821	0.09
2	13	X_55	0.94	0.054
5	12	X_77	0.916	0.048
4	13	X_148	0.737	0.042
6	13	X_67	0.657	0.037
3	7	X_116	0.916	0.028
11	6	X_155	0.845	0.022
18	6	X_110	0.836	0.022
10	5	X_168	0.924	0.02
22	6	X_173	0.76	0.02

- The graph and eigenvalue summary of Principal Components is displayed
- Cluster summary is displayed by order of total variation explained
- For each cluster, the proportion of variation explained by the cluster is also displayed
-

Cluster members Table

Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	X_9	0.949	0.458	0.093
1	X_8	0.945	0.458	0.102
1	X_32	0.91	0.433	0.159
1	X_31	0.907	0.425	0.161
1	X_24	0.895	0.408	0.177
1	X_25	0.894	0.412	0.18
1	X_30	0.894	0.428	0.185
1	X_26	0.888	0.398	0.186
1	X_35	0.886	0.407	0.193
1	X_34	0.872	0.39	0.209
1	X_21	0.871	0.412	0.219
1	X_20	0.85	0.373	0.239
1	X_22	0.875	0.484	0.242
1	X_27	0.824	0.305	0.254
1	X_23	0.86	0.496	0.278
1	X_11	0.848	0.531	0.325
1	X_10	0.846	0.531	0.329
1	X_33	0.768	0.357	0.361
1	X_29	0.76	0.337	0.362
1	X_28	0.765	0.356	0.365
1	X_18	0.754	0.352	0.38
1	X_19	0.744	0.339	0.387
1	X_17	0.764	0.45	0.429
1	X_16	0.746	0.484	0.492
1	X_124	0.202	0.102	0.889
2	X_55	0.974	0.27	0.036
2	X_54	0.971	0.267	0.039
2	X_56	0.969	0.271	0.043
2	X_53	0.966	0.269	0.047
2	X_57	0.963	0.259	0.05
2	X_48	0.948	0.247	0.069

