New Features in SAS/INSIGHT® in Version 7

Marc-david Cohen, Hong Chen, Yang Yuan, and Frederick Wicklin
Applications Division
SAS Institute Inc.
Cary NC

ABSTRACT

There are significant new features in Version 7 SAS/INSIGHT. These include new statistical analyses and enhancements to the graphics. INSIGHT now supports several multivariate statistical techniques including principal component rotation analysis, canonical correlation analysis, maximum redundancy analysis, and canonical discriminant analysis, and it also supports comparison of means. There are also new robust measures of scale and tests for normality for univariate data as well as tests for differences of means across groups. New graphical enhancements include 3D surface plots, contour plots, 3D response surfaces, comparison of means circles in the box plots, and color blending of up to 5 colors. Several methods for surface fitting are provided including linear interpolation, thin-plate spline, kernel estimation, and using a parametric model.

INTRODUCTION

Version 7 contains significant enhancements to SAS/INSIGHT in its statistical functionality and graphical features. This paper highlights the major new features and shows examples of many of them. In the first sections, the new statistical functions are discussed. In the later sections, the new graphical features, including response surfaces, are highlighted. All the graphs are shown in monotone, however, so please keep in mind that the software supports colors, and many of the features require color to be effective.

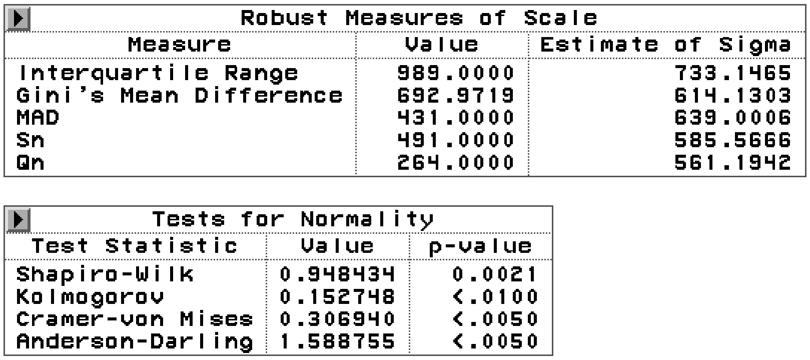
DISTRIBUTION ANALYSIS

There are several new univariate statistics in Version 7. These cover

- basic confidence intervals of mean, standard deviation, and variance
- robust measures of scale
- tests for normality

Under the ANALYZE:DISTRIBUTION(Y) pull-down menu you can specify the statistics of interest via the Output dialog. Figure 1 shows the tables produced when you request robust scale statistics and tests for normality.

Figure 1. New Univariate Statistics



Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	989.0000	733.1465
Gini's Mean Difference	692.9719	614.1303
MAD	431.0000	639.0006
Sn	491.0000	585.5666
Qn	264.0000	561.1942

Tests for Normality

Test Statistic	Value	p-value
Shapiro-Wilk	0.948434	0.0021
Kolmogorov	0.152748	<.0100
Cramer-von Mises	0.306940	<.0050
Anderson-Darling	1.588755	<.0050

MULTIVARIATE ANALYSES

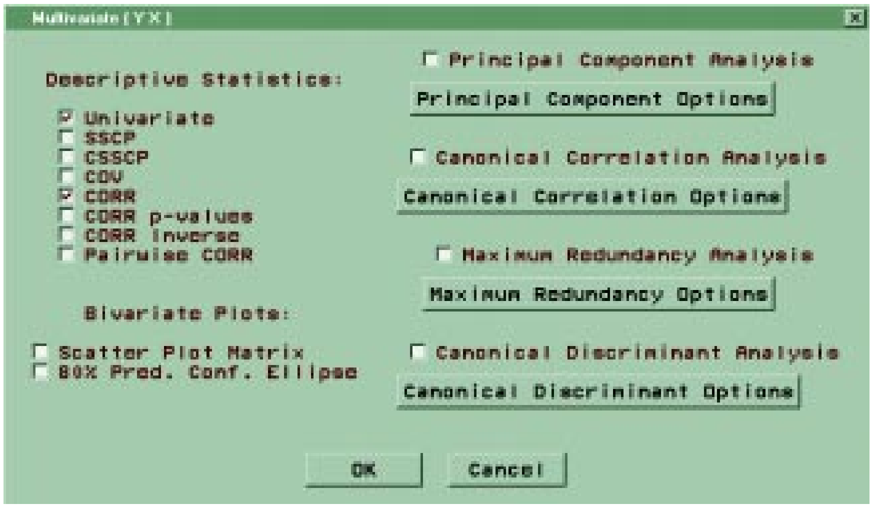
When you choose ANALYZE:MULTIVARIATE (YX) from the pull-down menus, you gain access to a variety of multivariate analyses. These methods provide a way of examining relationships among a set of variables and between two sets of variables.

In earlier versions of SAS/INSIGHT, limited multivariate analysis is supported. In Version 7 you can use principal component analysis to examine relationships among several variables, principal component rotation to obtain factors that are more easily interpretable, canonical correlation analysis and maximum redundancy analysis for relationships between two sets of interval variables, and canonical discriminant analysis for relationships between a nominal variable and a set of interval variables. The following table shows the requirements.

