Paper 268-27

Trees, Neural Nets, PLS, I-Optimal Designs and other New JMP® Version 5 Novelties

John P. Sall, Executive Vice President

Cathy Maahs-Fladung, JMP Technical Product Manager

JMP, A Business Unit of SAS

SAS Institute Inc., Cary, North Carolina

ABSTRACT

Version 5 of JMP® software for Macintosh and Windows is just emerging this spring, and this presentation covers the new features.

1 WHAT IS JMP?

JMP is a smaller sibling to SAS®, aimed at scientists, engineers, and other researchers who need to analyze data. JMP is to SAS like a spreadsheet is to a database, smaller and geared to interactive desktop uses, but able to merge into the larger enterprise easily. One of the most prominent uses of JMP is to design and analyze experiments. JMP has always been strongest in its graphical approach to analyzing data. There is a graph for almost every statistic, and most of the graphs are interactive.

JMP's largest user group consists of engineers and statistical support specialists in manufacturing, particularly in pharmaceuticals, semiconductors, chemicals, and consumer products. Often JMP is used in support of a Six Sigma or other quality improvement program. JMP is also heavily used at universities.

1.1 RECENT HISTORY

JMP was first released in 1989 and has been growing ever since, though it is still very small compared to SAS. There is just one unified product, though there is a student version, called JMPIN. The product was completely rewritten between version 3 and version 4, with version 4 taking a giant leap in the sophistication of the surface of the product, adding a scripting system, and completely redesigning the experimental design facility. For example, with Version 4 you could drag a statistical report into Word and the graphs would stay graphs, and the tables would stay as editable tables.

JMP also announced the first localized version of JMP, for Japan, the largest JMP market outside of the US. This version is now being updated for Version 5.

JMP Version 5 builds on the improved interface of version 4, adding new data access, statistics, graphics commands and drawing tools. JMP’s scripting language has been enhanced. The menu structure and tool bars are now fully customizable, and new drawing tools have been added to aid in presentation-style results. This means that JMP can also be used to help automate processes and to present meaningful results to a wide audience.

2 NEW JMP PLATFORMS

JMP V5.0 includes five new platforms:

PARTITION—a tree based recursive modeling tool,
NEURAL NET—simple neural network modeling with one hidden layer,

PLS—partial least squares,

DISCRIMINANT—now a separate platform that has features for stepwise selection, canonical plots and identification of rows by scoring profiles,

DIAGRAM for producing Ishikawa (fishbone charts) or cause and effect diagrams.

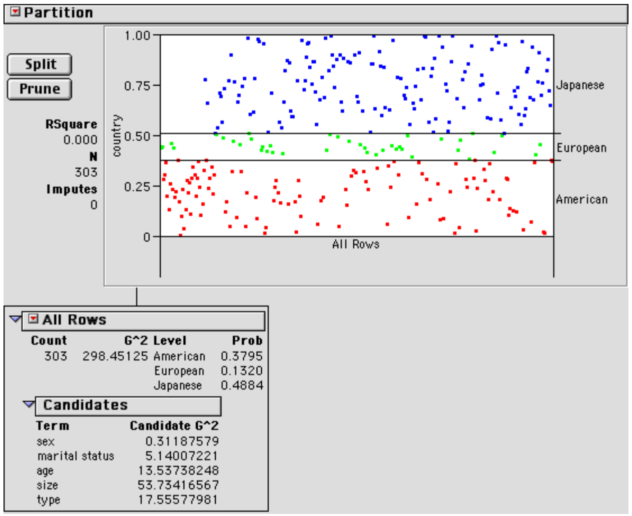
2.1 PARTITION

The Partition platform recursively partitions data creating a tree of partitions. Variations of this technique go by many names: decision trees, CART, CHAID, C4.5, C5, and others. The technique is often taught as a data mining technique, because it is good for exploring relationships without having a good prior model, it handles large problems easily, and the results are very interpretable.

The classic application is where you want to turn a data table of symptoms and diagnoses of a certain illness into a hierarchy of questions to ask new patients in order to make a quick initial diagnosis.

Partition features an innovative partitioned graph to show the results graphically. In contrast to data mining tools, the splitting is done purely interactively.