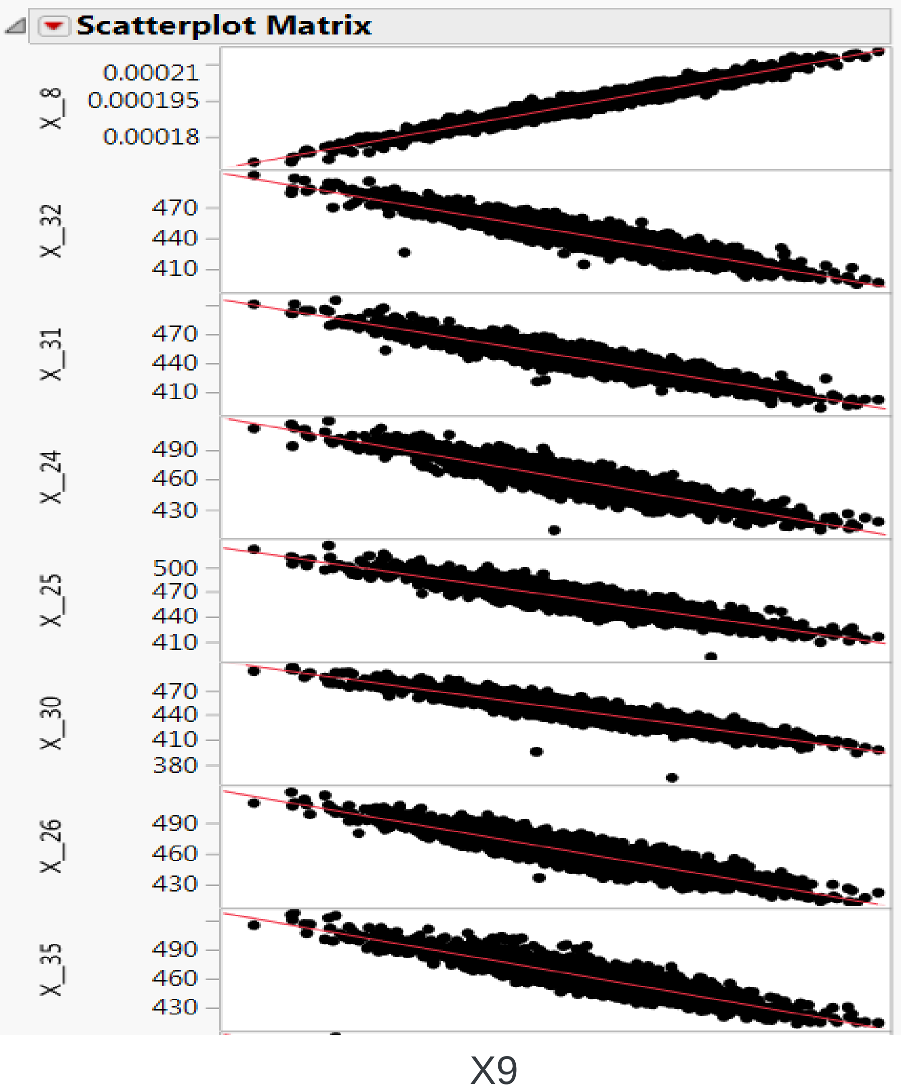
Cluster Members

Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
12	X_185	0.954	0.041	0.047
12	X_189	0.953	0.041	0.049
12	X_181	0.89	0.064	0.118
12	X_89	0.537	0.1	0.514
13	X_164	0.902	0.144	0.114
13	X_163	0.902	0.147	0.115
13	X_162	0.898	0.13	0.118
13	X_165	0.89	0.111	0.124
14	X_186	0.933	0.023	0.069
14	X_190	0.933	0.024	0.069
14	X_182	0.912	0.022	0.09
14	X_90	0.554	0.029	0.459
14	X_97	0.23	0.037	0.8
15	X_188	0.962	0.23	0.049
15	X_184	0.943	0.217	0.073
15	X_192	0.934	0.214	0.084
15	X_196	0.225	0.178	0.943
16	X_86	0.998	0.146	0.002
16	X_42	0.997	0.147	0.004
16	X_136	0.993	0.141	0.008
17	X_215	0.975	0.039	0.026
17	X_223	0.873	0.039	0.132
17	X_219	0.867	0.059	0.142
18	X_110	0.906	0.38	0.152
18	X_111	0.906	0.382	0.153
18	X_113	0.905	0.382	0.154
18	X_76	0.87	0.322	0.192
18	X_82	0.733	0.273	0.367
18	X_81	0.694	0.401	0.511
19	X_41	0.753	0.173	0.299
19	X_40	0.711	0.268	0.394

- The table displays the list of clusters with all the members and how well the test variable is correlated with its own cluster.
- Unlike Principal Components, clusters are not orthogonal
- This info can be extracted easily with « make into data table »

