Paper 357-2011

JMP® 9 and Interactive Statistical Discovery

John Sall, SAS Institute Inc., Cary, NC, USA

ABSTRACT

JMP 9 represents a major new revision of statistical discovery software from SAS® for engineers, scientists, and business analysts.

INTRODUCTION

Though there are many improvements throughout all the components in the JMP system, most of the excitement is focused on a few special features:

- Better scripting and a plugin system to make using scripts easy
- Improved presentation graphics
- More powerful interfaces to Excel, R, and SPSS as well as SAS
- A JMP Pro version that adds more data mining fits
- A new look on Windows

EXTENDING JMP THROUGH SCRIPTING

Suppose that you need to do an analysis of hyperspectral data. Our eyes and most of our cameras see in only three colors, but hyperspectral cameras can see in many colors, or wavelengths, including infrared and ultraviolet. Even the Landsat satellites send back images in seven wavelengths. Using the extra wavelengths helps discriminate land cover. Let's analyze an extreme example, Aviris data from NASA, which has 240 wavelengths. You can download some public Aviris data from a NASA website containing a number of 614-by-512 pixel images in 240 wavelengths. The script to do this is not hard, using some new features. Here is the housekeeping code to choose a directory and get the names of the rfl files.

```
NamesDefaultToHere(1);
defaultDir = "/C:/Users/sall/Documents/JMP9 Demo/Hyperspectral Data/";
inDir  = PickDirectory("Navigate to directory with AVIRIS rfl files",
        defaultDir,showEdit(1));
outDir = PickDirectory("Navigate to directory for resulting jpg PCA images",
        defaultDir,showEdit(1));
filelist = filesInDirectory(inDir);
ni = nItems(filelist);
rflFiles = {};
for(i=1,i<=ni,i++,if(EndsWith(fileList[i],".rfl"),InsertInto(rflFiles,fileList[i])));
   nRFL = NItems(rflFiles);
```

The first command, NamesDefaultToHere(1), says that all the names in this script will be kept in the local environment so that different scripts won't interfere with each other. This is one of the most important enhancements for JMP 9.

Now we sequence through all the files we found. Each one is brought in as one large binary chunk, a 'blob'—binary large object, using a new option in JMP 9. We figure out how many lines are in the image.

```
for(iFile=1,iFile<=nRFL,iFile++,
  file = rflFiles[iFile];
  fBlob = loadTextFile(inDir||file,blob);
  sz = length(fBlob);
  nchannel = 224;      // number of colors/wavelengths
  nsample = 614;       // number of pixels across each line
  nLines = 512;        // number of lines in the image
  rawData = Blob To Matrix( fBlob, "int", 2, "big", nchannel ); fBlob = 0;
  nLines = (sz/2)/(nsample*nchannel); // sometimes it's not 512
  n = nsample*nLines;  // number a data table rows
```

Next we standardize the data, taking care not to divide by zero for constant columns, and calculate the Correlation matrix. The Correlation function is new, and it is very fast, using multithreading.

```
stdv = vstd(rawData); stdv += stdv==0; // don't divide by zero
stdData = (rawData-J(n,1,1)*vmean(rawData)):/(J(n,1,1)*stdv); //standardize
rawData = 0;
corr = Correlation(stdData);
```

Next, we calculate the eigen decomposition of the correlation matrix and use the eigenvectors to score the first four principal components, with the scores reshaped back to the dimensions of the image.

```
{m,e} = eigen(corr); show(m[1::10,0]);
pc1v = shape(stdData*e[0,1],nLines,nsample)`;
pc2v = shape(stdData*e[0,2],nLines,nsample)`;
pc3v = shape(stdData*e[0,3],nLines,nsample)`;
pc4v = shape(stdData*e[0,4],nLines,nsample)`;
```

Now we scale the results so that we can colorize by values from 0 to 1.

```
pc1s = (.5+pc1v/(2*max(pc1v))):*(pc1v>0)+(.5-pc1v/(2*min(pc1v))):*(pc1v<0);
pc2s = (.5+pc2v/(2*max(pc2v))):*(pc2v>0)+(.5-pc2v/(2*min(pc2v))):*(pc2v<0);
pc3s = (.5+pc3v/(2*max(pc3v))):*(pc3v>0)+(.5-pc3v/(2*min(pc3v))):*(pc3v<0);
pc4s = (.5+pc4v/(2*max(pc4v))):*(pc4v>0)+(.5-pc4v/(2*min(pc4v))):*(pc4v<0);
stdData = u = v = 0;  // free memory
```

Now we convert to the principal images using default heatmap colorization: blue for low, to gray for medium, to red for high. Image objects and functions are new in JMP 9.

```
imagepc1 = NewImage(HeatColor(pc1s));
imagepc2 = NewImage(HeatColor(pc2s));
imagepc3 = NewImage(HeatColor(pc3s));
imagepc4 = NewImage(HeatColor(pc4s));
```

The images are saved on disk.

```
imagepc1<<SaveImage(outDir||rflFiles[iFile]||"_PC1.jpg");
imagepc2<<SaveImage(outDir||rflFiles[iFile]||"_PC2.jpg");
imagepc3<<SaveImage(outDir||rflFiles[iFile]||"_PC3.jpg");
imagepc4<<SaveImage(outDir||rflFiles[iFile]||"_PC4.jpg");
```

In addition to the principal images, we also create images of the eigenvector loadings so that we can make a legend across the top showing which wavelengths are contributing to each principal image.

```
imageEig1 = NewImage(DirectProduct(HeatColor(e[0,1]`/2+.5),J(12,2,1)));
imageEig2 = NewImage(DirectProduct(HeatColor(e[0,2]`/2+.5),J(12,2,1)));
imageEig3 = NewImage(DirectProduct(HeatColor(e[0,3]`/2+.5),J(12,2,1)));
imageEig4 = NewImage(DirectProduct(HeatColor(e[0,4]`/2+.5),J(12,2,1)));
```

Now we arrange the images into a presentation and show it:

```
    NewWindow(rflFiles[iFile]||" PCA",
            LineupBox(ncol(2),spacing(3),
                    imageEig1,imageEig2,
                    imagepc1,imagepc2,
                    imageEig3,imageEig4,
                    imagepc3,imagepc4));
```

This is a loop, which we need to close with a parenthesis.

```
);
```

If you run this script, you will get a window for each hyperspectral ‘rfl‘ file, looking like the images shown in Figure 1.