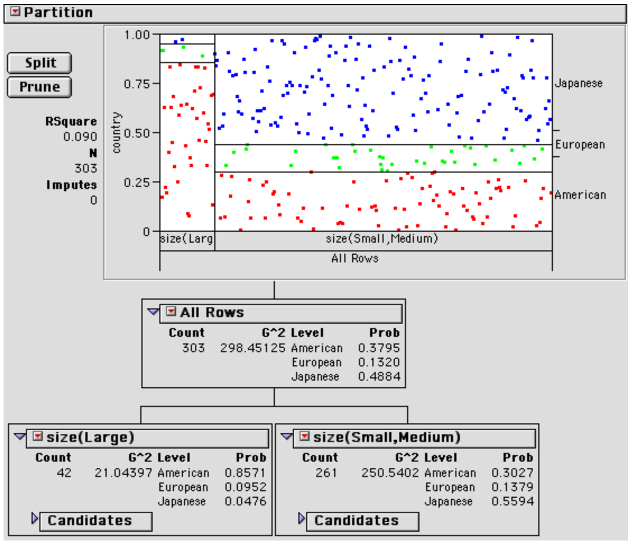
For example suppose that you want to find out what distinguishes the rates at which people buy Japanese, European, or American car brands. After specifying the columns of data, Partition starts with a graph where the rows are randomly ordered across a single population. The three response levels create vertical partitions representing the rates of the responses in the population. This is best seen in color with the points colored by response level.



```
All Rows
Count  G^2 Level  Prob
303  298.45125 American  0.3795
               European  0.1320
               Japanese  0.4884
Candidates
Term            Candidate G^2
sex             0.31187579
marital status  5.14007221
age             13.53738248
size            53.73416567
type            17.55577981
```

Figure 1. Carpoll Data's Initial Partition Report

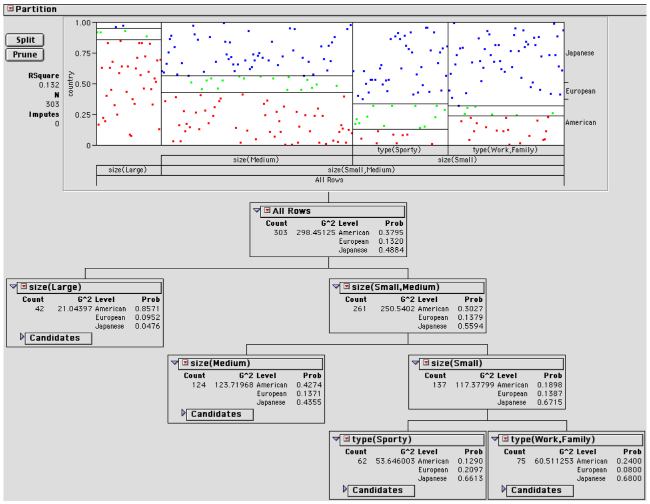
The user next interactively splits this group. The goal is to split the rows into two groups in which the rates across the response categories are more significantly different. It tests all the possible splits in each of the X columns and offers a Candidate report showing which column splits produce which test statistics (Chi-square or 2*entropy). Here you see that Size has the more significant split. Clicking the Split button produces this split of the population into two groups, one for “Large”, and one for “Small, Medium”.



```
All Rows
Count  G^2 Level  Prob
303  298.45125 American  0.3795
               European  0.1320
               Japanese  0.4884

size(Large)
Count  G^2 Level  Prob
42  21.04397 American  0.8571
             European  0.0952
             Japanese  0.0476
Candidates

size(Small,Medium)
Count  G^2 Level  Prob
261  250.5402 American  0.3027
              European  0.1379
              Japanese  0.5594
Candidates
```

Figure 2 Partition After First Split

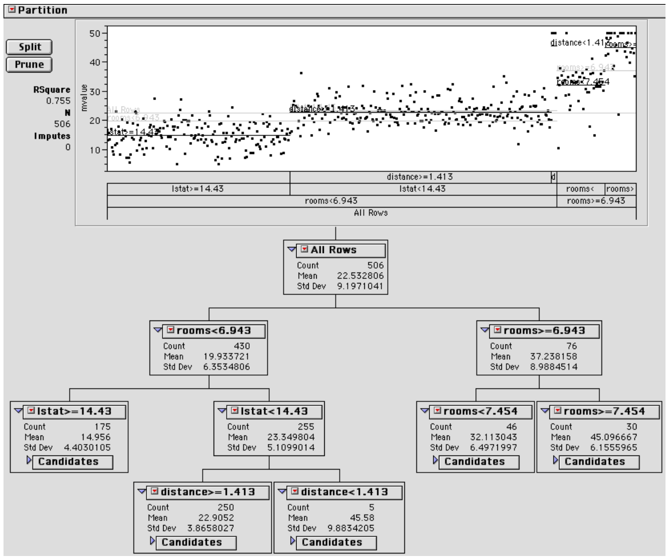
Splitting can continue interactively as far as you want.



```
All Rows
Count  G^2 Level  Prob
303  298.45125 American  0.3795
               European  0.1320
               Japanese  0.4884

size(Large)
Count  G^2 Level  Prob
42  21.04397 American  0.8571
             European  0.0952
             Japanese  0.0476
Candidates

size(Small,Medium)
Count  G^2 Level  Prob
261  250.5402 American  0.3027
              European  0.1379
              Japanese  0.5594

size(Medium)
Count  G^2 Level  Prob
124  123.71968 American  0.4274
               European  0.1371
               Japanese  0.4355
Candidates

size(Small)
Count  G^2 Level  Prob
137  117.37799 American  0.1898
               European  0.1387
               Japanese  0.6715

type(Sporty)
Count  G^2 Level  Prob
62  53.646003 American  0.1290
              European  0.2097
              Japanese  0.6613
Candidates

type(Work,Family)
Count  G^2 Level  Prob
75  60.511253 American  0.2400
              European  0.0800
              Japanese  0.6800
Candidates
```