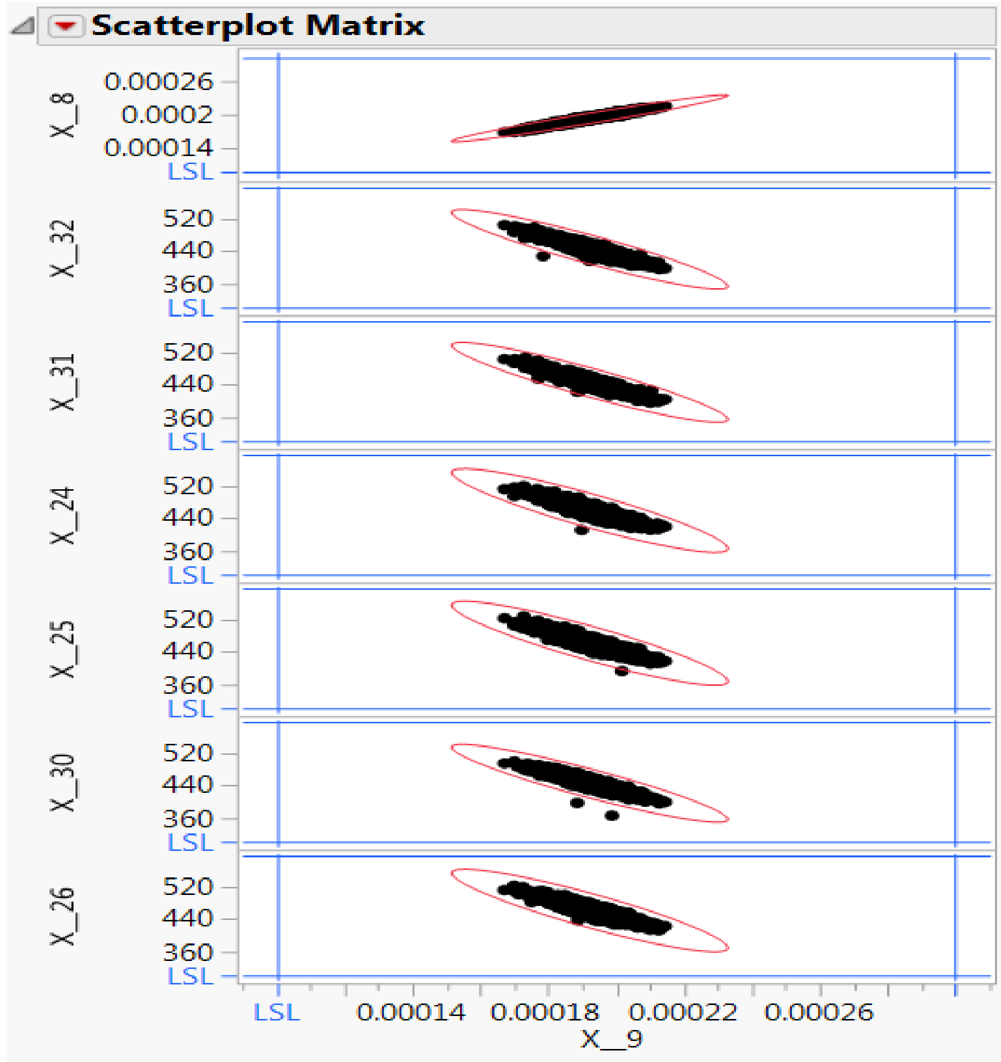
How well does the most representative variable correlates with others?

- For each group, a fit is performed between the most representative variable of the group and the rest of the group with the scatterplot matrix
- It is easy to see how well the most representative represents the rest



How can we use this information?

- It is interesting to do again this analysis on the same file with specifications adding a 0.999999999 confidence ellipse
- If we tighten X_9 limits, then we have a very high confidence that the remaining parameters will be in spec, providing we can explain the behavior of the outliers.

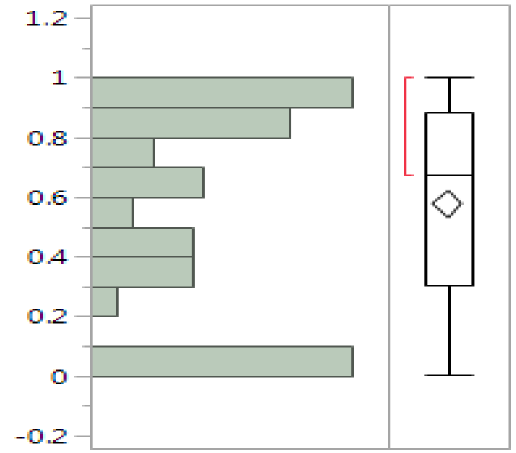


How well do most representative variables represent the rest

The histogram of Rsquare in each family show about 60% of the tests are explained by their most representative.