METHOD	X's	Y's
Canonical Correlation	Multiple	Multiple
Maximum Redundancy	Multiple	Multiple
Canonical Discriminant Analysis	Multiple	Single Nominal
Principle Component Analysis	None	Multiple

To demonstrate one of these methods, consider the data from the 1995 U.S. News & Report on American colleges and universities. They include demographic information on tuition, room and board costs, SAT or ACT scores, application/acceptance rates, student/faculty ratio, and graduation rate. If you select ANALYZE:MULTIVARIATE (YX) from the Analyze pull-down menu and press the Output button on the Multivariate dialog, you open the dialog shown in Figure 2. This shows the different multivariate analyses supported by the software. Select the type of analysis you want to perform in the appropriate check box then press the associated button to specify options for that analysis.

Figure 2. Multivariate (YX) Dialog



Multivariate (YX)

Descriptive Statistics:

☑ Univariate
□ SSCP
□ CSSCP
□ COV
☑ CORR
□ CORR p-values
□ CORR inverse
□ Pairwise CORR

Bivariate Plots:

□ Scatter Plot Matrix
□ 80% Pred. Conf. Ellipse

□ Principal Component Analysis

Principal Component Options

□ Canonical Correlation Analysis

Canonical Correlation Options

□ Maximum Redundancy Analysis

Maximum Redundancy Options

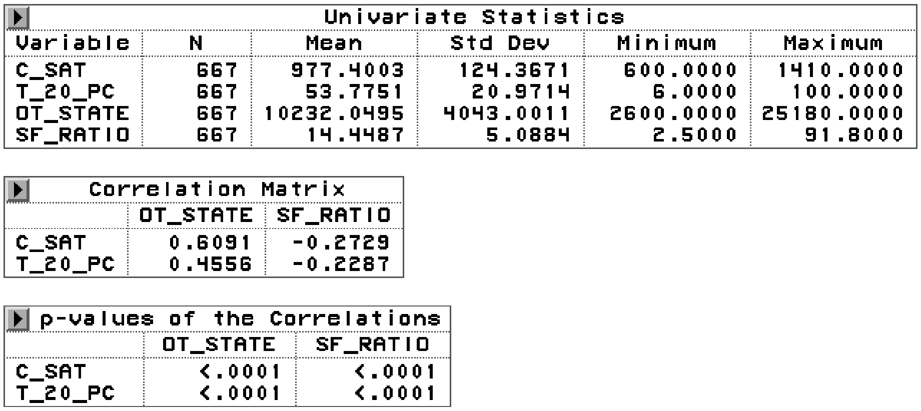
□ Canonical Discriminant Analysis

Canonical Discriminant Options
OK
Cancel

Suppose that you want to see how well the out-of-state tuition costs and the student-faculty ratio explain the variation in the quality of the student body as measured by the entering SAT scores and the percentage of the entering class that is in the top 20 percent of their graduating class. You select the appropriate variables and type of analysis.

The output first lists the two sets of variables upon which the analysis is being performed and then includes scatter plots with 80% confidence ellipses for each of the variables in each of the two groups. This is similar to the output in Version 6 of SAS/INSIGHT and is not shown here.

Figure 3. Univariate Statistics



Univariate Statistics

Variable	N	Mean	Std Dev	Minimum	Maximum
C_SAT	667	977.4003	124.3671	600.0000	1410.0000
T_20_PC	667	53.7751	20.9714	6.0000	100.0000
OT_STATE	667	10232.0495	4043.0011	2600.0000	25180.0000
SF_RATIO	667	14.4487	5.0884	2.5000	91.8000

Correlation Matrix

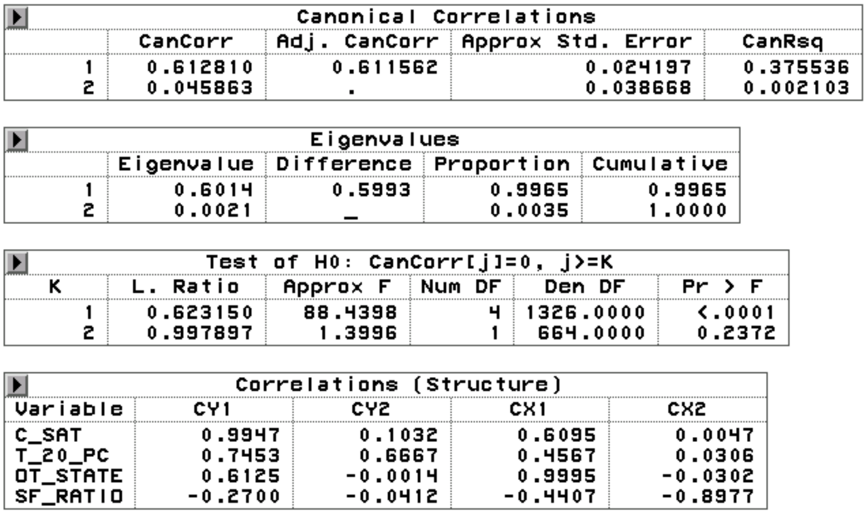
	OT_STATE	SF_RATIO
C_SAT	0.6091	-0.2729
T_20_PC	0.4556	-0.2287

p-values of the Correlations

	OT_STATE	SF_RATIO
C_SAT	<.0001	<.0001
T_20_PC	<.0001	<.0001

Next are univariate statistics, the correlation matrix, and the *p*-values associated with the correlations. Notice that, according to the *p*-values, all the correlations are significant (see Figure 3).