Figure 3 Further Splits

If you use continuous data, the graph shows the points in relation to the mean for each terminal tree node whose group the point is in.



```
All Rows
Count    506
Mean     22.532806
Std Dev  9.1971041

rooms<6.943
Count    430
Mean     19.933721
Std Dev  6.3534806

rooms>=6.943
Count    76
Mean     37.238158
Std Dev  8.9884514

lstat>=14.43
Count    175
Mean     14.956
Std Dev  4.4030105
Candidates

lstat<14.43
Count    255
Mean     23.349804
Std Dev  5.1099014

rooms<7.454
Count    46
Mean     32.113043
Std Dev  6.4971997
Candidates

rooms>=7.454
Count    30
Mean     45.096667
Std Dev  6.1555965
Candidates

distance>=1.413
Count    250
Mean     22.9052
Std Dev  3.8658027
Candidates

distance<1.413
Count    5
Mean     45.58
Std Dev  9.8834205
Candidates
```

Figure 4 Boston Housing Data, continuous response

2.2 NEURAL NET

The Neural Net platform implements a standard type of neural network. Neural nets are used to predict one or more response variables from a flexible network of functions of input variables. Neural networks can be very good predictors without needing to know the functional form of the response surface. JMP fits these models like it fits nonlinear regressions. As is typical for Neural Nets, there are facilities to try many fits automatically, since many iterations lead only to local, rather than global optima. Also, there is a weight-decay facility to help prevent the model from being over-flexible and over fitted.

In this example, data on tiretreads is used; this is a multiple-Y response surface experiment described in Derringer and Suich (1980). There are four Y output response variables, and three X input factors. When this data is fit using regression, a quadratic response surface model is used, with many additional terms for squares and crossproducts. When it is fit with neural nets, the activation (sigmoid) functions and the hidden layers provide for the flexibility and possibility of interactions, so only the three input variables are specified. The platform can format a structure diagram of the net, here using three hidden nodes.

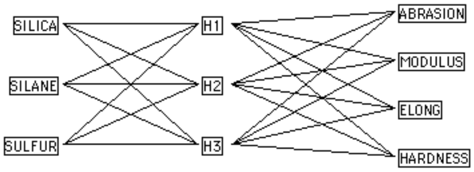
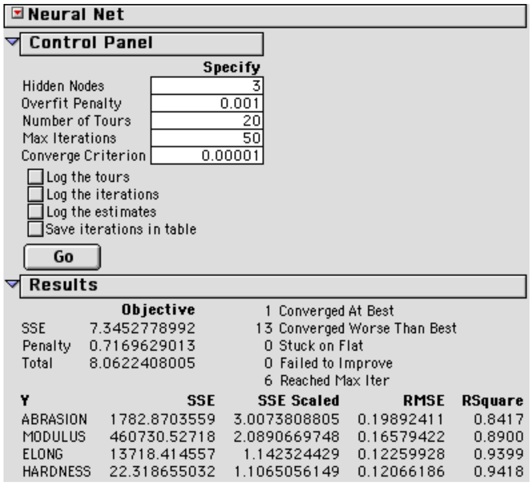


Figure 5 Neural Network Diagram

This yields a model with 28 parameters—far fewer than the 40 parameters of a complete quadratic response surface. After specifying the control values and clicking Go, the platform shows a summary of the iteration trials and displays the best fit obtained.



```
Neural Net
Control Panel
                          Specify
Hidden Nodes                    3
Overfit Penalty             0.001
Number of Tours                20
Max Iterations                 50
Converge Criterion        0.00001
□ Log the tours
□ Log the iterations
□ Log the estimates
□ Save iterations in table
Go
Results
            Objective          1 Converged At Best
SSE      7.3452778992         13 Converged Worse Than Best
Penalty  0.7169629013          0 Stuck on Flat
Total    8.0622408005          0 Failed to Improve
                               6 Reached Max Iter
Y                 SSE    SSE Scaled        RMSE  RSquare
ABRASION  1782.8703559  3.0073808805  0.19892411   0.8417
MODULUS   460730.52718  2.0890669748  0.16579422   0.8900
ELONG     13718.414557   1.142324429  0.12259928   0.9399
HARDNESS  22.318655032  1.1065056149  0.12066186   0.9418
```

Figure 6 Control Panel, and Results

A unique and valuable feature in JMP's Neural Net platform is the profiler, which shows slices of the response surface across each factor. This is the same facility that can be accessed from other platforms. The goal of the analysis might be to find a setting to optimize the desirability of the responses. To do this, select Desirability Functions in the drop-down menu of the profiler, then adjust the desirability profiles to maximize the desirability values.