It should be possible to enhance this correlation using the multivariate tools in JMP

Rsquare



Quantiles

100.0%	maximum	0.9998886844
99.5%		0.9998874778
97.5%		0.9994449796
90.0%		0.9827135243
75.0%	quartile	0.8843182918
50.0%	median	0.6747869106
25.0%	quartile	0.3051129362
10.0%		0.0524725292
2.5%		0.0072334235
0.5%		0.0035067597
0.0%	minimum	0.0034640597

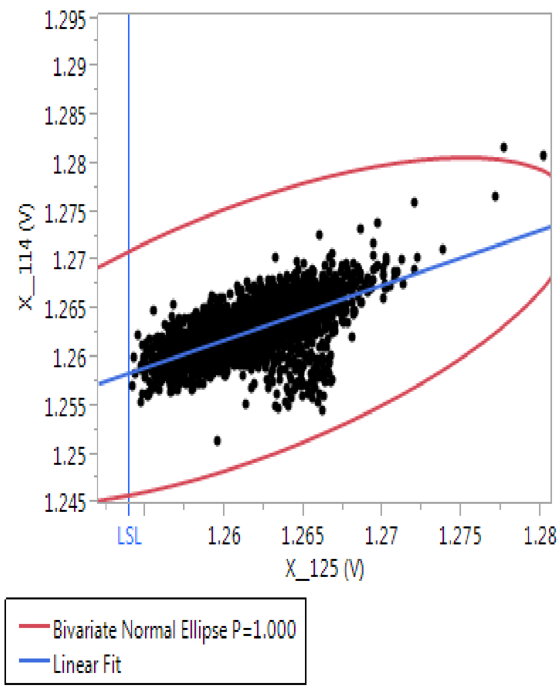
Summary Statistics

Mean	0.575294
Std Dev	0.3476792
Std Err Mean	0.0230256
Upper 95% Mean	0.6206653
Lower 95% Mean	0.5299227
N	228

Example of stepwise regression

- X_114 has a moderate correlation with X_125 with most representative variable in cluster 23
- The model given by stepwise regression between X_114 and the 45 "Most representatives" is greatly enhanced from 0.43 to 0.83.

Bivariate Fit of X_114 By X_125



Linear Fit

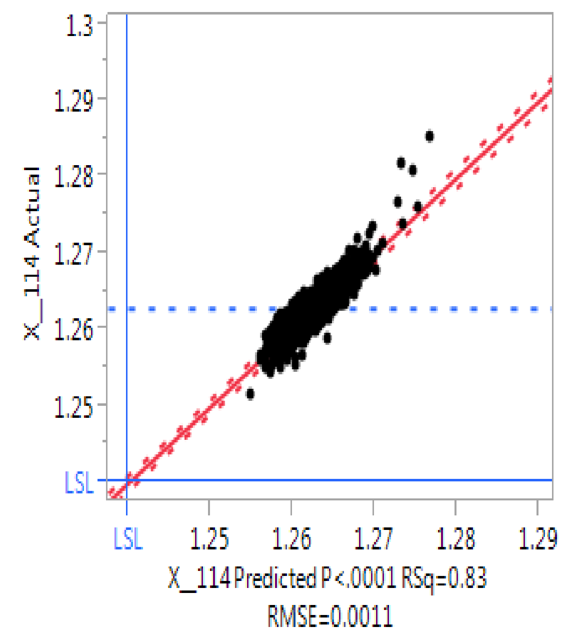
X_114 = 0.5481088 + 0.5665018*X_125

Summary of Fit

RSquare	0.43113
RSquare Adj	0.43094
Root Mean Square Error	0.002104
Mean of Response	1.262765

Response X_114

Actual by Predicted Plot



Summary of Fit

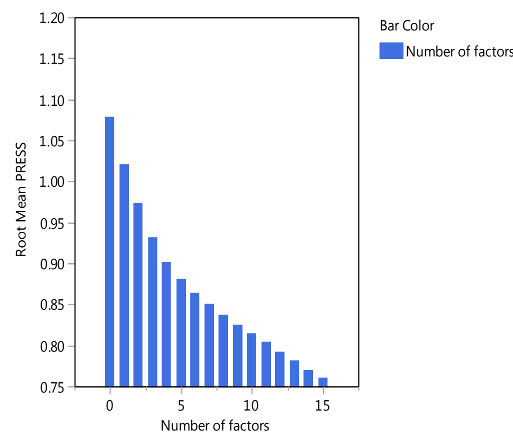
RSquare	0.833598
RSquare Adj	0.832591
Root Mean Square Error	0.001127
Mean of Response	1.262759

Enhancing the model with Partial Least Squares Regression

- Let’s try to model all the remaining test variables by the most representatives. PLS regression seems to be the best tool.
- PLS works very well to relate 2 groups of variables when the group of input variables shows some correlations.
- PLS does not show Rsquare but Percent Variation explained. It can be seen that some variables are well explained while a significant number still has a poor figure.

