Paper 241-25

Analysis of Environmental Data Using JMP®

Andy Mauromoustakos, University of Arkansas, Fayetteville, AR

ABSTRACT

The paper will focus on new features of JMP® (Version 4) by performing analysis on two different environmental datasets. The first dataset is an observational study that requires a Split-Plot type analysis and REML estimation of the two random error terms. Comparisons of the main-unit treatment factor for each sub-unit treatment level are done correctly in the new Fit Model platform a feature that did not exist in the Previous Version. The second example dataset contains information on the physical and chemical properties, morphological description in eight selected soil profiles in Arkansas. Multivariate exploration of the data with emphasis on the similar but vastly improved functionality of the new version in data manipulation, processing and presentation of the results. JMP Script Language (JSL) allows the researcher to code macros for performing new analysis and drawing customized graphs. Version 4 addresses most of the limitations of the old version and makes it an excellent discovery tool.

INTRODUCTION

At first glimpse, Version 4 looks different than Version 3, but you will soon find that it has the similar but vastly improved functionality as Version 3, with many new conveniences. These conveniences include the following list of improvements and additions:

I. Data Table
- Data Table Window Side Panel
- Formula Editor
- Accessing and Importing Data

II. Launching Platforms
- JMP Starter
- By Groups
- Role Selection
- Saving Scripts

III. Report Windows
- Popup Menus location
- Title and outlines in report windows
- Titles that don't scroll off the top of the window
- Improved Tools and Cursors
- Report Table Customization
- Graph Customization
- Journal
- Layout

IV. Design of Experiments
- Dialog Design
- DOE Metadata
- Custom Designs
- Screening Designs
- Mixture Designs

V. Analysis Platforms
- Distributions
- One way
- Bivariate
- Matched Pairs
- Time Series
- Cluster K-Means
- Multivariate
- Fit Model
- Random Effects REML
- Nominal Terms
- Screening Platform Combined
- LSMEANS comparisons
- Variability Chart
- Logistic

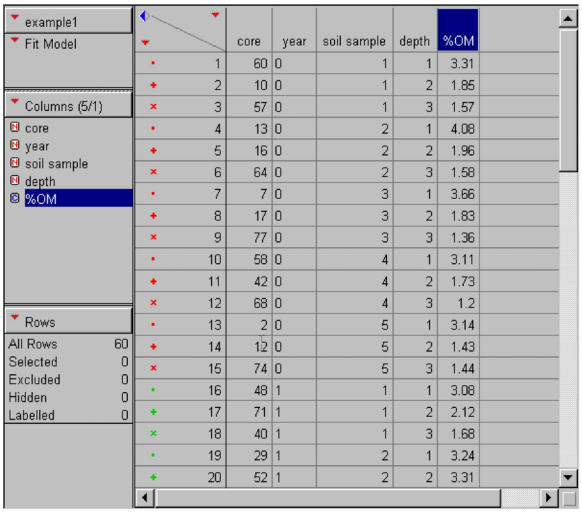
VI. Internals
- Language
- Host Interface
- Presentation
- Data Interfaces

VII. Scripting
- *Production Jobs*
- *Data manipulation*
- *Simulation*
- *Record Keeping*

We will attempt to illustrate some of these improvements in the context of the exploration, and analysis of two environmental datasets.

EXAMPLE 1

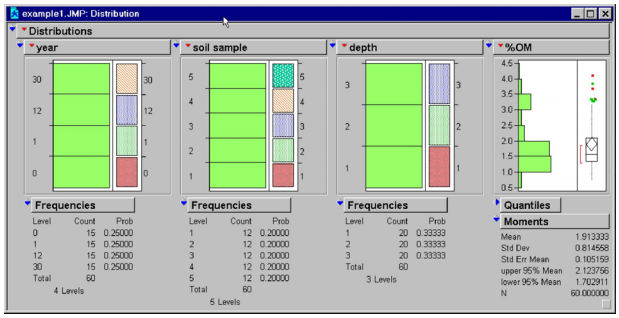
Four fields were sampled in Prairie County, Arkansas that has been on cultivation for different years. These fields included a prairie and three fields that have been on rotation of rice, soybean and wheat for 0, 1, 12 and 30 years respectively. Five soil core samples were selected from each field each at 3 depths within the same sample location (0-5cm, 5-10 cm and 10-15 cm). Several measurements were made on each soil sample but in this paper we will discuss the analysis of the percent organic matter content (%OM). Generally tillage practices are used to improve soil condition for crop growth and development. On short –term basis these practices may be beneficial to crop production and soil productivity. On the other hand, over many years the cumulative effect of these frequent tillage operations and cropping leads to changes in soil physical and biological properties. The analysis of the %OM follows a split-plot experiment analysis as a mixed model that involves two random error components. We will discuss features of the new version of JMP in the context of this analysis and in particular the new REML engine of Fit Model platform that newly available least squares means comparisons.