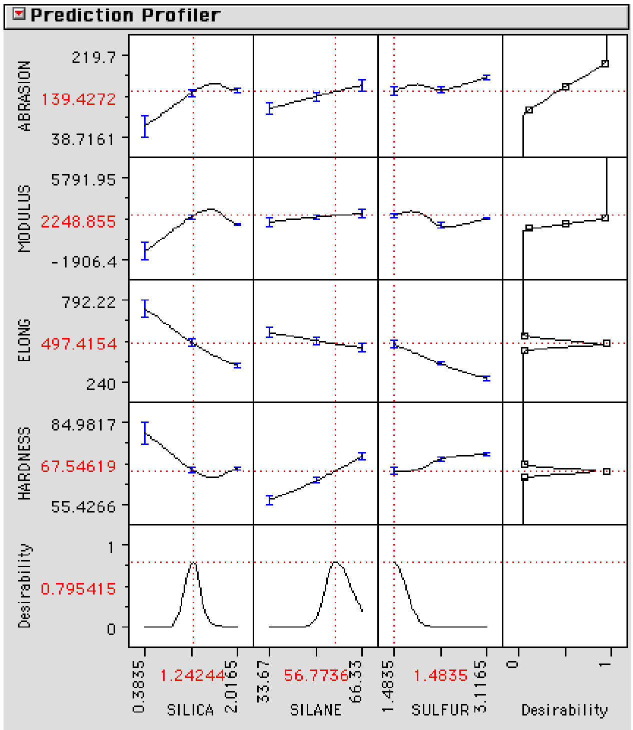


Figure 7 Prediction Profiler with Desirability

2.3 PLS - PARTIAL LEAST SQUARES

The PLS platform fits relationships between large numbers of variables when there are only a small number of observations. For example, five ingredients are studied in the relationship to a spectrum measured at 401 frequencies over 30 observations. First a set of graphs appears showing that six pairs of linear combinations seem worth looking at.

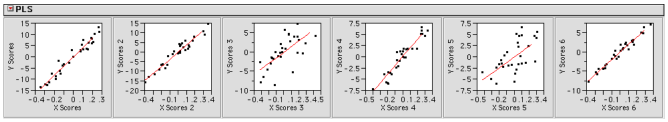
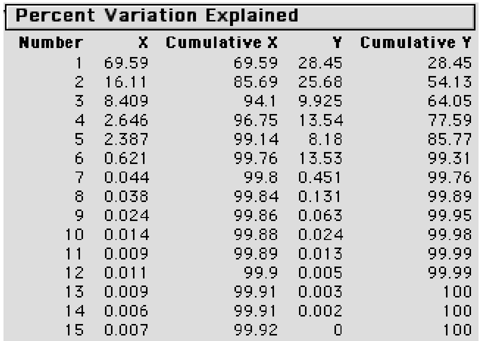


Figure 8 PLS Scores Plot



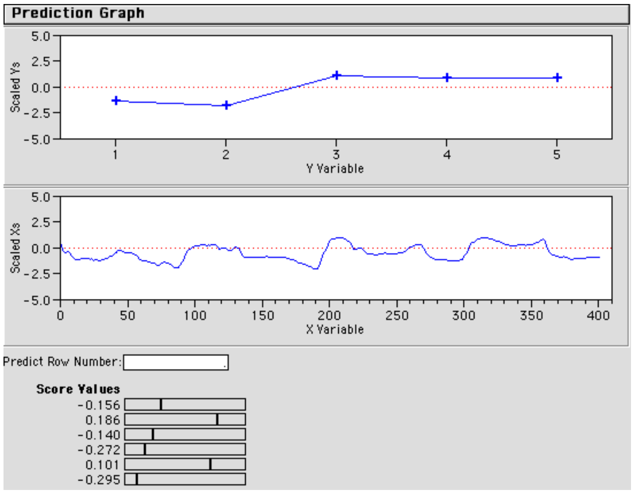
Percent Variation Explained

Number	X	Cumulative X	Y	Cumulative Y
1	69.59	69.59	28.45	28.45
2	16.11	85.69	25.68	54.13
3	8.409	94.1	9.925	64.05
4	2.646	96.75	13.54	77.59
5	2.387	99.14	8.18	85.77
6	0.621	99.76	13.53	99.31
7	0.044	99.8	0.451	99.76
8	0.038	99.84	0.131	99.89
9	0.024	99.86	0.063	99.95
10	0.014	99.88	0.024	99.98
11	0.009	99.89	0.013	99.99
12	0.011	99.9	0.005	99.99
13	0.009	99.91	0.003	100
14	0.006	99.91	0.002	100
15	0.007	99.92	0	100

Figure 9 Percent Variance Explained Report

Now you need to specify how many of these components or latent vectors to use in the final fit. A dialog helps you choose, suggesting a default choice.

Then a profiler shows the fitted Y's and X's as a function of the component values as specified in slider controls. You can enter an observation to see profiled, and then move the sliders to see the fitted relationships move.