Figure 4. Canonical Correlation Output



Canonical Correlations

	CanCorr	Adj. CanCorr	Approx Std. Error	CanRsq
1	0.612810	0.611562	0.024197	0.375536
2	0.045863	.	0.038668	0.002103

Eigenvalues

	Eigenvalue	Difference	Proportion	Cumulative
1	0.6014	0.5993	0.9965	0.9965
2	0.0021	_	0.0035	1.0000

Test of H0: CanCorr[j]=0, j>=K

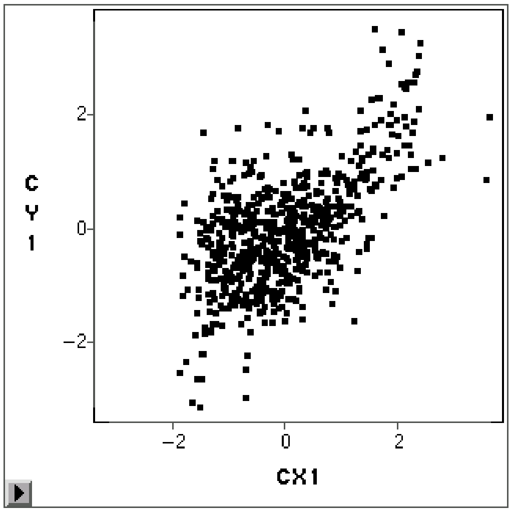
K	L. Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.623150	88.4398	4	1326.0000	<.0001
2	0.997897	1.3996	1	664.0000	0.2372

Correlations (Structure)

Variable	CY1	CY2	CX1	CX2
C_SAT	0.9947	0.1032	0.6095	0.0047
T_20_PC	0.7453	0.6667	0.4567	0.0306
OT_STATE	0.6125	-0.0014	0.9995	-0.0302
SF_RATIO	-0.2700	-0.0412	-0.4407	-0.8977

Next, as shown in Figure 4, the canonical correlations are printed along with eigenvalues from a matrix of cross products, the results of a test which show that the canonical correlations are zero, and the correlations of the original variables with the canonical variables. Much more detail can be included in the output, including the canonical coefficients.

Figure 5. Scatter Plot of Canonical Variable



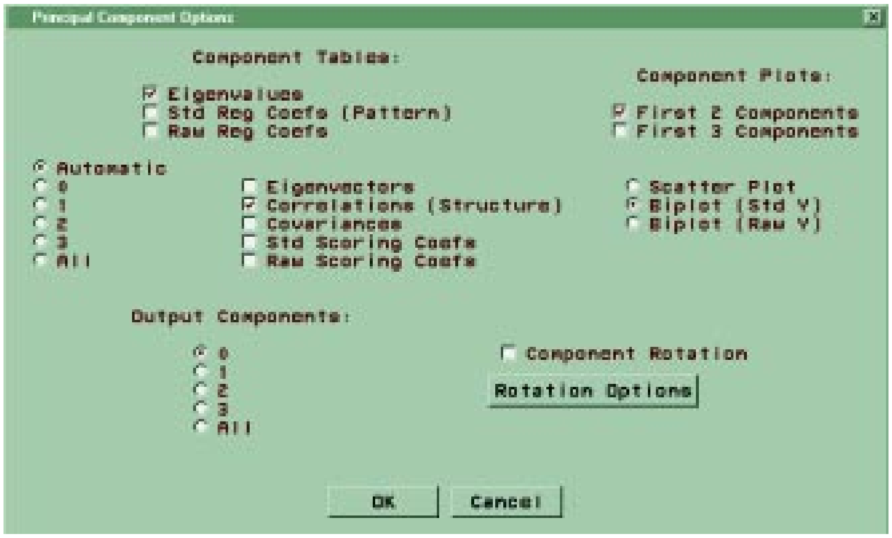
Notice that the second canonical variable fails the test of being different from zero. This means that much of the variation in the two sets of data can be captured in the first canonical variable. Figure 5 shows a scatter plot of the first canonical variable. This is a plot of the inner products of the two sets of standardized scoring coefficients associated with the first canonical variable and the standardized data.

Biplots

One key new feature in the multivariate analysis is the ability to create 2d and 3d biplots. A biplot is a way of plotting the variables and the observations together in a single plot; observations are points and variables are vectors. In a biplot the projection of an observation onto a variable shows the contribution of that variable to that point and the length of the axis is approximately proportional to the standard deviations of the variable. If the length of the projection is long then that variable had a large contribution to that observation. See Gower and Hand (1996) for a through discussion.

Consider data from the United States Department of Commerce on the lowest temperatures (in F) recorded in various months for cities in the US. Can the cities be characterized by these data? It is interesting to fit a two component model to these data and plot the principal components.