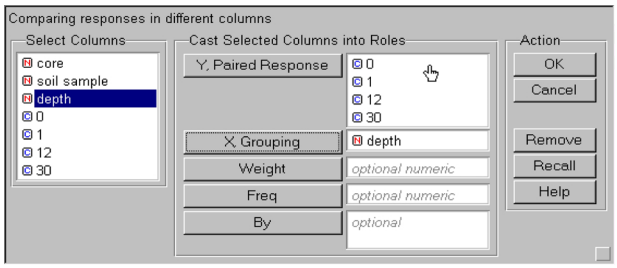
	core	year	soil sample	depth	%OM
1	60	0	1	1	3.31
2	10	0	1	2	1.85
3	57	0	1	3	1.57
4	13	0	2	1	4.08
5	16	0	2	2	1.96
6	64	0	2	3	1.58
7	7	0	3	1	3.66
8	17	0	3	2	1.83
9	77	0	3	3	1.36
10	58	0	4	1	3.11
11	42	0	4	2	1.73
12	68	0	4	3	1.2
13	2	0	5	1	3.14
14	12	0	5	2	1.43
15	74	0	5	3	1.44
16	48	1	1	1	3.08
17	71	1	1	2	2.12
18	40	1	1	3	1.68
19	29	1	2	1	3.24
20	52	1	2	2	3.31

WHOLE-PLOT AND SUB-PLOT EFFECT VISUALIZATION

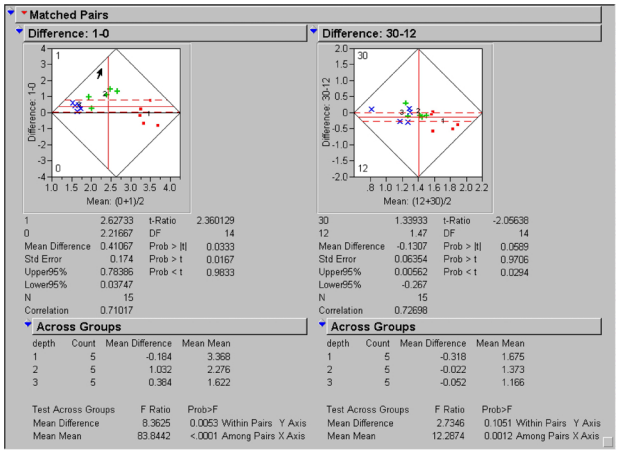
The Distribution of Y after selecting all the factors and response for the experiment is always useful for "checking" the data.



The new Matched Pairs Platform is utilized to produce a visual for the whole-plot (or between subjects) and split-plot part (or within subjects) parts of typical Split-Plot and Repeated Measures types of analysis. First we need to split the response %OM using year as the Col ID. The following Matched Pairs completed dialog and selecting since there is an even number of responses that we do not want all possible pairs provides us a series of pairs (Y2 by Y1 and separately Y4 by Y3 matched pairs).



Two analysis results help us visualize the data (year and depth main effects and interactions) via separate analysis of the 0 and 1 year cultivated fields and the 12 and 30 years cultivated fields respectively. The vertical axis becomes the split-plot or repeated measures axis, and the horizontal axis becomes the whole-plot or Between subjects axis. Thus differences in the Y direction represent (year or whole-plot differences) and differences in the X direction represent depth differences.

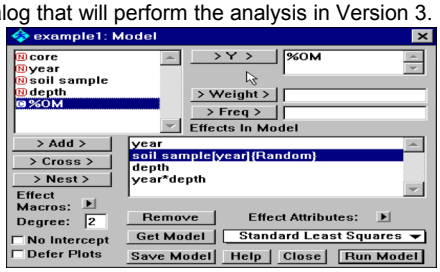


The results from 1-0 analysis on the left above indicate a 0.41 %OM significant increased in the organic matter content 1-year field (when averaged across depth). But The Across Groups Section suggests that there is these differences are significantly different for each depth (Mean Differences P=0.0053). In fact the interaction of the direction type since 1 vs. 0 years difference was -0.18% in the top 10 cm but 1.03 and .38 for the two subsequent depth. We did observe similar interactions from the pair of the12 and 30 years fields.

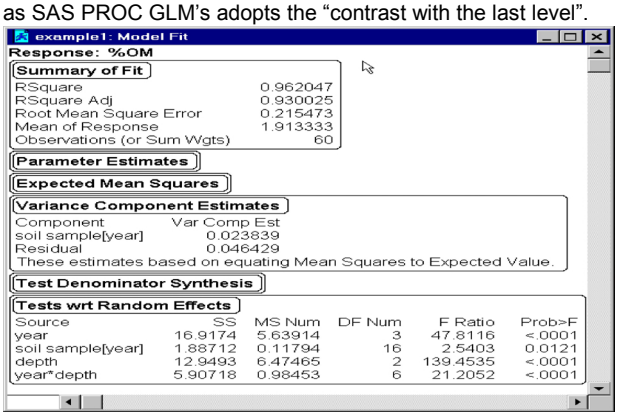
MIXED MODEL ANALYSIS (REML)

In JMP 3 there were some main issues with continuous terms the fitting tradition didn't address until now. First, if you crossed continuous terms with other terms, the tests on the main effects involved with the cross were testing hypotheses that were not meaningful. Second, the parameter estimates you obtained from continuous effects could not be judged for Effect size because they were scaled relative to the scaling of the continuous term. Lastly and most importantly because it applies in our example analysis there was not facility for producing multiple comparisons among the least squares means and only some of the contrast perform calculated appropriate standard errors for differences in the case of mixed model. In the case of the split-plot analysis example only contrast among depths for the same field (year) were done correctly among all possible interaction contrasts. These concerns were addressed with the new release. JMP now offers two methods for fitting models with random effects Method-of-Moments (Expected Mean Squares), and REML (REstricted Maximum Likelihood).