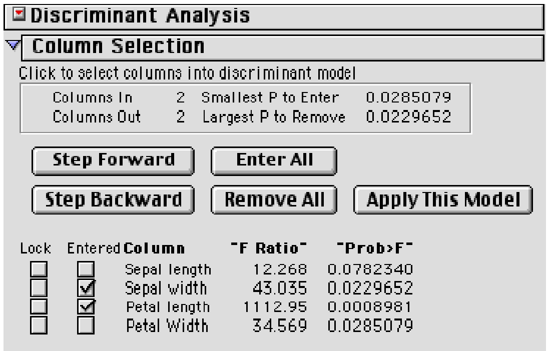
Predict Row Number:

Score Values
-0.156
0.186
-0.140
-0.272
0.101
-0.295

Figure 10 Prediction Graph and Sliders

2.4 DISCRIMINANT

Discriminant Analysis was available in previous JMP versions through the Manova fitting personality, but it didn't come with a full set of diagnostic plots. The new platform offers these, and also an interactive stepwise selection facility similar to the facility in JMP's stepwise regression.



Discriminant Analysis

Column Selection

Click to select columns into discriminant model

```
Columns In    2   Smallest P to Enter   0.0285079
Columns Out   2   Largest P to Remove   0.0229652
```

Lock	Entered	Column	"F Ratio"	"Prob>F"
☐	☐	Sepal length	12.268	0.0782340
☐	☑	Sepal width	43.035	0.0229652
☐	☑	Petal length	1112.95	0.0008981
☐	☐	Petal Width	34.569	0.0285079

Figure 11 Stepwise Control Panel

The canonical plot shows the points and means (centroids) in the (canonical) space that best separates the groups.

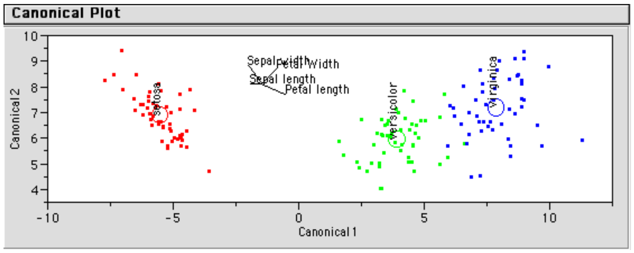
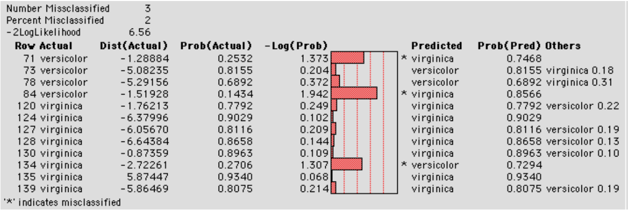


Figure 12 Canonical Plot



```
Number Missclassified        3
Percent Misclassified        2
-2LogLikelihood           6.56
 Row Actual      Dist(Actual)  Prob(Actual)  -Log(Prob)      Predicted     Prob(Pred) Others
  71 versicolor    -1.28884       0.2532       1.373    * virginica       0.7468
  73 versicolor    -5.08235       0.8155       0.204      versicolor      0.8155 virginica 0.18
  78 versicolor    -5.29156       0.6892       0.372      versicolor      0.6892 virginica 0.31
  84 versicolor    -1.51928       0.1434       1.942    * virginica       0.8566
 120 virginica     -1.76213       0.7792       0.249      virginica       0.7792 versicolor 0.22
 124 virginica     -6.37996       0.9029       0.102      virginica       0.9029
 127 virginica     -6.05670       0.8116       0.209      virginica       0.8116 versicolor 0.19
 128 virginica     -6.64384       0.8658       0.144      virginica       0.8658 versicolor 0.13
 130 virginica     -0.87359       0.8963       0.109      virginica       0.8963 versicolor 0.10
 134 virginica     -2.72261       0.2706       1.307    * versicolor      0.7294
 135 virginica      5.87447       0.9340       0.068      virginica       0.9340
 139 virginica     -5.86469       0.8075       0.214      virginica       0.8075 versicolor 0.19
'*' indicates misclassified
```

Figure 13 Scoring Report

2.5 DIAGRAM

The Diagram platform's is used to construct *Ishikawa chart*s, also called *fishbone chart*s, or *cause-and-effect diagrams*. These charts are useful to organize the sources (causes) of a problem (effect), perhaps for brainstorming, or as a preliminary analysis to identify variables in preparation for further experimentation.

Using sample data (Montgomery, 1996) which concerns defects in a circuit board we wish to produce the Ishikawa Cause and Effect diagram in Figure 19.