Figure 6. Principal Component Analysis Output Dialog



Principal Component Options

Component Tables:

☑ Eigenvalues
☐ Std Reg Coefs (Pattern)
☐ Raw Reg Coefs

● Automatic
○ 0
○ 1
○ 2
○ 3
○ All

☐ Eigenvectors
☑ Correlations (Structure)
☐ Covariances
☐ Std Scoring Coefs
☐ Raw Scoring Coefs

Component Plots:

☑ First 2 Components
☐ First 3 Components

○ Scatter Plot
● Biplot (Std Y)
○ Biplot (Raw Y)

Output Components:

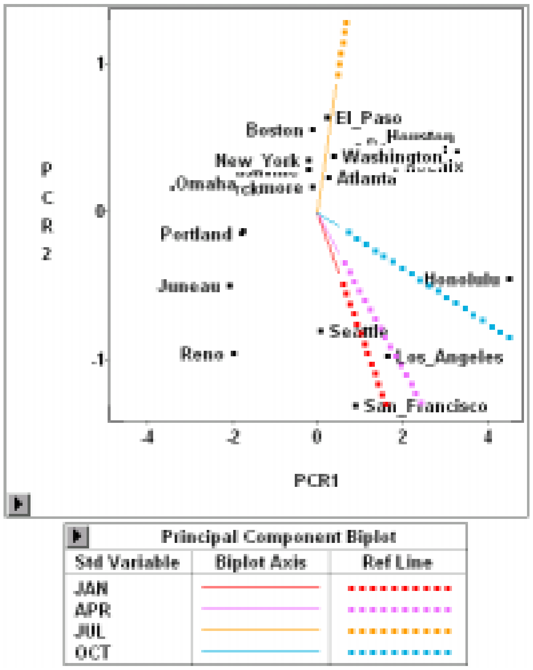
● 0
○ 1
○ 2
○ 3
○ All

☐ Component Rotation

Select ANALYZE:MULTIVARIATE (YX) and choose the 4 months as the Y variables. These are the variables containing the lowest temperature for each of four months. The Output button opens the dialog shown in Figure 6 and the Principal Component button opens the dialog shown in Figure 2. Selecting the Biplot radio button produces the desired plot.

SAS/INSIGHT shows the length of the vector as a solid line and extends the vector with a dotted line to clarify the vector's direction. Figure 7 shows a biplot in the principal component space. The output not shown here confirms that the Y variables are well approximated by the biplot.
Notice how the plot of the cities in the principal component space shows the division of cities graphically. The West Coast cities are grouped clearly. The biplot shows that Honolulu differentiates itself from the other Pacific Rim cities by its large distance from the July axis.

Figure 7. Principal Component Biplot

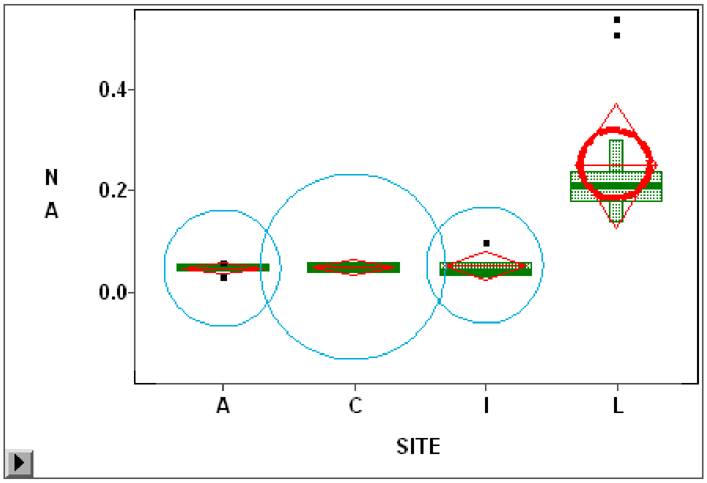


Principal Component Biplot		
Std Variable	Biplot Axis	Ref Line
JAN	—	········
APR	—	········
JUL	—	········
OCT	—	········

COMPARISON OF MEANS

Another area where the statistical content has been enhanced is in comparing means across groups. Consider data taken from *A Handbook of Small Data Sets* (Hand et al. 1994) on the result of chemical analysis of 26 samples of pottery found at kiln sites in Wales, Gwent and the New Forest. The variables are the percentages of oxides of the various metals indicated, and the sites are L: Lllanederyn, C: Caldicot, I: Island Thorns, A: Ashley Rails. The analysis includes the quantities of aluminum, iron, magnesium, calcium, and sodium. Select ANALYZE:BOX PLOT/MOSAIC PLOT from the pull-down menus and choose Site as the X variable and each of the minerals as the Y variable. As in previous releases of SAS/INSIGHT, a box plot is drawn for each of the five variables.

Figure 8. Box plot with Comparison of Means Circles



The box plot shown in Figure 8 suggests that the mean sodium level at site L is higher than at the other sites. INSIGHT now supports a variety of methods to determine whether this difference is statistically significant. Choosing the Comparison Circles option from the box plot's pop-up menu performs a test for the multiple comparison of means at a given confidence level. A comparison circle appears for each site: the circle's radius indicates how confident you can be of the location of the mean. Figure 8 shows the result of choosing the Bonferroni test and then selecting the circle for Site L. The selected circle is highlighted. Groups with means that are significantly different than the selected group are blue (light gray here). The software also supports the printing of a table of statistics (not shown here).