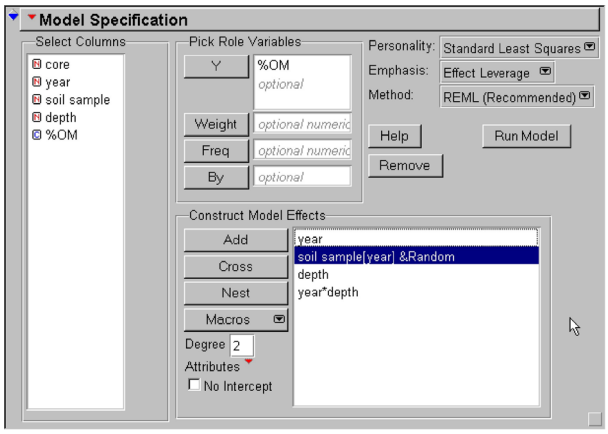
The Method-of-Moments is the way that is described in most statistics textbooks. If you need to match the method that most people have been taught, then this is the choice. JMP 3 handles Random effects like the SAS GLM procedure with a Random statement and the Test option. But JMP 4 by default and correctly so uses the REML approach as in PROC MIXED that leads to correct calculations of standard error. Below see the Fit model dialog that will perform the analysis in Version 3.



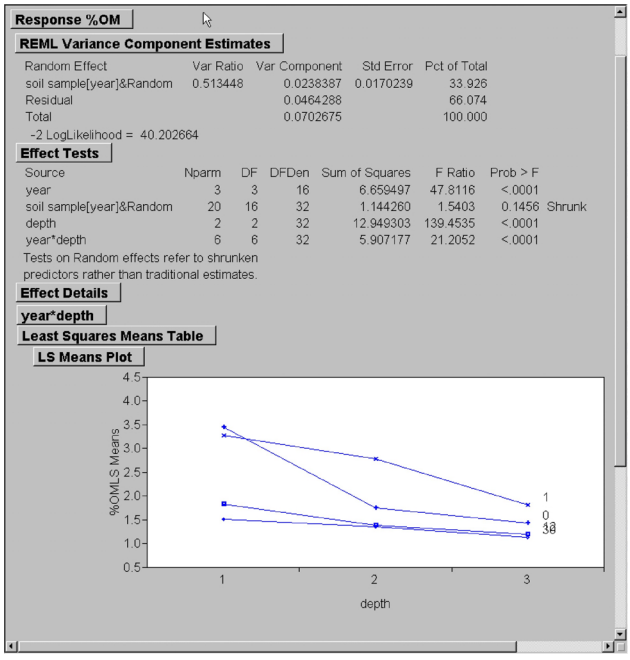
The results for the Method of Moments estimates of the variances components and the synthesized tests for each effect using Satterwaite's method are given below. It is also worth noting that In version 3, JMP labeled parameters for categorical effects in a way that was misleading to many. It labeled by the design column rather than by the meaning of the parameter, e.g. depth[1-3] was the difference between %OM in depth 1 (0-5 cm) to the overall average %OM in the 3 depths (0-15cm), rather than the difference between the depth 1 (0-5cm) and depth 3 (10-15cm). This has been corrected in Version 4; it's labeled just with the level, e.g. depth[1]. Also remember that JMP 3 parameter estimates below correspond to "zero sum parameterization" were as SAS PROC GLM's adopts the "contrast with the last level".



It turns out that in balanced designs, the REML F test values will be the same as with the Method of Moments (Expected Means Squares) approach. The degrees of freedom could differ in some cases. There are a number of methods of obtaining the degrees of freedom for REML F tests; the one that JMP uses is the smallest degrees of freedom associated with a containing effect (which corresponds to the option DDFM=CONTAIN of the MODEL statement in PROC MIXED). The Fit Model Dialog for version 4 analysis of the data is shown below.



JMP 4 provided the same P values for testing both main- plot (year) and sub-plot (depth) main effects and their interaction. A profile plot of confirms that although there is are significant differences in %OM due to year in cultivation and depth there is also an significant interaction.