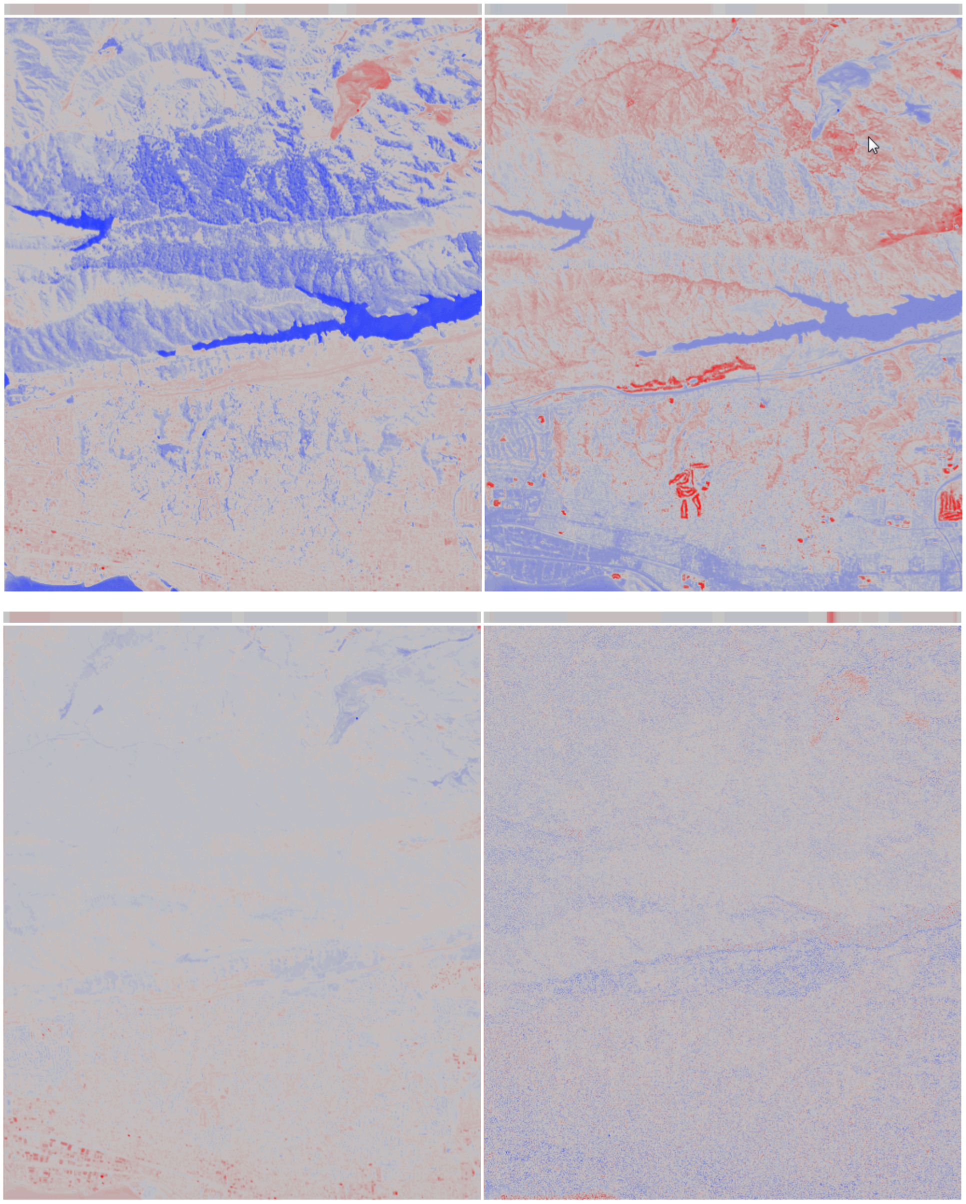
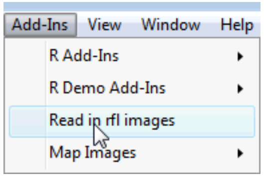


Figure 1. Images Created from Principal Components Computed from .rfl Hyperspectral Data

Notice that the first principal image concentrates on general intensity (the loading legend shows reddish values across the spectrum). The second principal image compares red wavelengths and blue (the legend loads positively in the lower middle wavelengths and negatively on the higher ones). The third principal image concentrates on infrared (legend positive at the low wavelengths). The fouth principal image loads in a very narrow band in the blue range of the spectrum, and the image looks rougher than the others. Each image emphasizes different features in the scene.

Considering that this is a 240 column by 314,368 row problem, it is amazing that it only takes about 30 seconds on my 8-core laptop to do this and everything else in this script per image.

Now consider that you want to enable this script to be easy to use for yourself. You install a menu item to invoke it.



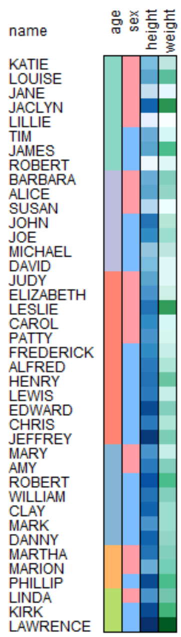
Also, you want it easy to install for many users. JMP 9 now supports 'add-ins' which zip together a script with some information about how to install it in the menus. This add-in file can now be put on a website. Any user can download it and drag it into JMP, and it automatically installs with a menu item to invoke it.

EXPLORING DATA

Exploring data is JMP's special talent.

COLOR JMP DATA TABLE CELLS

One of the graphs in JMP is a cellplot, sometimes called a heatmap as shown on the left in Figure 2, which shows the data with colored rectangles. With JMP 9, the spreadsheet itself can be colorized, either by value across a column or individually per cell.



	name	age	sex	height	weight	mean weight by age
1	KATIE	12	F	59	95	99.000
2	LOUISE	12	F	61	123	99.000
3	JANE	12	F	55	74	99.000
4	JACLYN	12	F	66	145	99.000
5	LILLIE	12	F	52	64	99.000
6	TIM	12	M	60	84	99.000
7	JAMES	12	M	61	128	99.000
8	ROBERT	12	M	51	79	99.000
9	BARBARA	13	F	60	112	94.714
10	ALICE	13	F	61	107	94.714
11	SUSAN	13	F	56	67	94.714
12	JOHN	13	M	65	98	94.714
13	JOE	13	M	63	105	94.714
14	MICHAEL	13	M	58	95	94.714
15	DAVID	13	M	59	79	94.714
16	JUDY	14	F	61	81	100.833
17	ELIZABETH	14	F	62	91	100.833
18	LESLIE	14	F	65	142	100.833
19	CAROL	14	F	63	84	100.833
20	PATTY	14	F	62	85	100.833
21	FREDERICK	14	M	63	93	100.833
22	ALFRED	14	M	64	99	100.833
23	HENRY	14	M	65	119	100.833
24	LEWIS	14	M	64	92	100.833
25	EDWARD	14	M	68	112	100.833
26	CHRIS	14	M	64	99	100.833
27	JEFFREY	14	M	69	113	100.833
28	MARY	15	F	62	92	108.286
29	AMY	15	F	64	112	108.286
30	ROBERT	15	M	67	128	108.286
31	WILLIAM	15	M	65	111	108.286

Figure 2. Color using a Cell Plot (left) and JMP Data Table Cells (right)

FEATURES IN GRAPH BUILDER

Graph Builder has many new skills. In the Graph Builder example on crime data (Figure 3 and Figure 4), I can click on a column name 'State' which identifies U.S. States, drag it to the new Shape drop zone, and it brings up a map and sets the graph coordinates to latitude by longitude. Graph Builder has looked through all the maps it has to find out which supports the values in the column.

Then I drag the column name 'Violent Rate' to the Color drop zone and the states are colorized, with red for high and blue for low.