SURFACE PLOTS and CONTOUR PLOT

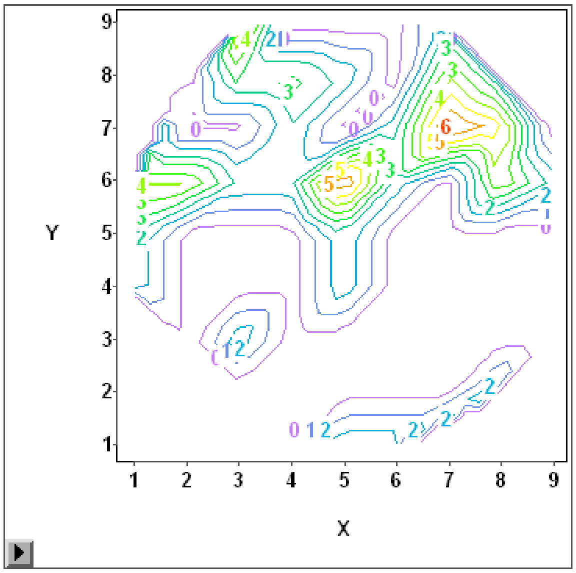
The addition of surface and contour plots is an important enhancement to SAS/INSIGHT in Version 7. These appear in the product in various ways. You can fit a surface plot to the ANALYZE:ROTATING PLOT (ZYX), you can see a response surface from ANALYZE:FIT (YX), and you can select the ANALYZE:CONTOUR PLOT (ZYX) entry in the Analyze pull-down menu.

The following example uses data on the number of metal short failures per chip. In order to measure process stability, a set of metal comb structures is used to gather information on lateral shorts on a defect density test chip. The chip contains 16 equally sized comb structures that are sized so that usually only one fault per comb is counted. There are 64 chips per wafer on a 150 millimeter diameter wafer. From these data you can make a wafer map of defects by fitting a surface to the defect data and displaying the contours of that surface.

There are two methods available for fitting the surface: linear interpolation and thin-plate spline. The linear interpolation fit is the default. It is a simple and fast method and, in many cases, sufficient. However, a spline surface can be fit by pressing the Method push button on the Contour dialog box. Although the linear fit is much faster than the spline method, it does not produce a smooth surface.

Figure 9 shows a linear fit to the wafer defect data. Notice that each contour is labeled with a value of the modeled surface. You can use the hand tool to highlight and move all contours corresponding to a given level. You can also add contour levels by selecting a point through which no contour is drawn. This gives you great flexibility in how the contours are displayed.

Figure 9. Wafer Map Contour Plot



The pop-up menu on the contour plot, shown in Figure 10, enables you to fill the areas between the contour lines with color. You simply select Fill Areas and the areas fill with the colors that have been selected in the Tools window (see the section "Color Palette" later in this paper). The Color Blending selection colors the contour lines themselves (without filled areas) with the colors selected in the tools palette. Also, you can turn the contour lines off and just show the filled areas for a smoother visual transition between regions.

Figure 10. Contour Pop-up Menu

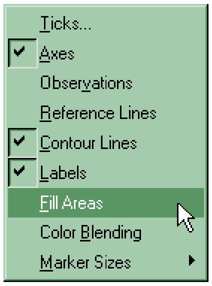
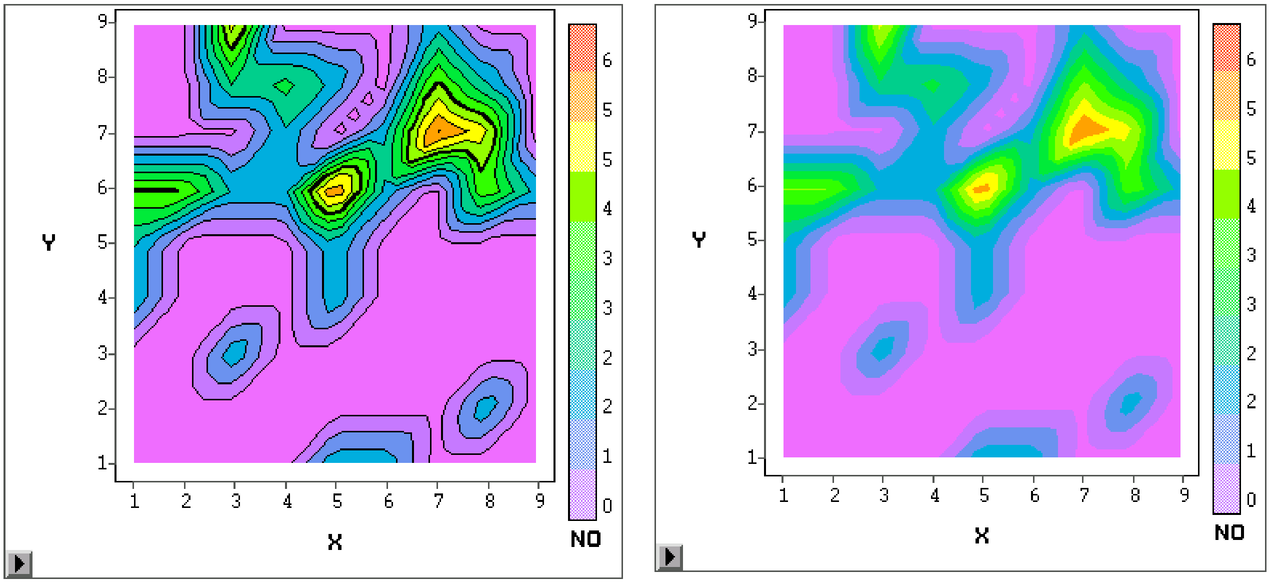