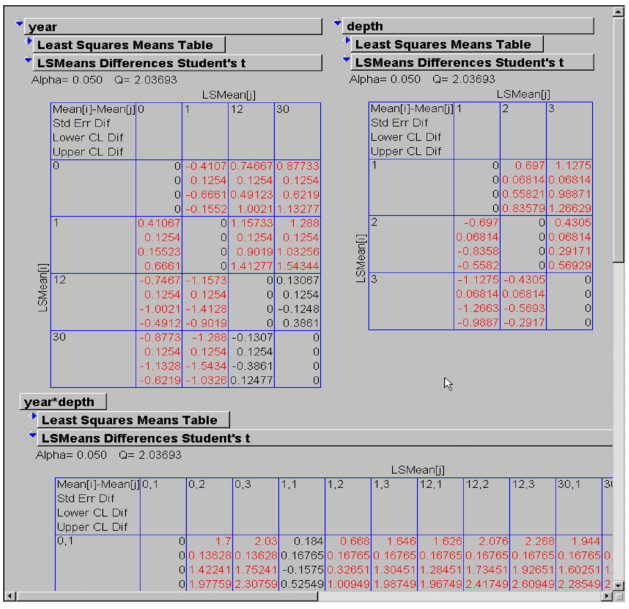


Where the methods disagree will be in the effect tests for random effects themselves. Here they will disagree by much, since the REML estimates are shrunken and the traditional estimates are not. So are the new tests right? What do they test? These tests are valid tests based on the Henderson (1984) model framework, but for a statistical hypothesis that is not interesting: that the effect sizes of the levels you randomly drew from the population happened to all be zero. Because you are interested in the population of levels, rather than just in the levels you happened to draw, you should be looking at the variance component instead, with its confidence interval. Effect sizes that are all zero in your sample would happen from a population with a variance component of zero, or by a very rare event that would bring the variance estimates to zero. These shrunken effect tests on random effects will always be smaller and less significant than the old tests based on fixed effect estimates and corrected later.

LSMEANS COMPARISSONS

The biggest reason that we recommend the REML approach is that it gets correct answers. For example, when you have contrasts across interaction effects, different comparisons have different variances. REML is the straightforward way to get them all correct. It is worth mentioning here that there are two approaches, two different statistical traditions for parameterizing the variance components: the unrestricted and the restricted approaches. JMP 4 and SAS use the unrestricted approach. A good reference that explains both sides is Cobb (1998) section 13.3. Though variances are always positive, it is possible to have a situation where the unbiased estimate of the variance is negative. This happens in experiments when an effect is very weak, and by chance the resulting data causes the estimate to be negative. Our advise is in those cases to also try to use the method of moments AOV results or go back into to the model dialog and remove the term that was corresponding to the random component and refit. JMP reports these negative estimates, and you treat them as if they were zero. The REML method in an attempt to prevent the estimates from going negative, sometimes convergences to a solution doesn't occur in a reasonable number of iterations and hangs near the boundary of zero.

In Version 3, the journal was an append-only word processing window. The new version's journal is made of the same display stuff as the original report. You can save it. You can reload it. You can extend it. And you can edit it in a variety of ways. Layout is an alternative way to collect and edit output, behaving more like a drawing package, rather than as an outliner. The Layout command copies all or selected areas of output into a new window. The layers in the output can be ungrouped and then dragged around to the desired arrangement. An example is shown below with an LSD comparison for both main effects and two-factor interaction. All possible pairwise simple effects because of the significant interactions can de displayed as shown in the layout all the possible main effects comparisons for both units below the main effects. The standard errors of differences in years for the same depth or different depths reflect correctly both error variances and not just the sub-plot error.



The researcher main question can now be correctly addressed in terms of differences in the %OM content of the topsoil (0-5 cm depth) among the fields that have been cultivated for various durations. Thus we can say that there was a significant decrease in top 5 cm % OM as years in cultivation. This decrease was not statistically significant between the prairie and the 1-year field. In fact in two depths below there a significant increase from 0 to 1-year followed by a significant decrease from 1 to 12. For the same two lower depths the changes were insignificant between the 12 and 30-year fields.

We wished that JMP will include and option for reversing the screen output formats for a blank and white printing so that the significant differences stand out instead of the insignificant in a piece of paper. We rather see the options for choosing among confidence intervals and P-values for each pairwise difference. Also we hope that eventually JMP 4 will incorporate a general Satterthwaite approximation for the denominator degrees of freedom. We also don't know why the –2 log likelihood reported by JMP is different that the one in PROC MIXED although the variance components and standard errors for the interaction contrast do agree.