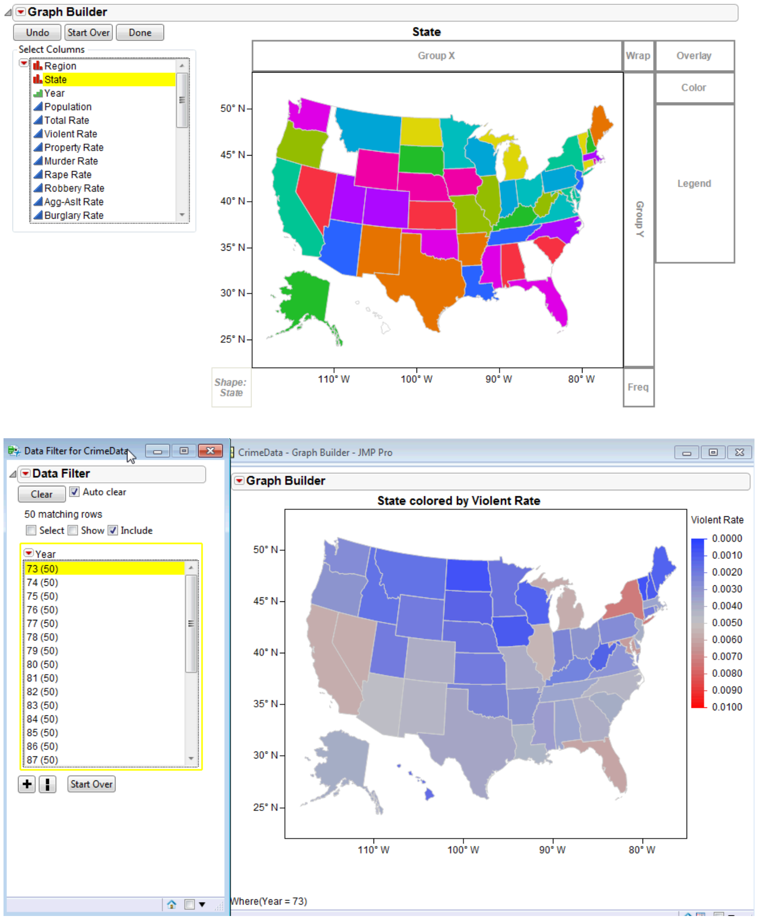


Figure 3. Color Map Data in Graph Builder by Crime Rates

Here is the crime picture for 1991, the height of the 'crack' epidemic, then in 1999 when violent crime had calmed down. This can be played like a movie, letting you watch violent-crime rate patterns both geographically and across time.

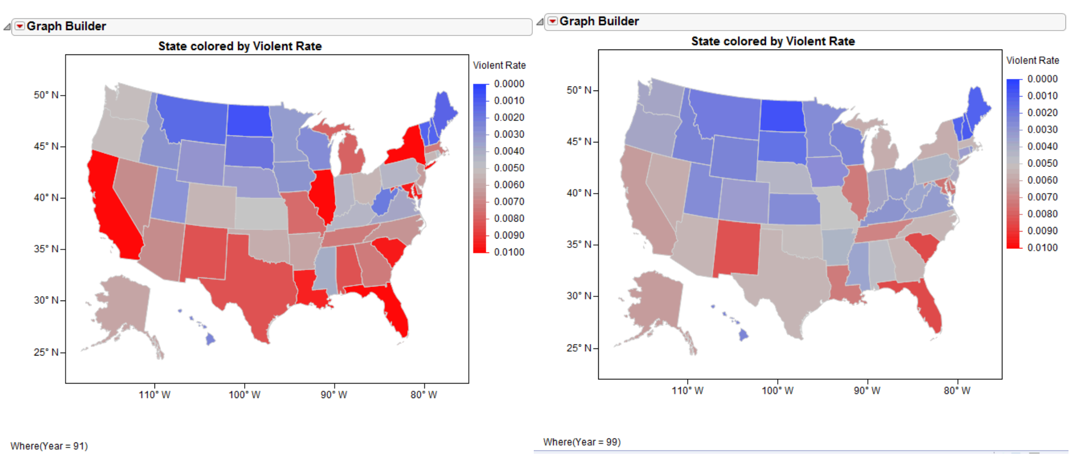


Figure 4. Look at Changing Crime Rates over Time

EXPLORING MULTIVARIATE DATA

Suppose you need to characterize the multivariate distribution of several variables, in this case 30,000 rows of FCS data. The KMeans platform has many features for doing this. The data may be noisy, so as a first step you find the distance of each point to the k^{th} nearest neighbor, as shown in Figure 5. Selecting the 900 or so points that are far from the 15^{th} nearest neighbor can be done by dragging a rectangle in the 15^{th} plot below. Notice that in JMP 9 the points that are not selected are faded, where in earlier versions of JMP the selected points were bigger. The fade selection mode is much better for when there are many points and you need to spot where the selected points are.

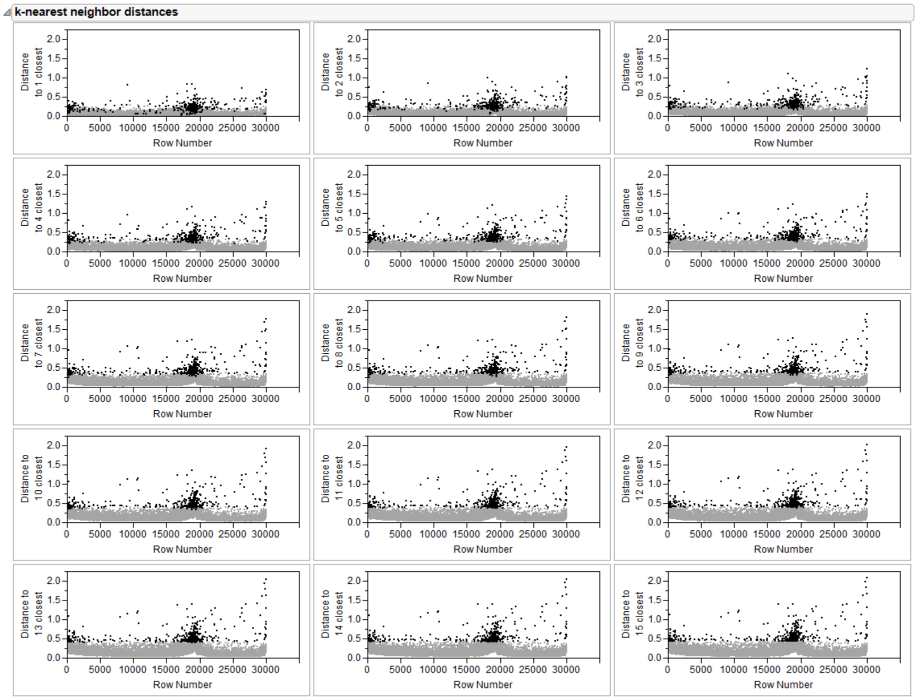


Figure 5. Nearest Neighbor Plots

Once these points are selected, you can see where they are in the space, using the 3D scatterplot, and the scatterplot matrix.

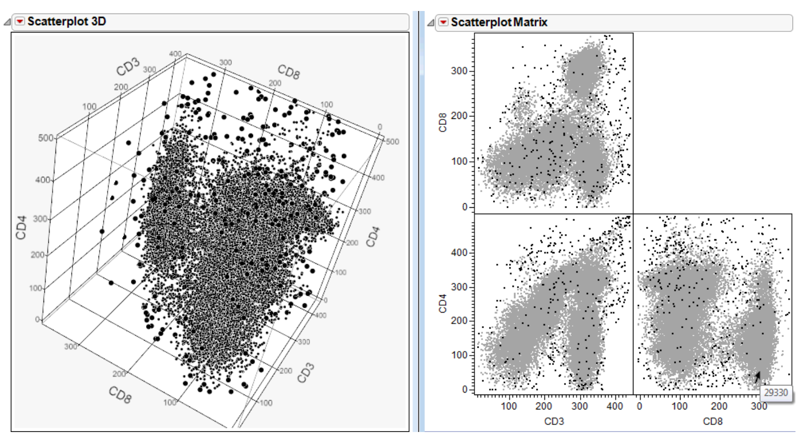


Figure 6. Noisy Points Showing in 3D Plot and Scatterplot Matrix

Now that we have concentrated the sample to points that are near other points, it is ready to find the structure of the distribution, using normal mixtures. JMP 9 has a much faster Normal Mixtures facility. There are a variety of ways to see the resulting mixture. There is the principal component score plot (left-hand plot in Figure 7) with shaded ellipses showing the normal contours and with circles indicating the portions in each cluster. The same idea can be done in 3D with semi-transparent ellipsoids (middle plot in Figure 7).

Parallel coordinates plots show the structure of each cluster, assigning points to the closest cluster in the Mahalanobis sense (right-hand plot in Figure 7).