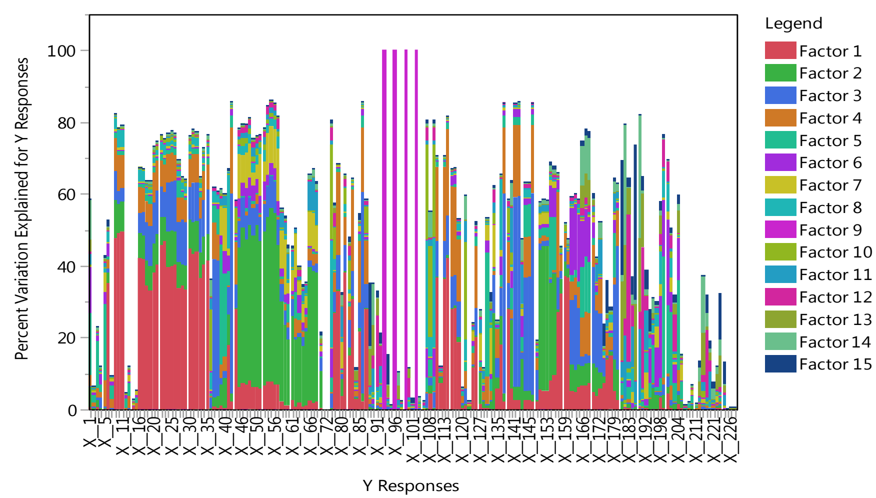
KFold Cross Validation with K= 7 and Method=NIPALS

Number of factors	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2
0	1.078223	2616.6493	0.0100*
1	1.020067	2584.5243	0.0250*
2	0.973634	2537.6667	0.3870
3	0.931166	2499.3824	0.2180
4	0.900931	2461.6392	0.0070*
5	0.880109	2395.8322	0.1430
6	0.864114	2340.2318	0.0780
7	0.850534	2273.4195	0.0300*
8	0.837453	2178.6409	0.1070
9	0.824628	2180.8305	0.1840
10	0.814317	2060.4196	0.0040*
11	0.803707	1952.9725	0.0140*
12	0.792525	1769.1244	0.3890
13	0.781051	1522.0839	0.1670
14	0.770214	1041.8675	0.6310
15	0.759962	0.000000	1.0000



Note: The minimum root mean PRESS is 0.75996 and the minimizing number of factors is 15.