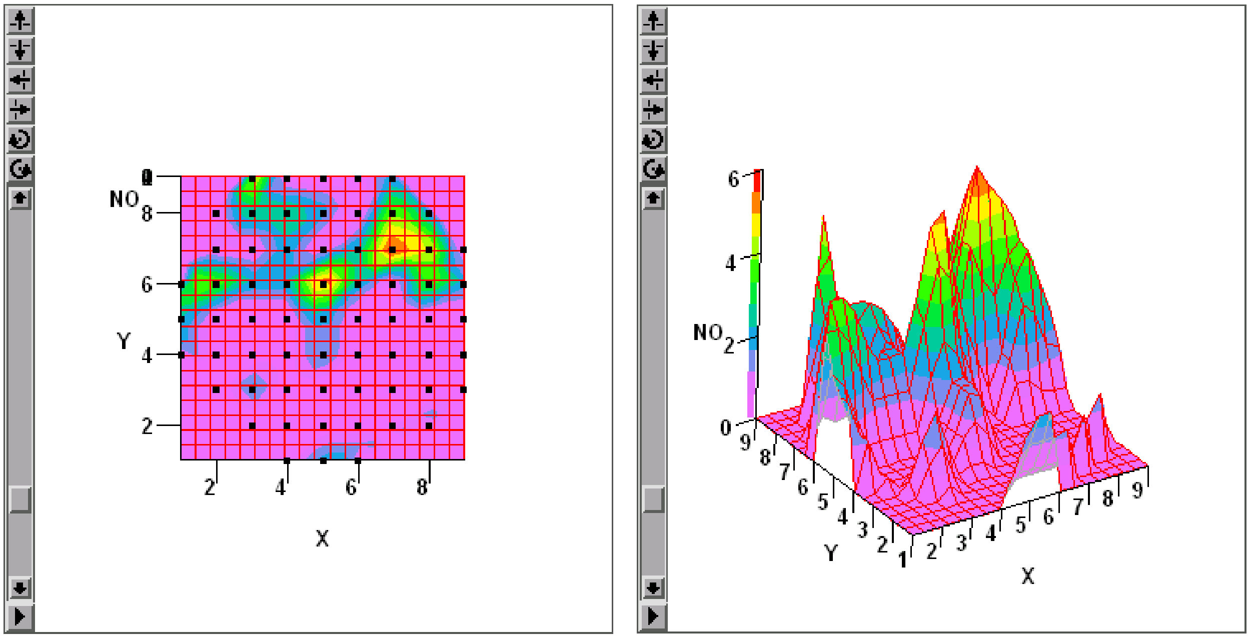
Figure 11 shows two such plots. The one on the left has the contour level 4 selected. Notice how the contour line is highlighted. Also, the legend bar to the right of the plot has level 4 highlighted. The plot on the right is the same plot with the contour lines not shown.

Figure 11. Two Area Filled Contour Plots Showing Highlighted Level and No Contour Lines



The ANALYZE:ROTATING PLOT (ZYX) also supports a surface that can rotate. The type of fit is selected in the Method dialog, and the surface is selected in the Output dialog. Figure 12 shows the rotating plot with the wafer defect data with a linear fit. The default color palette is used. On the left is the plot as it initially appears with the observations shown. On the right is the same plot

Figure 12. Two Surface Plots on the Rotating Plot Shown in Different Positions



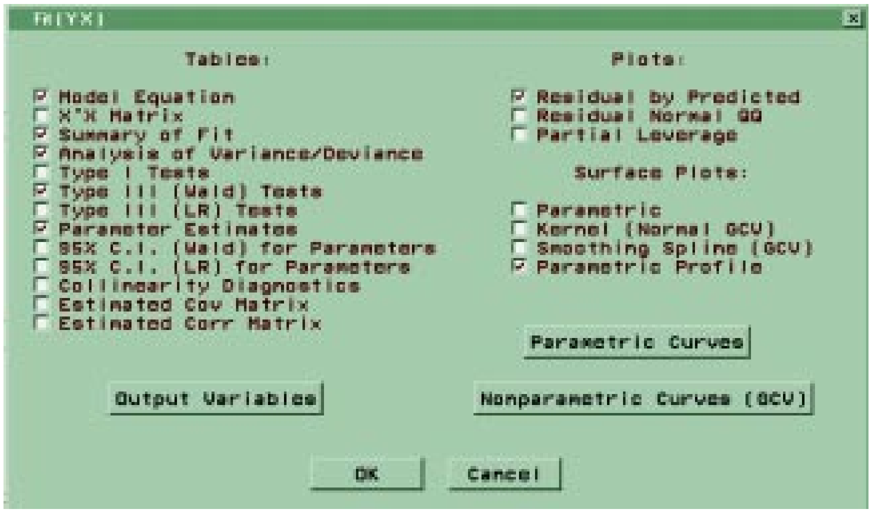
rotated slightly and with the observations turned off. Notice that the Z axis acts as a color legend.