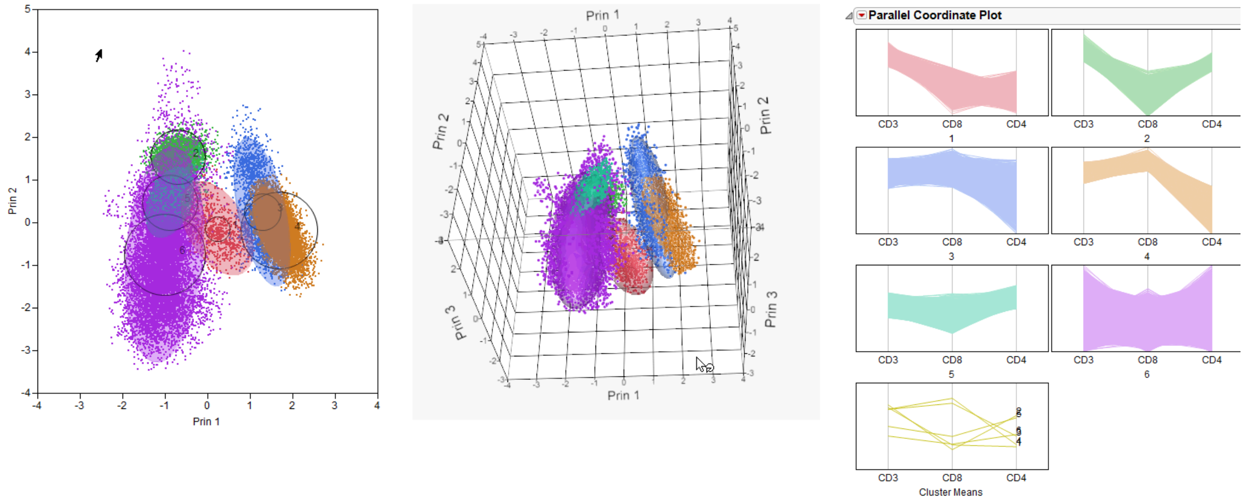


Figure 7. PCA Score Plot (left), 3D Plot with Ellipsoids (middle), Parallel Coordinate Plot (right)

Also, there is a feature to output the scatterplot matrix with ellipsoids projected into each component pair's space. Here we faded the points by changing the transparency setting.

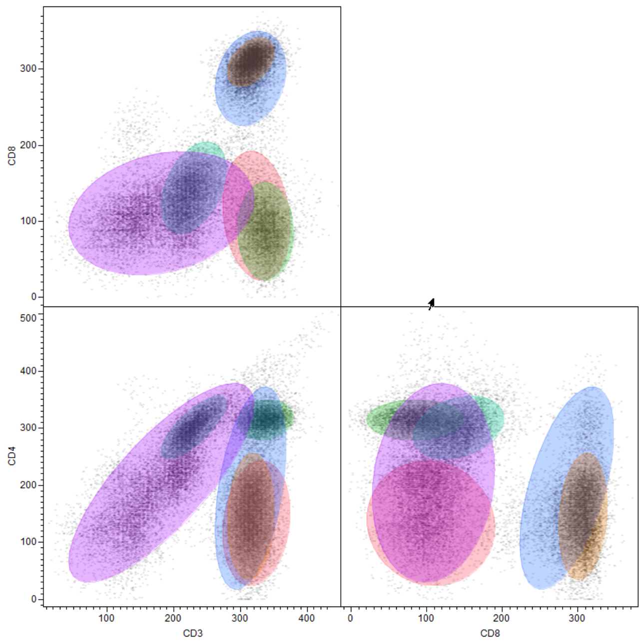


Figure 8. Scatterplot Matrix of Normal Mixture Clusters using Transparency Features

ALIAS-OPTIMAL EXPERIMENTAL DESIGN

JMP has a tradition of strong support for experimental design. Suppose that you have to make a screening experiment for six factors in 12 runs. The plain d-optimal design is efficient for the main effects, but if there is a big two-factor interaction lurking that is not estimable here, it will contaminate the main effects estimates due to the large number of 1/3 and -1/3 correlations confounding main effects with two-factor interactions (left picture in Figure 9). But if I change to an 'alias optimal' criterion, all these will be cleaned up—to zero in this situation, as seen in the all-zero (all-blue) section in the upper right and lower left (right picture in Figure 9).

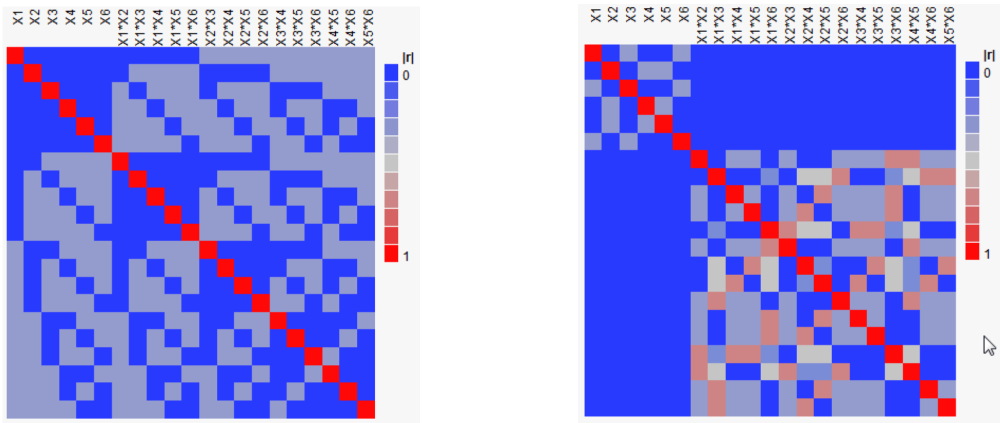


Figure 9. Correlation Maps for D-Optimal Design (left) and Alias Optimal Design (Right)

The alias-optimal design is trading some efficiency in the main effects with robustness to two-factor interactions. This is not always a trade you will want to make, but if you suspect strong two-factor interactions, but don't have the runs to make them estimable, this kind of design is worth considering.

MODELING

The Concrete Slump Test data, contributed to the UCI Machine Learning lab, is a very large-scale response surface experiment with three responses, 7 factors, and 103 runs. When expanded to all the quadratic terms, the model has 36 parameters for each response. To validate the predictive ability of fits, I created validation and test sets, identified by a Validation column in the data table. Using the Stepwise personality, if you fit all the terms for the first response (Slump), you get an Rsquare of 0.79 for the training data. However, the Rsquare for the validation data is very terrible (-1.415)—i.e. the validation data is fit far worse than just a simple mean (left-hand report in Figure 10). If you clear the terms and then click Go to step in terms until the crossvalidation Rsquare is maximized, then the resulting model validates much better with a validation Rsquare of 0.38. All these cross validation features are new to JMP 9.