Percent Variation Explained for Y Responses

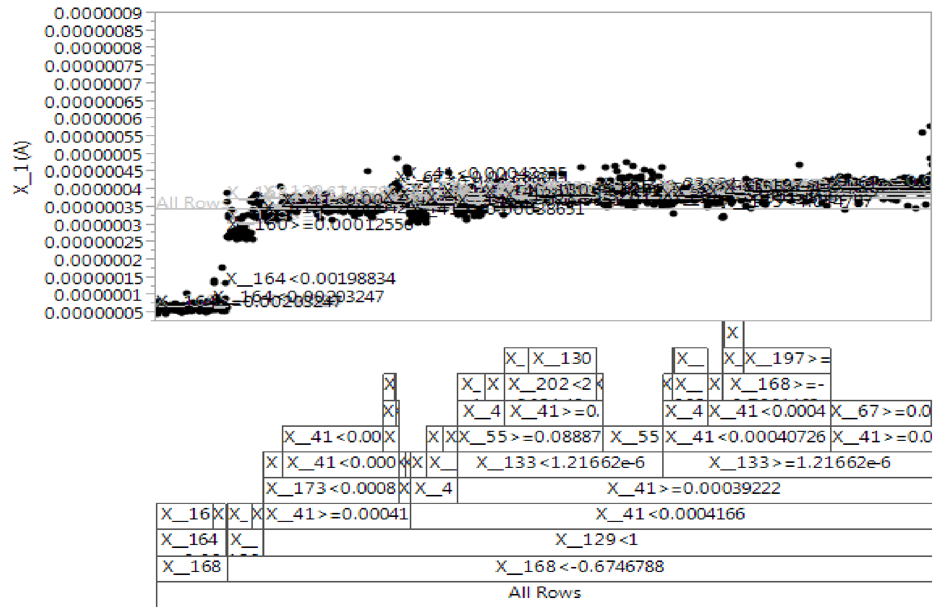


Comments on PLS results and a the possible alternative

- JMP pro PLS module is limited to 15 Factors
- While this is sufficient for many applications, it may not be enough when the numbers of variables is large.
- PLS takes into account all the variabilities of the output and input variables, therefore it is sensitive to data problems (example : variables with far outliers)
- We are looking for a multivariate technique that can give good results in terms of explaining correctly the « trivial many » by the « vital few », while keeping a good computing performance. The partition platform is another way to build our model

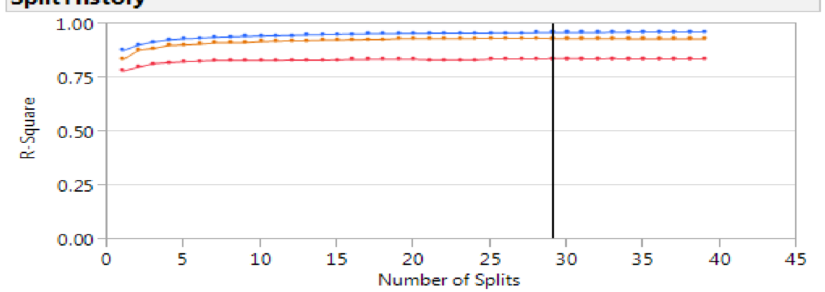
Partition platform

- A partition decision tree is performed to model all the 182 remaining variables by the 45 « most representatives » after adding a validation column.
- The CTRL Go broadcasting makes it really easy
- Computation is very fast



	RSquare	RMSE	N	Number of Splits	AICc
Training	0.962	1.8427e-8	1800	29	-58943
Validation	0.841	3.7996e-8	600		
Test	0.934	2.3353e-8	600		

Split History



Validation Data in Red

Models comparison for stepwise regression, PLS and Partition

- Models for Stepwise Regression, PLS and Partition Decision Tree have been run on the 182 variables by the 45 most representatives. For each analysis, the "save prediction formula" is used.
- Prediction formulas are entered in the Model Comparison platform
- "Make into combined data table" command builds a table with all Rsquare

Measures of Fit for X_141

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_141 Predictor	Partition		0.9892	0.0038	0.0027	3000
Pred Formula X_141 2	Fit Least Squares		0.9929	0.0031	0.0022	3000
Pred Formula X_141 3	Partial Least Squares		0.8589	0.0137	0.0108	2994

Measures of Fit for X_142

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_142 Predictor	Partition		0.9881	0.0040	0.0029	3000
Pred Formula X_142 2	Fit Least Squares		0.9924	0.0032	0.0023	3000
Pred Formula X_142 3	Partial Least Squares		0.8607	0.0137	0.0107	2994

Measures of Fit for X_143

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_143 Predictor	Partition		0.7633	0.0041	0.0031	3000
Pred Formula X_143 2	Fit Least Squares		0.8006	0.0037	0.0029	2994
Pred Formula X_143 3	Partial Least Squares		0.4886	0.0060	0.0046	2994

Measures of Fit for X_144

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_144 Predictor	Partition		0.6380	0.0411	0.0316	3000
Pred Formula X_144 2	Fit Least Squares		0.7229	0.0360	0.0282	2994
Pred Formula X_144 3	Partial Least Squares		0.6462	0.0406	0.0316	2994