EXAMPLE 2

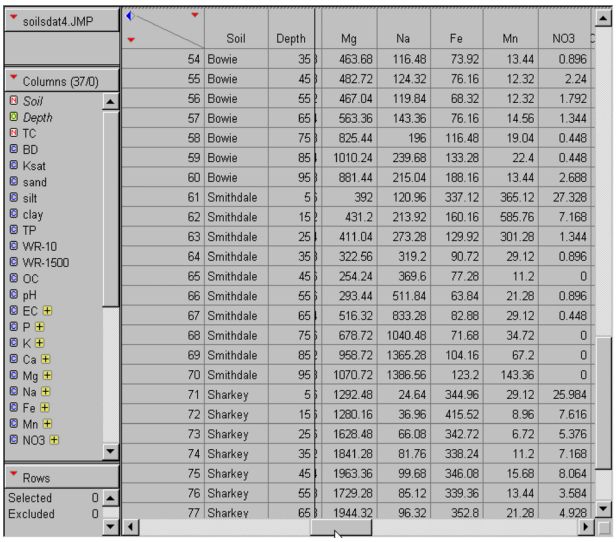
The physical and chemical properties, morphological description in eight selected soil profiles in Arkansas were determined in 10 cm increments to the 1.0 m depth. Soil physical properties determined included saturated hydraulic conductivity, water retention, bulk density, particle size and organic carbon content. Soil chemical properties included pH, electrical conductivity, and elemental concentrations of NO_3-N, P, Ca, Mg, K, Na, and Fe. Statistical relationships between several of these properties were determined. Soils are natural resources that develop from the interactions of the five soil forming factors: climate, parent material, topography, vegetation and time. Optimum use and productivity of soils for land use management practices, storage for water and chemicals, and water use and scheduling systems requires knowledge of soil characteristics. Development of quantitative information on soil properties, however, is both time consuming and expensive. The objectives of this work were to characterize selected dominant soils in Arkansas for their physical and chemical properties, morphological description, and characteristics of transport of a conservative tracer. Statistical relationships between several of the soil properties in the profiles were developed.

DATA TABLE WINDOW SIDE PANEL

A partial view of the data on the data table below illustrates the basic improvements on the data window. Because this dataset contains a large number of variables it helps point out some of these changes. It has a left border area with three panels showing table attributes, columns, and row information.

The tables panel shows the name of the table, and the table variables and table properties, which are new. Table variables are for storing context data, for example the source where the data is from, or other notes. Table properties are for storing metadata and scripts. The columns panel is for showing the columns with some icons showing the modeling type. Users with hundreds of columns will appreciate being able to see dozens of column names, rather than just the few that would fit across the top of the spreadsheet grid. The rows area shows the number of rows and the number selected, excluded, etc. The modeling type and role icons are no longer on the column heading, but in the columns side panel. This allowed us to expand the name field so that it can wrap long names into two lines. Other data table changes that we found extremely useful that cut down significantly on scrolling thru the data window included that:

- Highlighted cells are now highlighted, e.g. in white if the background is gray.
- You can scroll-lock label columns (such as the first two variables Soil and Depth in our example) so that they stay on the left, and are not scrolled.



	Soil	Depth	Mg	Na	Fe	Mn	NO3
54	Bowie	35	463.68	116.48	73.92	13.44	0.896
55	Bowie	45	482.72	124.32	76.16	12.32	2.24
56	Bowie	55	467.04	119.84	68.32	12.32	1.792
57	Bowie	65	563.36	143.36	76.16	14.56	1.344
58	Bowie	75	825.44	196	116.48	19.04	0.448
59	Bowie	85	1010.24	239.68	133.28	22.4	0.448
60	Bowie	95	881.44	215.04	188.16	13.44	2.688
61	Smithdale	5	392	120.96	337.12	365.12	27.328
62	Smithdale	15	431.2	213.92	160.16	585.76	7.168
63	Smithdale	25	411.04	273.28	129.92	301.28	1.344
64	Smithdale	35	322.56	319.2	90.72	29.12	0.896
65	Smithdale	45	254.24	369.6	77.28	11.2	0
66	Smithdale	55	293.44	511.84	63.84	21.28	0.896
67	Smithdale	65	516.32	833.28	82.88	29.12	0.448
68	Smithdale	75	678.72	1040.48	71.68	34.72	0
69	Smithdale	85	958.72	1365.28	104.16	67.2	0
70	Smithdale	95	1070.72	1386.56	123.2	143.36	0
71	Sharkey	5	1292.48	24.64	344.96	29.12	25.984
72	Sharkey	15	1280.16	36.96	415.52	8.96	7.616
73	Sharkey	25	1628.48	66.08	342.72	6.72	5.376
74	Sharkey	35	1841.28	81.76	338.24	11.2	7.168
75	Sharkey	45	1963.36	99.68	346.08	15.68	8.064
76	Sharkey	55	1729.28	85.12	339.36	13.44	3.584
77	Sharkey	65	1944.32	96.32	352.8	21.28	4.928