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation
1403.1247	46	5.5229259	0.7940	0.6373	36	36	603.4735	628.613	-1.415	10.66509

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	-1533.7907	1	0	0.000	1
	✓	Cement	0.44508705	8	580.8067	2.380	0.0309
	✓	Slag	0.65997917	8	678.0405	2.779	0.01354
	✓	Fly ash	0.43233725	8	569.0628	2.332	0.03414
	✓	Water	1.80457186	8	510.1638	2.091	0.05618
	✓	SP	2.39753315	8	138.463	0.567	0.7989
	✓	Coarse Aggr	0.59524008	8	522.8882	2.143	0.05046
	✓	Fine Aggr	0.5938577	8	537.1101	2.201	0.04475
	✓	(Cement-229.894)*(Cement-229.894)	0.03796629	1	55.39926	1.816	0.18437
	✓	(Cement-229.894)*(Slag-77.9738)	0.11483328	1	60.74711	1.992	0.16491
	✓	(Slag-77.9738)*(Slag-77.9738)	0.08591237	1	65.45108	2.146	0.14977
	✓	(Cement-229.894)*(Fly ash-149.015)	0.07748005	1	54.98588	1.803	0.18598
	✓	(Slag-77.9738)*(Fly ash-149.015)	0.1171556	1	60.2541	1.975	0.1666
	✓	(Fly ash-149.015)*(Fly ash-149.015)	0.0386775	1	52.09633	1.708	0.19775
	✓	(Cement-229.894)*(Water-197.168)	0.25697904	1	61.77071	2.025	0.16147
	✓	(Slag-77.9738)*(Water-197.168)	0.38829935	1	67.66328	2.218	0.14321
	✓	(Fly ash-149.015)*(Water-197.168)	0.26057479	1	60.584	1.986	0.16547
	✓	(Water-197.168)*(Water-197.168)	0.41366541	1	61.73472	2.024	0.16159
	✓	(Cement-229.894)*(SP-8.53981)	0.31183423	1	55.52712	1.820	0.18387
	✓	(Slag-77.9738)*(SP-8.53981)	0.45112675	1	56.82853	1.863	0.17891
	✓	(Fly ash-149.015)*(SP-8.53981)	0.31519215	1	51.26471	1.681	0.2013
	✓	(Water-197.168)*(SP-8.53981)	0.985527	1	54.29124	1.780	0.18873
	✓	(SP-8.53981)*(SP-8.53981)	0.59565572	1	48.5381	1.591	0.2135
	✓	(Cement-229.894)*(Coarse Aggr-883.979)	0.09818998	1	58.39278	1.914	0.17316
	✓	(Slag-77.9738)*(Coarse Aggr-883.979)	0.14778863	1	63.48324	2.081	0.15589
	✓	(Fly ash-149.015)*(Coarse Aggr-883.979)	0.09982857	1	57.5688	1.887	0.17616
	✓	(Water-197.168)*(Coarse Aggr-883.979)	0.32366626	1	61.68809	2.022	0.16174
	✓	(SP-8.53981)*(Coarse Aggr-883.979)	0.39121749	1	55.10365	1.807	0.18552

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation
2288.9563	67	5.8449574	0.6640	0.5938	23.041079	15	546.064	576.2022	0.3888	5.365885

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	-354.27088	1	0	0.000	1
	✓	Cement	0.07444718	3	555.1947	5.417	0.00214
	✓	Slag	0.0678975	5	1097.075	6.422	6.25e-5
	✓	Fly ash	0.10284375	2	143.8339	2.105	0.12982
	✓	Water	0.60503766	4	792.265	5.798	0.00046
	✓	SP	0.90774862	2	206.755	3.026	0.05519
	✓	Coarse Aggr	0.13099398	2	215.2303	3.150	0.04926
	✓	Fine Aggr	0.13205391	1	64.43526	1.886	0.17422
		(Cement-229.894)*(Cement-229.894)	0	1	8.922592	0.258	0.613
	✓	(Cement-229.894)*(Slag-77.9738)	-0.0007732	1	524.1694	15.343	0.00021
	✓	(Slag-77.9738)*(Slag-77.9738)	-0.0014563	1	694.9113	20.341	2.68e-5
		(Cement-229.894)*(Fly ash-149.015)	0	1	12.01915	0.348	0.55704
	✓	(Slag-77.9738)*(Fly ash-149.015)	-0.0003485	1	72.86743	2.133	0.14884
		(Fly ash-149.015)*(Fly ash-149.015)	0	1	38.47533	1.128	0.292
		(Cement-229.894)*(Water-197.168)	0	1	23.56243	0.686	0.41035
	✓	(Slag-77.9738)*(Water-197.168)	0.00200571	1	321.9582	9.424	0.00309
		(Fly ash-149.015)*(Water-197.168)	0	1	0.036998	0.001	0.97404
	✓	(Water-197.168)*(Water-197.168)	-0.0086204	1	492.3669	14.412	0.00032
	✓	(Cement-229.894)*(SP-8.53981)	0.00517808	1	48.12945	1.409	0.23945
		(Slag-77.9738)*(SP-8.53981)	0	1	38.00015	1.114	0.29502
		(Fly ash-149.015)*(SP-8.53981)	0	1	3.775202	0.109	0.74229
		(Water-197.168)*(SP-8.53981)	0	1	38.26307	1.122	0.29334
		(SP-8.53981)*(SP-8.53981)	0	1	6.348772	0.184	0.66972
		(Cement-229.894)*(Coarse Aggr-883.979)	0	1	6.634793	0.192	0.6628
		(Slag-77.9738)*(Coarse Aggr-883.979)	0	1	3.899897	0.113	0.73822
		(Fly ash-149.015)*(Coarse Aggr-883.979)	0	1	44.04305	1.295	0.25927
	✓	(Water-197.168)*(Coarse Aggr-883.979)	-0.0012516	1	128.0872	3.749	0.05705
		(SP-8.53981)*(Coarse Aggr-883.979)	0	1	[illegible]	[illegible]	[illegible]

Figure 10. Stepwise Example of Cross Validation

In order to fit all the responses in one click on the Go button, I hold down the control key to broadcast that Go. Then I can control-click the Run Model button to run all the resulting models in the Standard Least Squares platform. The fits are assembled in the new Fit Group, allowing different models to support a single combined Profiler to explore the response surface across all the factors and responses (see Figure 11). In earlier versions of JMP, this would have involved many more steps and saving lots of prediction formulas as data table columns, and the result would not have included confidence intervals.

JMP 9 and Interactive Statistical Discovery, continued

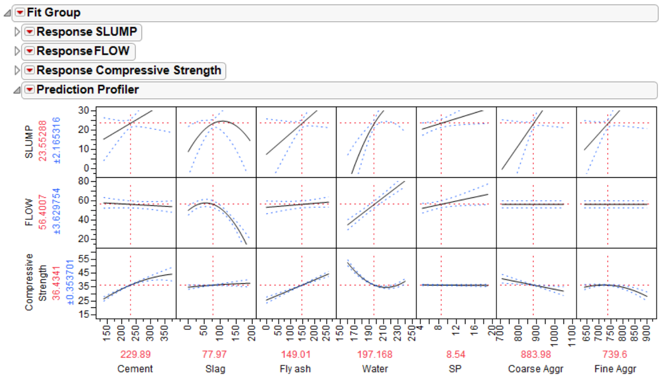


Figure 11. Profiler for Concrete Slump Test Data

SURVEY ANALYSIS

To see JMP at work analyzing surveys, I went to the website of a very large survey on Civic Engagement in America (http://www.bowlingalone.com/data.htm), the ‘Bowling Alone’ data. A convenient form of the data is in an SPSS save file. You can now easily download that and open it in JMP, and it preserves all the value labels.

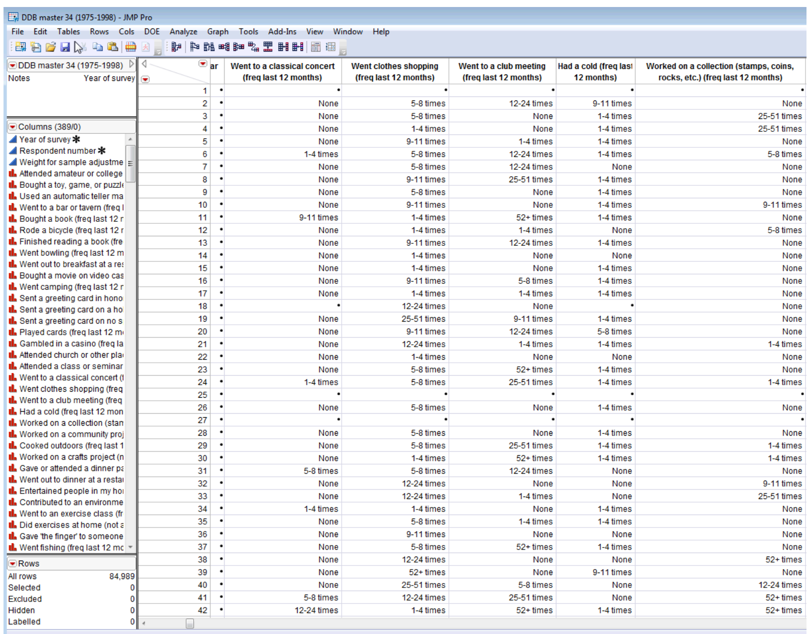


Figure 12. Partial Listing of the Bowling Alone Data

Notice that I opted to use the long names, which are very descriptive. However, these long names have a common phrase “(freq last 12 months)” repeated that will distort the compactness of the results (see Figure 12). It is easy to change them. I select all 104 variables for these frequency value and bring up the Search dialog to change them, clicking ‘Replace All‘. While all these columns are still highlighted, I can click the analysis type and change them to Ordinal so that their values will be charted with an ordinal theme. Also I check the column properties to verify the value labels from SPSS (see Figure 13).