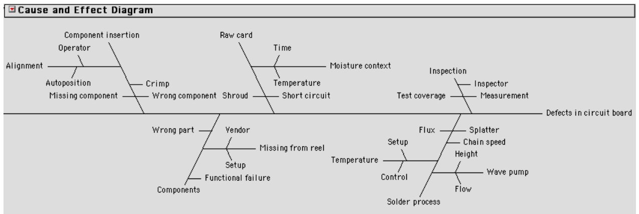


Figure 14 Ishikawa Cause and Effect Diagram

Charts can also be built interactively. Right click on any node in the chart to bring up a context menu that allows a chart to be built piece-by-piece.

3 NEW FOR EXPERIMENTAL DESIGN

I-Optimal Designs Version 4 introduced an exceptionally powerful custom experimental designer that made it easy to produce D-Optimal experimental designs for the specific needs of an experimenter. However, often instead of wanting to minimize the variance of the parameter estimates (D-Optimal), you should want to minimize the variance of the predictions across the design space (I-Optimal). Now I-Optimal becomes the default in cases where you have second order continuous terms in a model, which is usually when you are doing response surface optimization.

Bayesian D-Optimal Often you want to estimate more effects that you can afford runs for. Though you can't optimize a D-Optimal design directly for them, there is a new feature in JMP Custom Designer where you specify extra effects to be estimable if possible. The optimizer constructs designs which minimize; the correlation between the extra terms, even though they may not all estimable. If makes as many estimable as possible.

Supersaturated This feature also makes JMP a good tool for constructing supersaturated designs. These are designs which have more effects than runs, however when you have only a few strong effects, it can pick them out of the effect population for screening situations.

Other DOE improvements include:

- For certain situations with covariate factors, the row exchange algorithm using a candidate set is available in the custom designer.
- Desirability Functions are much improved. New desirability functions have been implemented for both maximizing targets and minimizing targets.
- JMP now (optionally) shows a table with D, G, and *A* efficiencies for a custom design.
- The design matrix can be stored in a data table.
- Random starts have been implemented when searching for optimal designs.

4 DATA ACCESS

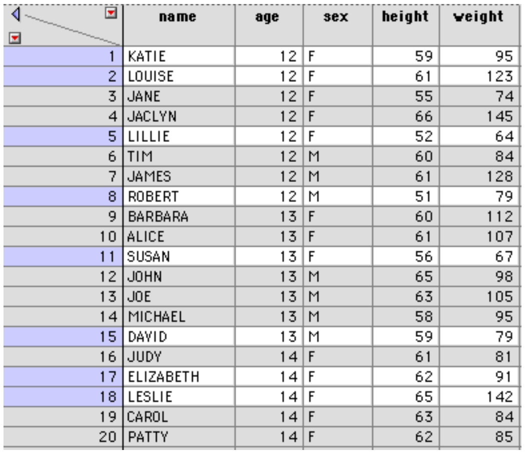
Internet Access (Windows Only): JMP can now get HTML, JMP or text files from a http address. In addition, web pages can be browsed from within JMP using a built-in browser. A File>Internet open command makes access easy.

Native SAS Files (Windows Only): When opening SAS data sets, there is now an option to use the SAS variable names for the column names in the resulting JMP data table.

Column Selection: You can select which columns are imported from other JMP or SAS Files.

5 DATA TABLE

- The cells are now highlighted in accordance with which cells would be copied with a copy command.
- You can select a random selection of rows, either a random number or random proportion of rows.



	name	age	sex	height	weight
1	KATIE	12	F	59	95
2	LOUISE	12	F	61	123
3	JANE	12	F	55	74
4	JACLYN	12	F	66	145
5	LILLIE	12	F	52	64
6	TIM	12	M	60	84
7	JAMES	12	M	61	128
8	ROBERT	12	M	51	79
9	BARBARA	13	F	60	112
10	ALICE	13	F	61	107
11	SUSAN	13	F	56	67
12	JOHN	13	M	65	98
13	JOE	13	M	63	105
14	MICHAEL	13	M	58	95
15	DAVID	13	M	59	79
16	JUDY	14	F	61	81
17	ELIZABETH	14	F	62	91
18	LESLIE	14	F	65	142
19	CAROL	14	F	63	84
20	PATTY	14	F	62	85

Figure 15 Random selection showing new highlighting

6 PLATFORM IMPROVEMENTS

Distribution

- The Alt-clicking a histogram bar will narrow a selection, for use in *And*-queries.
- The P-value and Power animations, accessible after testing a mean, are enhanced to allow sample size and alpha levels to be changed.
- Tolerance Intervals can now be computed

Capability Analysis

Capability analysis can now use four different options for estimating :

- Long-term
- Specified
- Short-term grouped by fixed subgroup size
- Short Term, grouped by column.

Fit Y By X

- The Bivariate platform now provides a way to turn off polynomial centering, using the Fit Special command.
- The Version 3 paired t-test is accessible by holding the Shift key down when opening the Bivariate Platform menu.
- The Fit Spline command now results in a report with an attached slider, used to vary the stiffness of the fitted curve.
- The text box for the equation in bivariate reports is now editable.
- A CDF plot has been added as an option in one-way analysis of variance.
- When you do multiple comparisons on means in the Oneway or Fit Model platforms, a connecting letter report (similar to the old reports in SAS PROC ANOVA) shows.