ACCESSING AND IMPORTING DATA

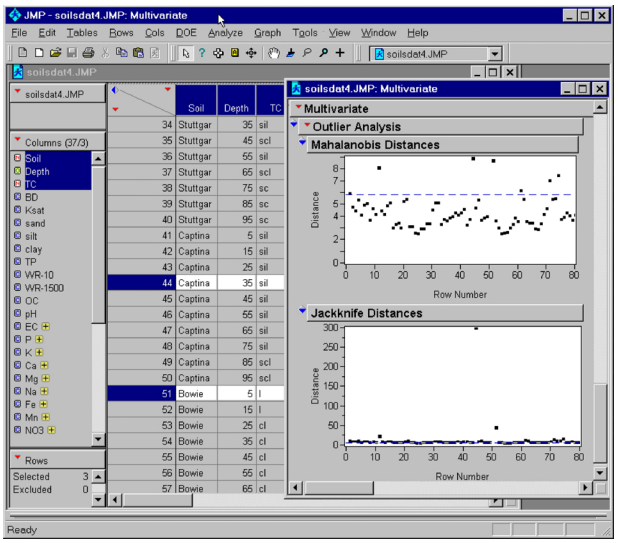
In Version 3, there were separate menu items for importing data, compared with opening a JMP data file. In the new version there is just one *open* menu item, but the dialog lets you open the file in many ways. If you choose to open a data table from text, then the dialog expands into a previewer where you can set up how it imports. By comparison, the version 3 imports were blind and primitive. The Open command can open Excel such as the above file provided to us. In the old version you could have cut and past into a blank open table but you still needed to enter all the variables names replacing the generic Colum1 – Column N.

BY GROUPS

One of the top five improvements and least confusing thing in the new version is that by groups are created in the launch dialogs in a separate By list. This is exactly like it is done in SAS, although in JMP the data does not need to be sorted first. In Version 4, all the analyses come out in one window, with an outline item for each value so they can be printed or send to the journal with a single command.

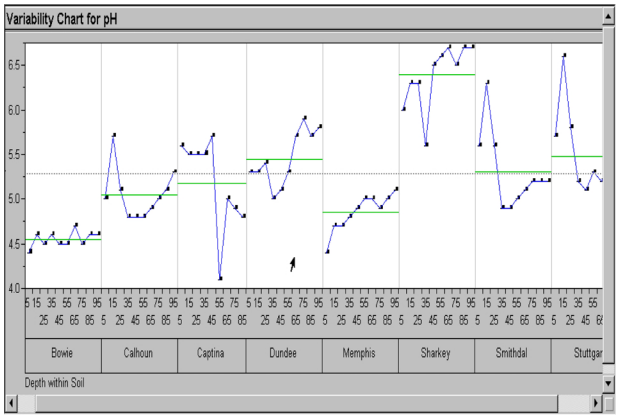
MULTIVARIATE DATA AND OUTLIER EXPLORATION

We used multivariate outlier utilizing the outlier analysis option of the Multivariate platform (old correlations) to help identify troubling data points that can be necessarily spotted in the one-dimensional box plots and two-dimensional Y by X views. Those points really stand out in the Spinning Plot multi-dimensional views as points far away from the centroid of the data.

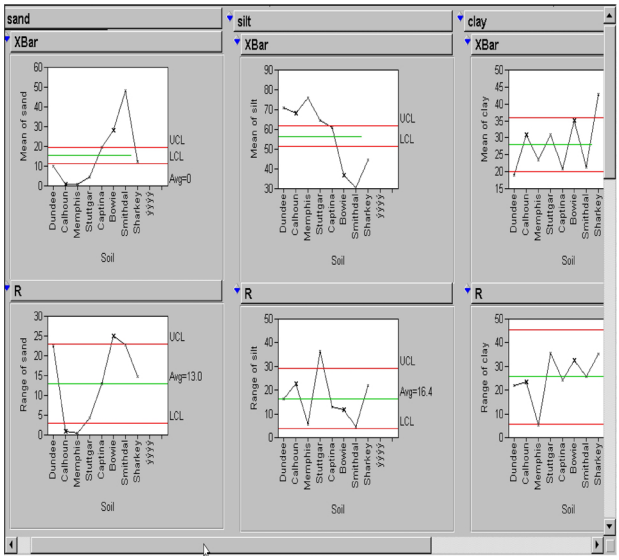


SUMMARY VIEWS (CONTROL AND VARIABILITY CHARTS)

Many users have not discovered how useful are the summary statistics that the Control Chart platform and especially the variability chart platform. A the entire dataset for pH across the 1 m soil profile for each of the eight soils can be examined and the variability by specifying ph as the Y response and Soil and Depth as the X grouping respectively. We chose to customize by the default chart by option selecting to (Connect cell means, Show group means and Show grand mean). The chart help identify Bowie as the most acidic soil with an average pH of about 4.5 and the most stable as far as pH concern across the 1 m profile depth.