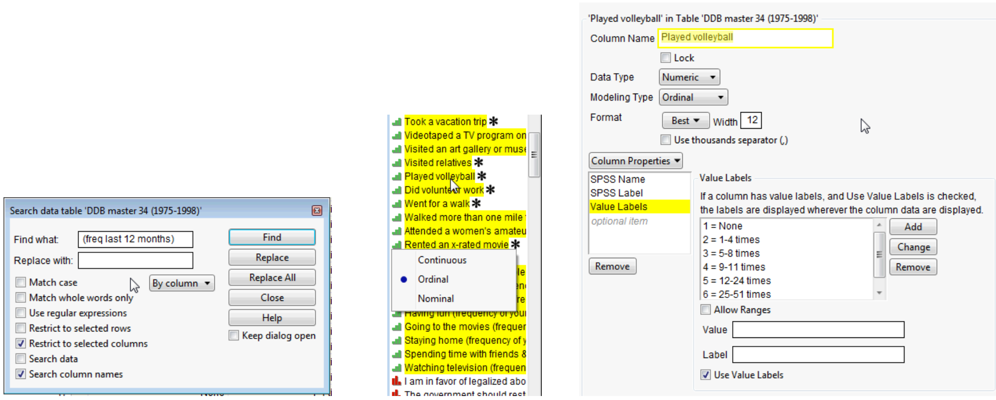


Figure 13. Search Dialog (left), Change Modeling Type (middle), Assign Value Labels (right).

Now I invoke the Categorical platform and click Separate Responses for the selected 104 columns, and then Year of Survey for my X grouping column.

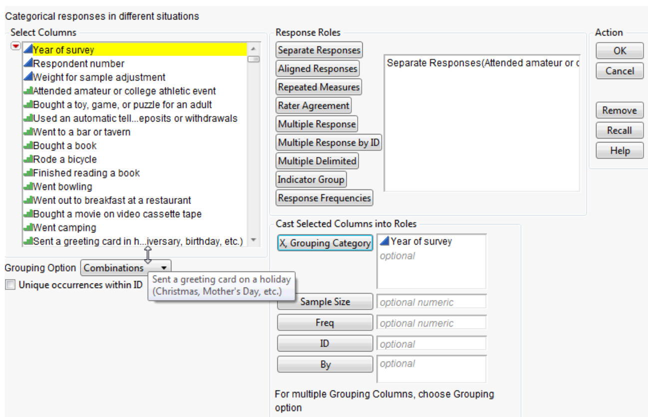
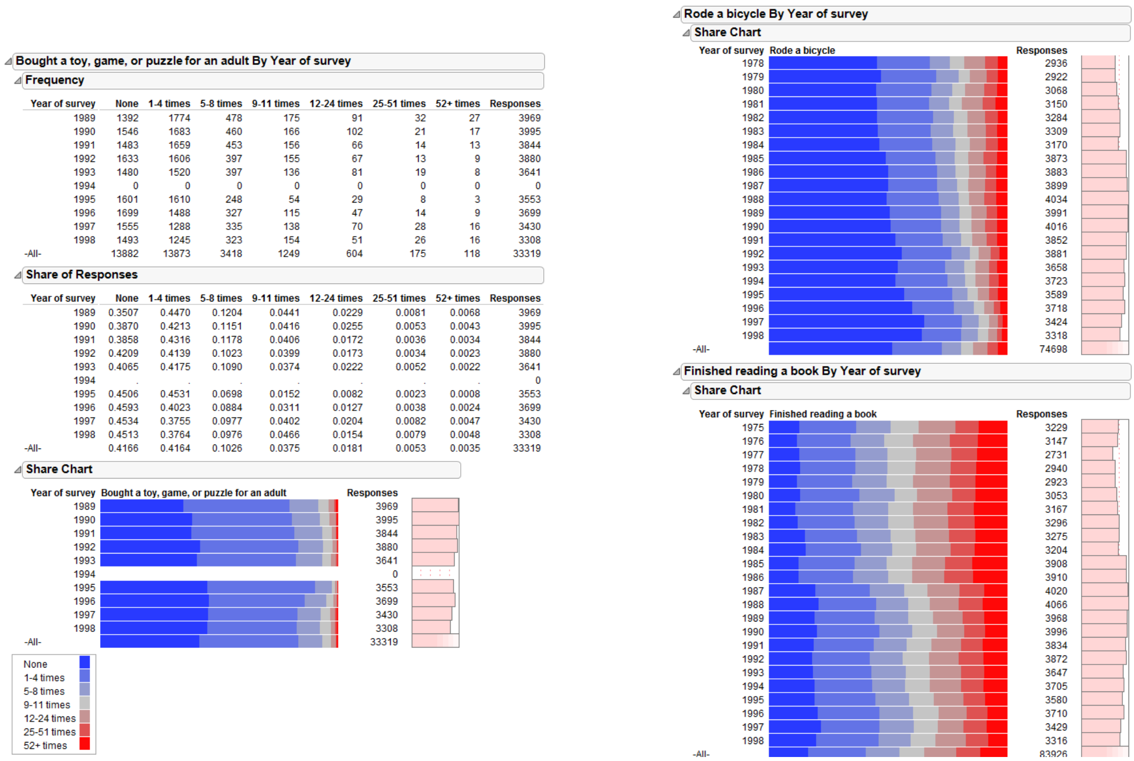


Figure 14. Categorical Analysis Launch Dialog

The default report is to show the frequency tables and response rate tables. To make the result more compact, I can uncheck the Frequencies and Share of Responses and Legend to get only the Share Charts, as shown by the right-hand charts in Figure 15. On the right, you see the change over time in two of the variables, riding a bicycle and reading a whole book. There are 102 more like this.