Fit Model

- In Stepwise, a new All Possible Regressions command shows the results of running all possible subsets of a linear model.
- CV has been added as a hidden column in REML and EMS results.

Clustering: Several new commands have been added including Ordering, Colormap, Geometric X Scale, Save Cluster Hierarchy, Orientation. The algorithm has been speeded up considerably so that it is practical to do 5000 rows, more if you have lots of memory and patience. This makes it useful for DNA microarray analysis.

In the k-means clustering, Biplots and Self Organizing Maps (SOMS) have been added.

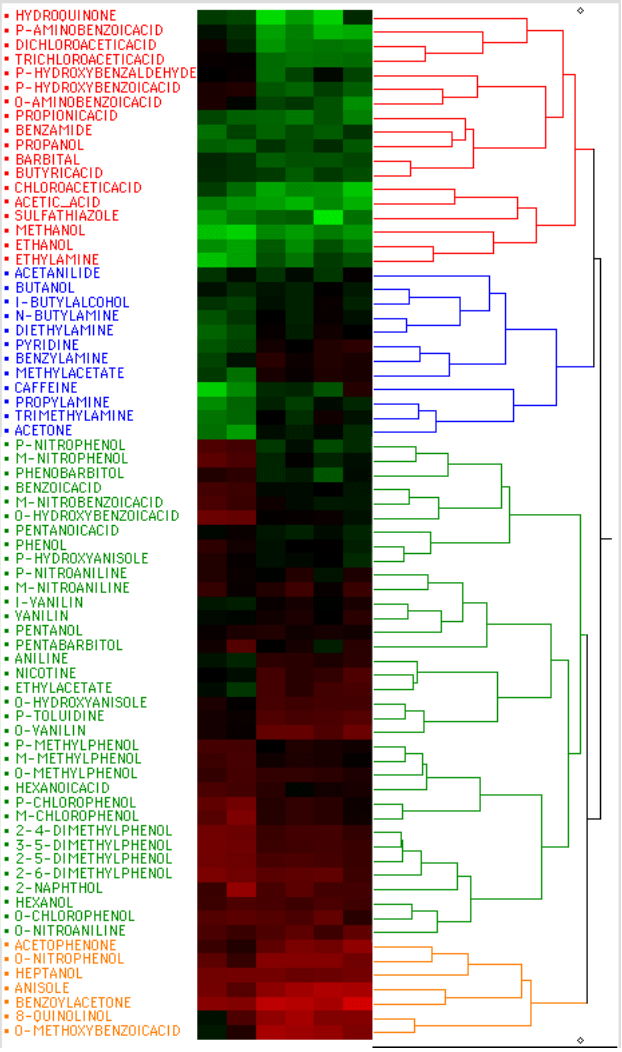


Figure 16 Hierarchical clustering with color map

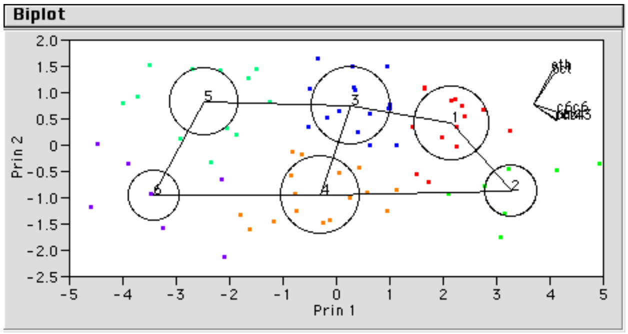


Figure 17 The SOM grid in principal component space

Survival and Reliability Analysis:

- Interval censoring is supported both in the univariate fits, and for regression. Turnbull estimates are used in place of Kaplan-Meier estimates in this case.
- Failure Plots have been added to univariate analyses.
- The Fitted Distribution Plots command, in conjunction with the fitted distributions, shows Survival,Density, and Hazard plots corresponding to the fitted distributions.
- JMP can now constrain the values of the Beta (Weibull) and Sigma (LogNormal) parameters for use in Weibayes Analysis.
- The Proportional Hazards platform now supports a Frequency column.
- Parametric survival regression graphs with respect to the regressor, such as in an accelerated failure model. You can obtain specific quantiles or survival or failure probability estimates and their confidence intervals.