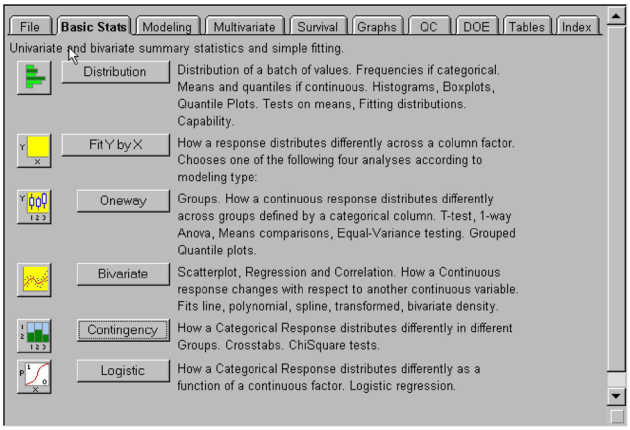


The control chart below also can help us better visualize our data and is used to identify the composition of each soil as a percent of Sand Silt and Clay. Selecting Sand, Silt and Clay as Process variables, Soil as the Sample Label and checking the Sample Grouped by Sample Label produces the control charts. We leave the rest of the options about the chart type and K sigma at their default values since the 10 values for each soil depth profile are considered a small sample so the Xbar chart will be useful to control process location and the R chart to control process variability. It is good to remind users here that there is an option in the Xbar chart under Chart options to select a Box Chart. This is the other not so obvious way of creating side by side box plots for comparing continuous distributions other that in the Fit Y by X platform. It is very easy to observe in the graph below that the first 4 soils Dundee, Calhoon, Memphis and Stuttgard in the data table are mainly "silty soils" Bowie and Smithdale (the two before the last in the graph below) are "very sandy" and Sharkey is ver high on clay.

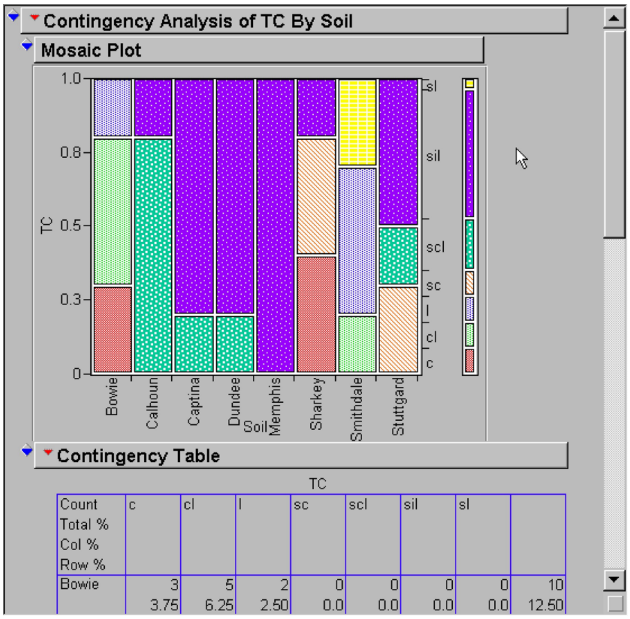


CORRESPONDENCE ANALYSIS PLOT

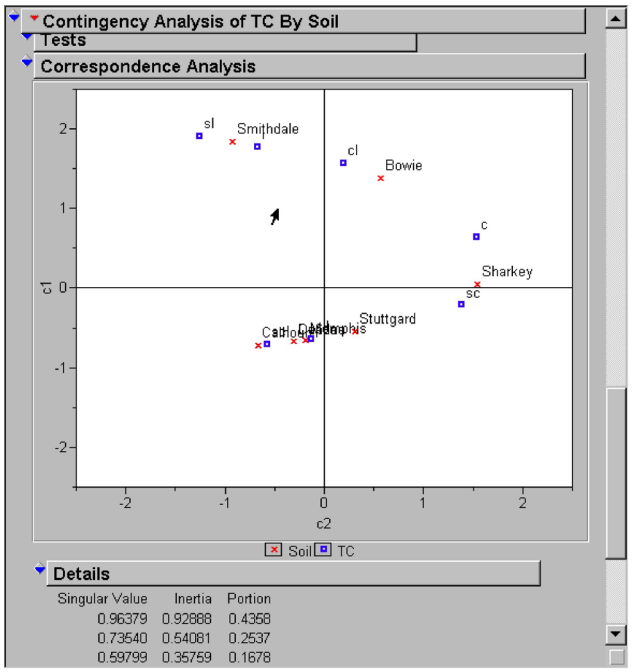
The relationships among the nominal variable for the eight different soils (Soils: Dundee, Calhoon Dundee, Memphis, Stuttgard, Captina, Bowie, Smithdale and Sharkey) and texture class (TC: with 7 levels; "sil" for silt, "scl" for silt clay loam, "l" for loam, "cl" for clay loam, "c" for clay and "sl" for sandy loam) can be explored in Fit Y by X that use to have four different personalities in the previous version depending on the types of variables. Navigation was made a lot easier in the new version and the result is called the JMP Starter. This is essentially the launch Menu items from the menu bar, but in a greatly expanded and helpful form. The JMP Starter is organized slightly differently from the Main menu. The main difference is that analyses that appear in one slot in the main menu sometimes appear as two choices in the JMP Starter. Fit Y by X becomes five buttons: Generic Fit Y by X, Oneway, Bivariate, Contingency, and Logistic as sheen below.