Another new feature in the rotating plot is the Three Section axes. Making this selection causes the axes to move around the exterior of the plot so that all three axes are always visible and none are hidden by the plot itself.

Response Surfaces

Suppose you fit a quadratic model to the 1995 U.S. News & Report college data using the ANALYZE:FIT(YX) pull down selection. Here you predict the out of state tuition costs with the entering SAT scores, percent of entering class in the top 20% of their high school class, the student-faculty ratio, and the percentage of accepted from those that apply. The details of the model are not important to this discussion, but note that there are four independent variables. If you select the Output button on the Fit dialog, then you open a dialog from which you can select the type of response surface to show. Figure 13 shows this dialog with Parametric Profile selected.

Figure 13. Output Dialog for Specifying Response Surfaces



Fit(YX)

Tables:

☑ Model Equation
☐ X'X Matrix
☑ Summary of Fit
☑ Analysis of Variance/Deviance
☐ Type I Tests
☑ Type III (Wald) Tests
☐ Type III (LR) Tests
☑ Parameter Estimates
☐ 95% C.I. (Wald) for Parameters
☐ 95% C.I. (LR) for Parameters
☐ Collinearity Diagnostics
☐ Estimated Cov Matrix
☐ Estimated Corr Matrix

Plots:

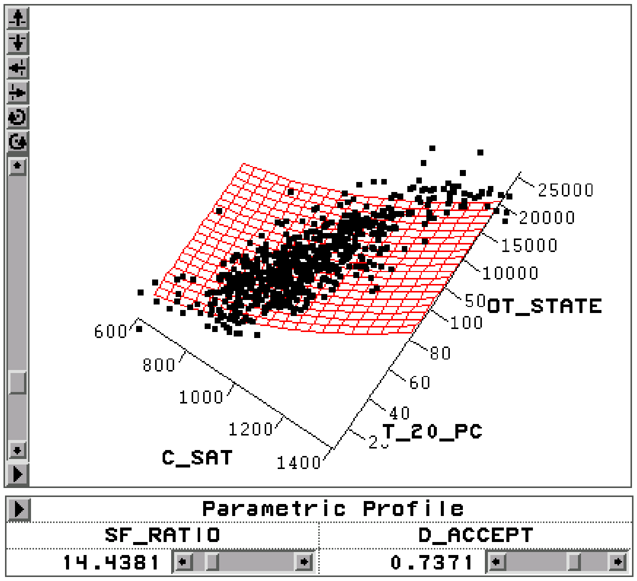
☑ Residual by Predicted
☐ Residual Normal QQ
☐ Partial Leverage

Surface Plots:

☐ Parametric
☐ Kernel (Normal GCV)
☐ Smoothing Spline (GCV)
☑ Parametric Profile

This results in the same model being fit as in Version 6 but with the additional rotating plot,

Figure 14. Output from Parametric Profile response surface



Parametric Profile

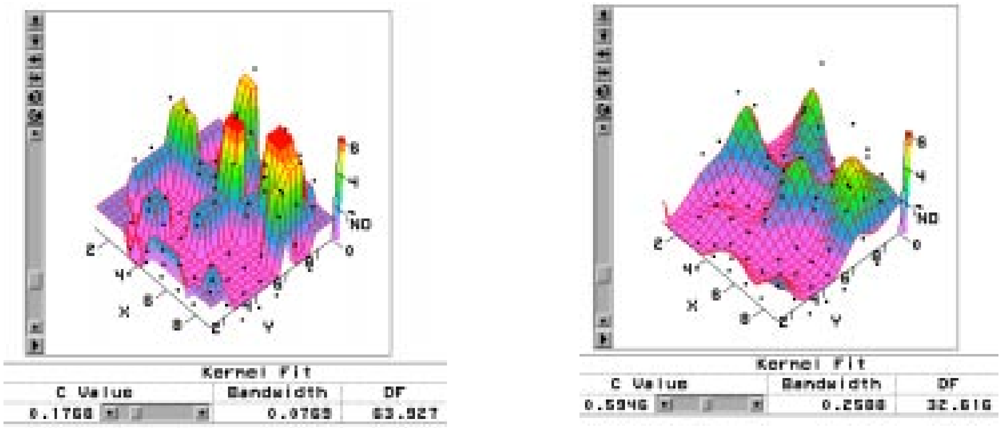
SF_RATIO 14.4381

D_ACCEPT 0.7371

shown in Figure 14, displayed in the output. The surface shown is the response surface from the regression model. All the features of the surface plot discussed previously apply to this rotating plot with response surface. Note that, since the parametric model has four independent variables, the surface shown is a section for fixed values of two of the variables, namely student-faculty ratio and percent students accepted. The sliders below the plot show the current values for these two variables and give you the flexibility to change those values and see how the surface changes dynamically.