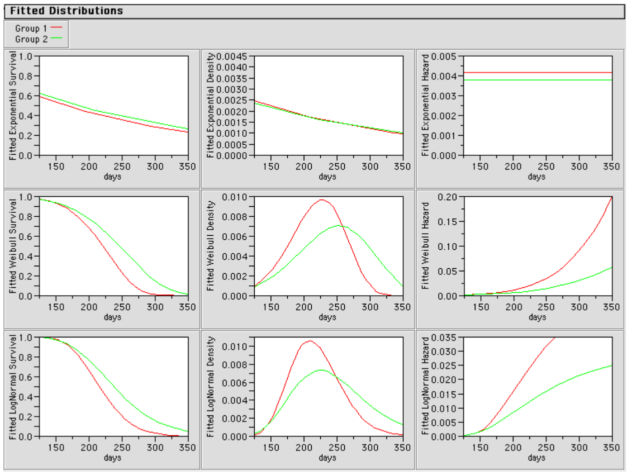
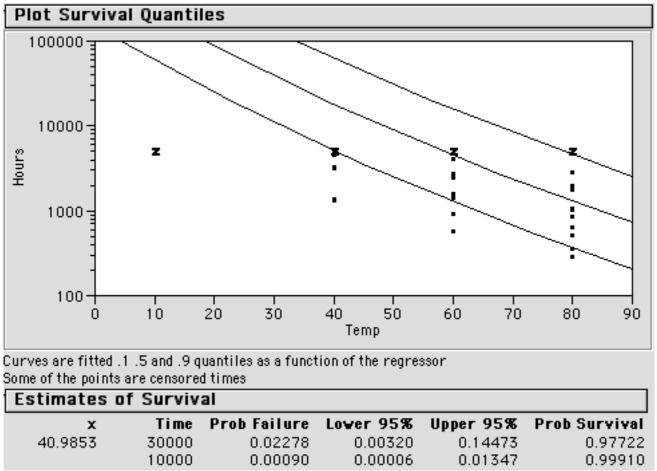


Figure 18. Fitted Distribution Plots



Curves are fitted .1 .5 and .9 quantiles as a function of the regressor
Some of the points are censored times

Estimates of Survival

x	Time	Prob Failure	Lower 95%	Upper 95%	Prob Survival
40.9853	30000	0.02278	0.00320	0.14473	0.97722
	10000	0.00090	0.00006	0.01347	0.99910

Figure 19. Accelerated Failure Regression Graph

7 USER INTERFACE

In addition to JMP's traditional annotation tool, JMP Version 5 has tools for drawing lines, ovals, rectangles, and polygons. These graphics can appear on reports, in journals, and in layouts.

The buttons and other controls in JMP have been updated to give them a more modern appearance, more consistent with Windows-XP and MacOS X.

On Windows, the way in which the user customizes menus and toolbars has been dramatically improved. At the bottom of the Edit menu, there is a new submenu named Customize. This brings up a drag-and-drop editor for easy customization of menus and toolbars to suit the audience, application or purpose.

8 OTHER IMPROVEMENTS

- The executable has been virtually halved in size. It now takes up approximately 6 megabytes of disk space.
- The help system was reworked to reduce image sizes and to include a more useful navigation structure.
- Formulas that calculated across the rows of a data table (using, for example, subscripted variables or the Lag function) stressed the formula dispatcher and dependency system for large data tables. This resulted in a marked decrease in performance. The formula dependency system has been rewritten completely and now performs quickly.
- ODBC is much faster for importing large database tables, including Excel files.
- The JSL Try function now intercepts errors better.
- Random numbers are now generated using the Mersenne-Twister technique [Matsumoto and Nishimura, 1998]. This technique has a period length of 2^{19937}-1 (as opposed to 2 31 -1 for the former generator). The new generators are verified to pass all the DIEHARD tests as documented in Marsaglia (1996). The routines are also in SAS
- JSL and the OLE Automation interface are a number of new features.

CONCLUSION

The JMP development team is pleased to announce the introduction of JMP Version 5 with many new platforms, features and enhancements. For further details consult our website http://www.jmpdiscovery.com, our brochures and whitepapers and forthcoming documentation.

REFERENCES

Derringer, D. and Suich, R. (1980), “Simultaneous Optimization of Several Response Variables,” *Journal of Quality Technology*, 12:4, 214–219.

Marsaglia, G. (1996) DIEHARD: A Battery of Tests of Randomness". http://stat.fsu.edu/~geo.

Matsumoto, M. and Nishimura, T. (1998)"Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator", *ACM Transactions on Modeling and Computer Simulation*, Vol. 8, No. 1, January 1998, pp 3--30.

Montgomery, D. C. (1996) *Introduction to Statistical Quality Control*, 3rd edition. New York: John Wiley.

ADDITIONAL CO-AUTHORS

Development: John Sall, Katherine Ng, Michael Hecht, Richard Potter, Brian Corcoran, Annie Zangi, Bradley Jones, Charles Soper, Craige Hales, Bob Hickey, Kevin Hardman, and Chris Gotwalt. *Documentation*: Ann Lehman, Lee Creighton. *Testing*: Nicole Jones, Jianfeng Ding, Jim Borek,and Kyoko Tidball. *Localization*: Erin Vang and SDL. *Japanese support:* Kyoko Takenaka and Noriki Inoue.

CONTACT INFORMATION

Contact the authors at SAS Institute Inc., Cary NC 27513

John.Sall@sas.com

Cathy.Maahs-Fladung@sas.com