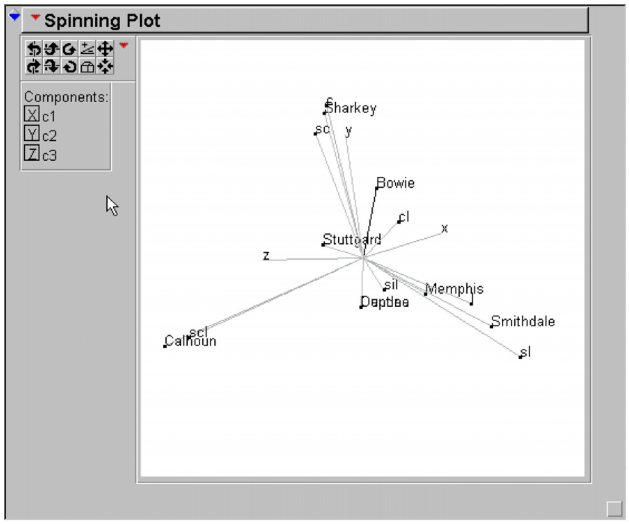
The results of the Fit Y by X with Soil as X and TC as Y follow. You can see from the mosaic plot below that the soil texture districations for each of the eight soils are not homogeneous. The most extreme case been Memphis that is 100% silt loam.



Correspondence analysis is a graphical technique to show which rows (Soils) or columns (TC) of the frequency table above have similar patterns of counts. In the correspondence analysis plot there is a point for each soil and each texture class follow.

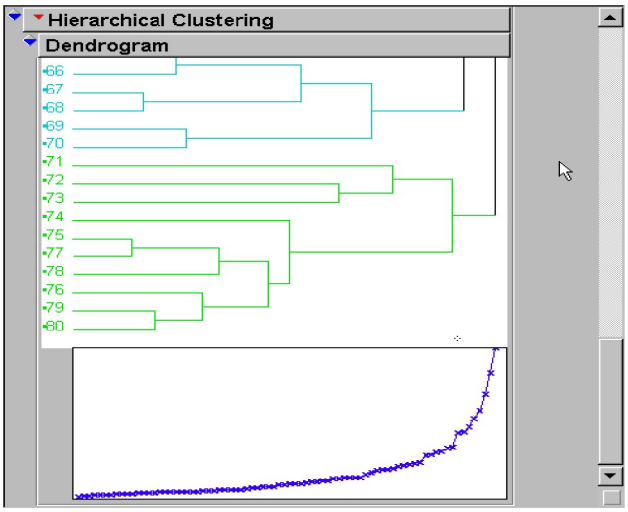


Row and column profiles are associated with each other if their points fall in approximately the same direction away from the origin and are located in approximately the same region as seen by Calhoun, Dandee and Memphis soils. In correspondence analysis, a variation of principal components for categorical data is performed on the row and column profiles of the contingency table. The details section of the plot suggests that the first second and third dimension accounts for 87% of the total inertia (measure of variation of the row and column profiles). The right mouse click in the new version when pointed on the principal components of the details section allows the option of saving the dimensions in a new table (Make Into Data Table). After repeating it for both sets of rows and columns and concatenating the tables after renaming the first variable as ID (use it as label). Now utilizing the Spinning plot graph we can display the results nicely as shown below where you can still see the relationships among Dandee, Smithdale and Memphis not seen in 2D.



DENDOGRAM

Clustering is a technique of grouping rows together that shares similar values across a number of variables. It is a wonderful exploratory technique to help you understand the clumping structure of your data. JMP provides three different clustering methods: (hierarchical for smaller tables such as ours, and k-means and normal mixtures). The expectation maximization (EM) algorithm is used to obtain estimates for k-means and "new" normal mixtures clustering. After the clustering process is complete, you can save the cluster assignments to the data table or use them to set colors and markers for the rows. Hierarchical clustering based on 9 physical soil properties BD--OC in the table) and 10 chemical properties (pH--NO3) that is 19 total Y's created the following dendogram below. Also note here that the scree plot in the bottom suggests nicely the need for 8 clusters and in fact clusters all the Sharkey's in the same cluster.



CONCLUSION

JMP 4 has come a long way from just being an excellent exploratory discovery tool as seen by the exhaustive list of significant improvements listed in the introduction. The software has changed for the better not only in its functionality that is already impressive but also in its capability. In the context of the two environmental examples we attempted to illustrate and address (some way) the improvements. We would like to have more space to provide the steps by steps commands and screen captures of the dialog boxes that produce this analysis but we run out of space. Unfortunately we did not have space to discussed a third example that includes a lot more variables than example two and it calls for the use of JSL to code Spatial statistics macros for variography, kriging and mining operations. Version 4 is definitely statistically significantly better than Version 3 and should be on most researchers toolkit

REFERENCES

Cobb, G. W. 1998. *Introduction to Design and Analysis of Experiments*. Springer-Verlag New York, Inc.

ACKNOWLEDGMENTS

Acknowledgments to Dr. H. Don Scott, University Professor of the Department of Crop Soils and Environmental Sciences of the University of Arkansas for the use of his data.

CONTACT INFORMATION

Author Name: Andy Mauromoustakos
Company: University of Arkansas
Address: Agricultural Statistics Laboratory, 101 AGRX
City State ZIP: Fayetteville, AR 72701
Work Phone: (501)575-5678 Fax: (501) 575-8643
Email: andym@comp.uark.edu