Measures of Fit for X_145

Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_145 Predictor	Partition		0.6502	0.0404	0.0310	3000
Pred Formula X_145 2	Fit Least Squares		0.7241	0.0359	0.0281	2994
Pred Formula X_145 3	Partial Least Squares		0.6475	0.0405	0.0316	2994

Measures of Fit for X_147

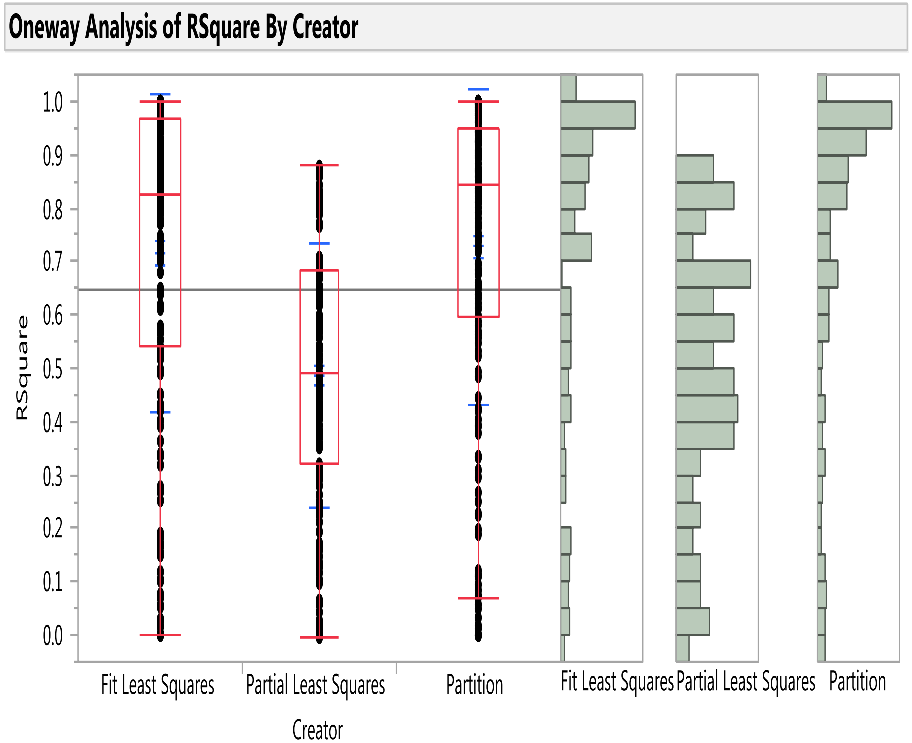
Predictor	Creator	.2.4.6.8	RSquare	RASE	AAE	Freq
X_147 Predictor	Partition		0.9990	0.0011	0.0008	3000
Pred Formula X_147 2	Fit Least Squares		0.9995	0.0008	0.0006	3000
Pred Formula X_147 3	Partial Least Squares		0.8631	0.0133	0.0104	2994

Measures of Fit for X_149

Models comparison for stepwise regression, PLS and Partition (2)

Method	Compution time	Modeling ability
Stepwise	40 mn	Good
PLS	30 s	Average
Decision Tree	15 s	Good

Stepwise and Decision Tree improve the average Rsquare over PLS but also over simple regression (0.57 to 0.72)



Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Fit Least Squares	183	0.715330	0.299101	0.02211	0.67171	0.75896
Partial Least Squares	170	0.487653	0.247064	0.01895	0.45025	0.52506
Partition	183	0.727032	0.295992	0.02188	0.68386	0.77020

Conclusion

- Principal Components Clustering procedure combined with multivariate techniques are very useful tools for variable reduction.
- Partition Decision Tree is particularly efficient. It combines fast computation with good modeling accuracy
- Overall, the Pareto Principle works on this data set. It is possible to represent fairly well 80% of the variables by 20%.

- This information is useful to have a better picture of the product. It can be used for quality improvement as well as cost savings.