The ANALYZE:FIT(YX) also can produce a response surface calculated from a thin-plate spline fit or from kernel estimation. This approach to fitting a surface gives you access to the fitting parameters. For example, consider the wafer data discussed previously. In Figure 15 you use kernel fit with two different fitting parameters. The surface on the left shows a smaller bandwidth used in the kernel fit, which results in a surface that shows higher frequency information than the

Figure 15. Two Surface Plots in Rotating Plots Showing Affects of Varying Bandwidths



Kernel Fit

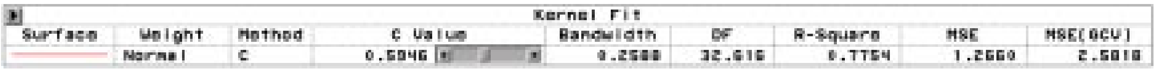
C Value	Bandwidth	DF
0.1766	0.0765	63.327

Kernel Fit

C Value	Bandwidth	DF
0.5946	0.2500	32.616

surface on the right. Since this parameter is controlled by a slider, you can dynamically adjust the bandwidth and see how the surface changes with the changing parameter.

Figure 16. Kernel Fit Output



Kernel Fit

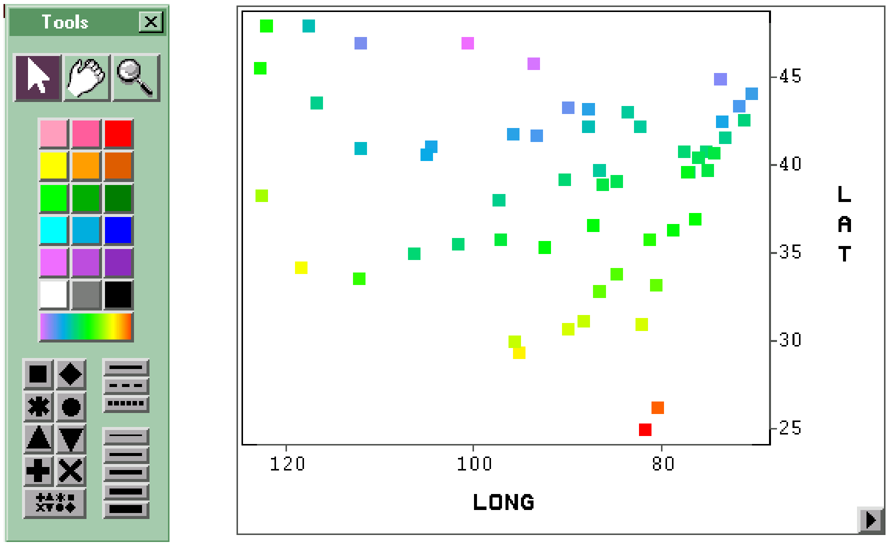
Surface	Weight	Method	C Value	Bandwidth	DF	R-Square	MSE	MSE(GCV)
	Normal	C	0.5046	0.2588	32.616	0.7754	1.2660	2.5018

Note that not all of the table for the kernel fit is shown in Figure 15. In addition to the parameter, bandwidth, and degrees of freedom, the R-square, and mean square errors for the fit are also given as in Figure 16. These dynamically update with the changing parameter and resulting surface.

Color Palette

The color palette in the Tools window has been enhanced to support the contour and surface plots. In Version 6 the color palette supported a two color mixture. It now supports a five color mixture. Colors are added to the mixture as before with one exception; you pick up a color **with the shift key pressed** and drop it on the mixing strip in one of five positions. If you press the mixing strip and select a variable, it is colored according to its value as in Version 6, but because of the multiple color mixture you can display more information. For example, Figure 17 shows average temperature in several cities across the United States. The hotter the average temperature, the hotter the color.

Figure 17. Tools Window and Scatter Plot Showing Enhanced Multi-color Highlighting



FUTURE DIRECTIONS

This paper has highlighted some of the new features in Version 7 SAS/INSIGHT. This release is a major step forward in the product, as representative of SAS Institute’s commitment to it and as one step in a continuing development process that has several projects for the future. These include

- continued enhancement of Statistical content
  - robust methods
  - modern Non-parametric methods for smoothing and regression
- continued enhancement of Graphical methods
  - projecting a contour plot from the surface plot
  - cutting planes through surface plots
- improvements in the User Interface
- merging of SAS/IML® and SAS/INSIGHT

The last of these projects is perhaps the most exciting in that it will provide a rich scripting language for SAS/INSIGHT and a comprehensive set of graphical tools for SAS/IML.

REFERENCES

Collica, R.S., (1992), "The Effect of the Number of Defect Mechanism on Fault Clustering and Its Detection Using Yield Model Parameters," *IEEE Trans. On Semi. Manuf.* 5 (3), 189--195.

Hand, D.J., F. Daly, A.D. Lunn, K.J. McConway and E. Ostrowski, (1994), *A Handbook of Small Data Sets*, London: Chapman & Hall.

Gower J.C., and D.J. Hand, (1996), *Biplots*, London: Chapman & Hall.

Tubb, A., Parker, A.J. and Nickless, G., (1980), "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry," *Archaeometry*, 22, 153--171.

US Bureau of the Census (1988), *Statistical Abstract of the US*, United States Department of Commerce, Bureau of the Census.

ACKNOWLEDGMENTS

We thank Randy Collica for suggesting a source of wafer defect data and Robert Rodriguez and Donna Sawyer for editing assistance.