Bought a toy, game, or puzzle for an adult By Year of survey

Frequency

Year of survey	None	1-4 times	5-8 times	9-11 times	12-24 times	25-51 times	52+ times	Responses
1989	1392	1774	478	175	91	32	27	3969
1990	1546	1683	460	166	102	21	17	3995
1991	1483	1659	453	156	66	14	13	3844
1992	1633	1606	397	155	67	13	9	3880
1993	1480	1520	397	136	81	19	8	3641
1994	0	0	0	0	0	0	0	0
1995	1601	1610	248	54	29	8	3	3553
1996	1699	1488	327	115	47	14	9	3699
1997	1555	1288	335	138	70	28	16	3430
1998	1493	1245	323	154	51	26	16	3308
-All-	13882	13873	3418	1249	604	175	118	33319

Share of Responses

Year of survey	None	1-4 times	5-8 times	9-11 times	12-24 times	25-51 times	52+ times	Responses
1989	0.3507	0.4470	0.1204	0.0441	0.0229	0.0081	0.0068	3969
1990	0.3870	0.4213	0.1151	0.0416	0.0255	0.0053	0.0043	3995
1991	0.3858	0.4316	0.1178	0.0406	0.0172	0.0036	0.0034	3844
1992	0.4209	0.4139	0.1023	0.0399	0.0173	0.0034	0.0023	3880
1993	0.4065	0.4175	0.1090	0.0374	0.0222	0.0052	0.0022	3641
1994	.	.	.	.	.	.	.	0
1995	0.4506	0.4531	0.0698	0.0152	0.0082	0.0023	0.0008	3553
1996	0.4593	0.4023	0.0884	0.0311	0.0127	0.0038	0.0024	3699
1997	0.4534	0.3755	0.0977	0.0402	0.0204	0.0082	0.0047	3430
1998	0.4513	0.3764	0.0976	0.0466	0.0154	0.0079	0.0048	3308
-All-	0.4166	0.4164	0.1026	0.0375	0.0181	0.0053	0.0035	33319

Figure 15. Frequency Table and Share Charts from Categorical Platform

Now I want to see how the response ‘Finished reading a book’ breaks down by region, sex, and educational level. To find these columns among the 389 columns in the table, I use the new search control in the column selection list. Then I select ‘Each Individually’ to get a separate breakdown for each of the three X variables.

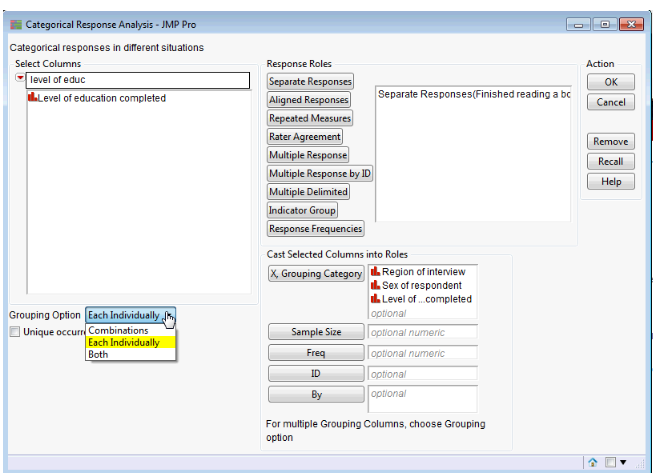


Figure 16. Categorical Launch Window Specifying Grouping Variables and Grouping Option

To the resulting analysis, I make similar option selections to get just the charts, and now I see which regions, which genders, and which educational levels are associated with more frequent reading (Figure 17).

JMP 9 and Interactive Statistical Discovery, continued

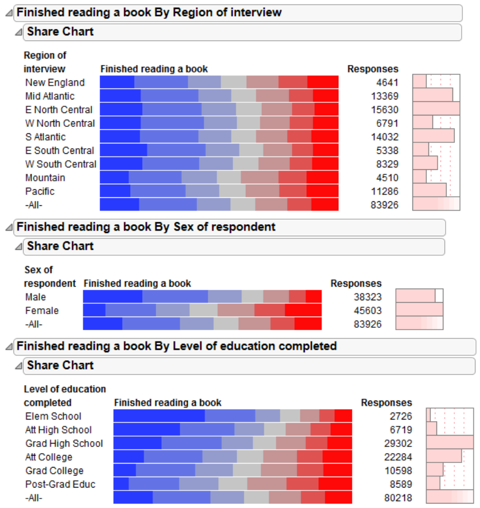


Figure 17. Share Charts from Categorical Analysis Results

EXCEL MODELING

When you estimate models in JMP, the Profiler is a great way to explore the response surface. But what if your model is a set of business calculations in an Excel spreadsheet? Moving all those calculations to JMP data table formulas is too laborious to be practical in most cases. With JMP 9, you can profile a model that is still in Excel. Here is a financial model for the Airbus 380 in Excel. Note that additional command items have been added for JMP. With the 'Edit/New Model' button, I can specify which cells are dependent responses, and which cells correspond to inputs that I want to profile against (Figure 18).

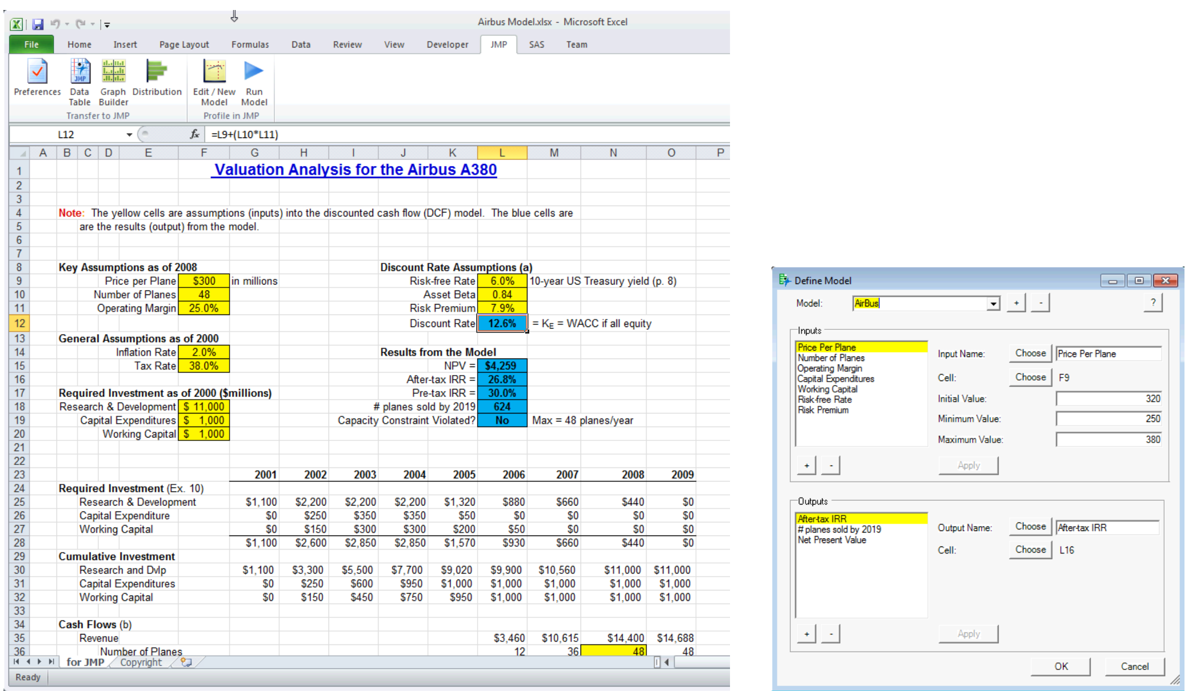


Figure 18. Excel Spreadsheet with Formulas and Define Model Dialog for JMP

When I click Run Model, it connects the model with JMP and JMP repeatedly has Excel re-evaluate the model under many different factors settings, resulting in the Profiler in Figure 19. You can drag the current values around to see how the other profile traces are affected. This gives you a great view of how changes in any variable would affect all the responses.

You can even bring up a simulator to explore distributions of the responses with respect to random distributions in one or more factors.

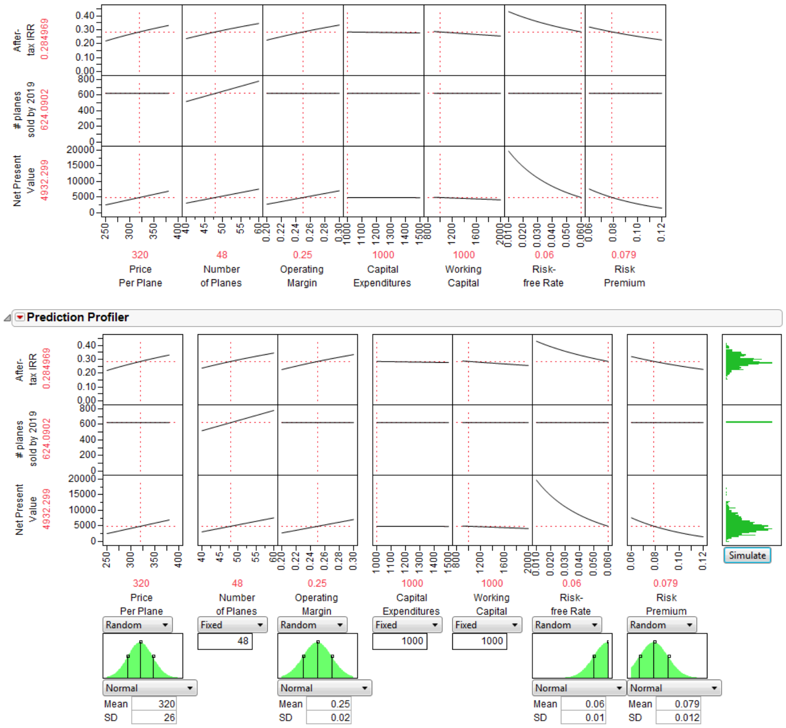


Figure 19. Profiler Display and Simulator Using Excel Formula

DATA MINING

In JMP Pro, Version 9, the Partition platform, which creates decision trees, includes new crossvalidation features—the splitting stops when the validation measures of fit stop improving the fit.

In addition, there are two new measures implemented. In bootstrap forest,

- a number of decision trees are built based on different bootstrap samples
- candidate set columns are selected randomly

These features usually perform better than a straight decision tree. For example, lets use several methods to fit the Wisconsin breast cancer data from the UCI repository (Figure 20 and Figure 21).

	ID	Clump Thickness	Cell Size Uniformity	Cell Shape Uniformity	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	2
2	1002945	5	4	4	5	7	10	3	2	1	2
3	1015425	3	1	1	1	2	2	3	1	1	2
4	1016277	6	8	8	1	3	4	3	7	1	2
5	1017023	4	1	1	3	2	1	3	1	1	2
6	1017122	8	10	10	8	7	10	9	7	1	4
7	1018099	1	1	1	1	2	10	3	1	1	2
8	1018561	2	1	2	1	2	1	3	1	1	2
9	1033078	2	1	1	1	2	1	1	1	5	2
10	1033078	4	2	1	1	2	1	2	1	1	2
11	1035283	1	1	1	1	1	1	3	1	1	2
12	1036172	2	1	1	1	2	1	2	1	1	2
13	1041801	5	3	3	3	2	3	4	4	1	4
14	1043999	1	1	1	1	2	3	3	1	1	2
15	1044572	8	7	5	10	7	9	5	5	4	4
16	1047630	7	4	6	4	6	1	4	3	1	4

Figure 20. Partial Listing of Wisconsin Breast Cancer Data

JMP 9 and Interactive Statistical Discovery, continued

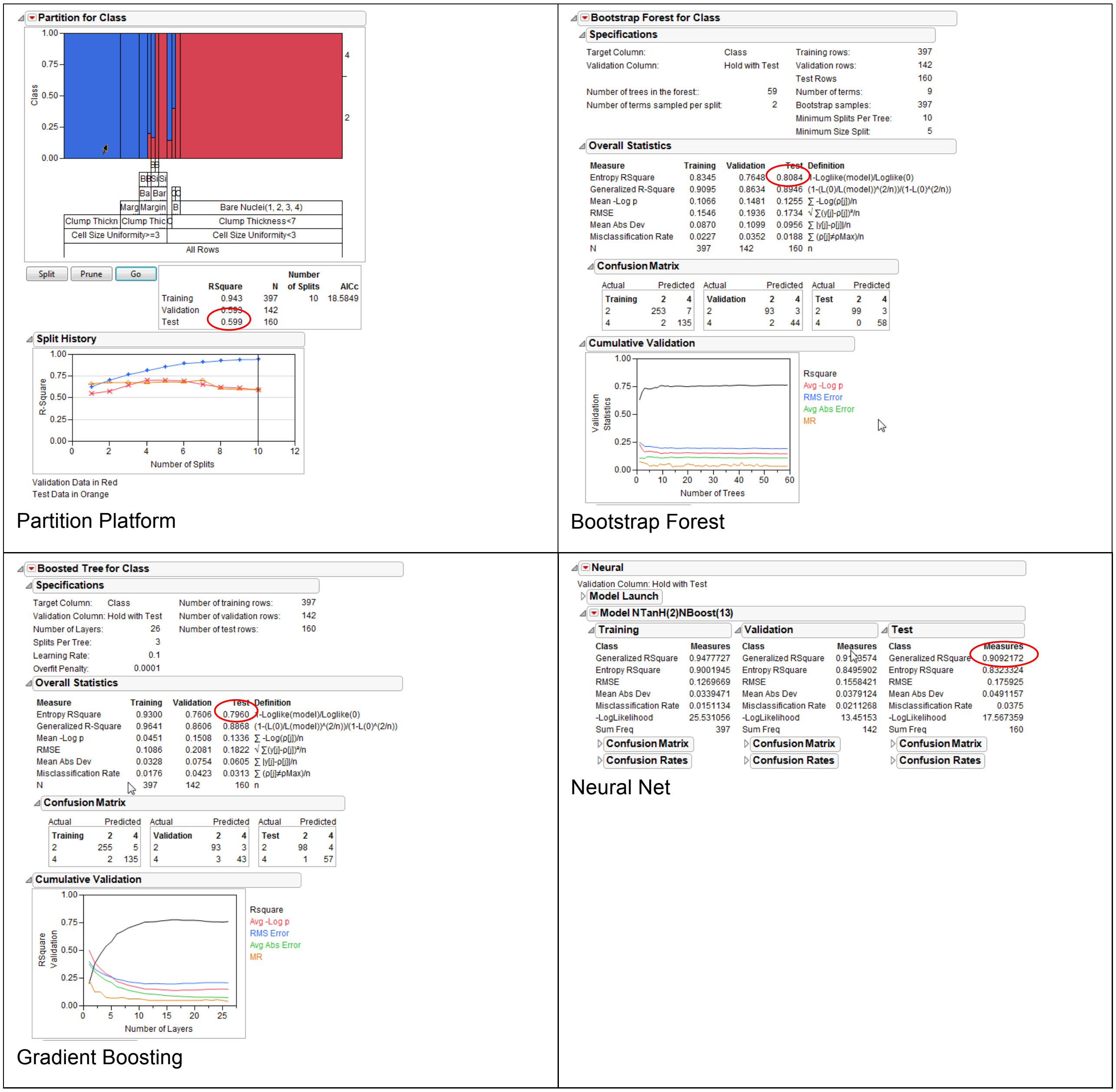


Figure 21. Several Methods of Wisconsin Breast Cancer Data

The best performing method differs for different data sets, and even for how you randomly choose the training, validation, and test sets, and differs each time due to randomness in the methods. But, usually Neural Net is the best and a single decision tree is the worst. Table 1 summarizes the fitting results as reflected by R Square values.

Method	Test RSquare (entropy)
Decision Tree	0.599
Bootstrap Forest	0.808
Gradient Boosting	0.796
New Neural Net	0.909

Table 1. Summary of Fitting Results

CONCLUSION

JMP 9 has many new features, and only a few of them are described in this paper. With renewed features in scripting, we expect JMP to be a good implementation language for many new applications. In addition to JMP's traditional user base of engineers and scientists, we expect other researchers and analysts to discover JMP, both as a stand-alone application, and as a front-end and explorer part of the SAS system.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

John Sall
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 677-8000
john.sall@jmp.com
www.jmp.com