

# REVIEWS

## Order from chaos, at the click of a mouse



Have haystack, will search for needle. *Felix Grant* looks at Umetrics' latest contribution to multivariate analysis via SIMCA-P

Multivariate analysis is the new black. Everybody's doing it. Not that it is suddenly new in itself but, in just a few years (fuelled by falling RAM prices and ever rising processor speeds), it has become, first, economic, and then, the tool of choice across a wide variety of scientific work. Only comparatively recently has seeking a needle in a haystack become viable as a routine activity on any run-of-the-mill personal computer. As a label, it is slightly vague, covering an activity pursued through a number of techniques usually applied in combination. Some of those techniques pre-date the term; some have grown up under its umbrella; others again have developed in parallel.

Generic statistics software tends to emphasise the informed combination of separate multivariate techniques from the first two groups – particularly discriminant, factor and cluster analyses, canonical correlation and multiple ANOVA.

The result is a programme of surgical examination, purpose-built for each problem at hand. There is a growing market, though, for tools that move away from specifics to provide packaged examination of overarching models. This doesn't mean that the various techniques used by other packages are necessarily abandoned, but they are no longer the focus, and may be



A collage of views of SIMCA-P+ in use.

removed from the user's conscious view altogether.

SIMCA-P from Umetrics, now at release 10.0, focuses on principal components or a generalised extension (projection to latent structures, or PLS), in contiguous time series. Many of the traditional multivariate terms are absent from, or appear only tangentially in, its help file.

The SIMCA-P screen is split on opening into the now familiar arrangement of left pane tree and right pane results display. Both are resizable but also have 'hot spots' on the boundary which, if clicked, hide one pane and enlarge the other. Below both is an audit trail window, with similar hot spots. There is an option to prompt for extended user information, which makes this a rich documentation seam if properly used. Work is

organised into the equally widespread project structure, with all related information organised under a folder named for the primary dataset (that is, the dataset to be analysed).

Apart from the primary dataset, a project may contain optional secondary datasets (used for validation and prediction), 'workset' and models. A workset is a portion of the primary dataset (subset or entire), which is subject to a specific variable treatment in terms of prediction/response role, scaling, transformation, lagging, and so on. Once a workset has been defined, a corresponding unfitted model is created under the same name; this also happens if the active model type of a fitted model is changed.

The variant reviewed was the more powerful SIMCA-P+. Although the most significant

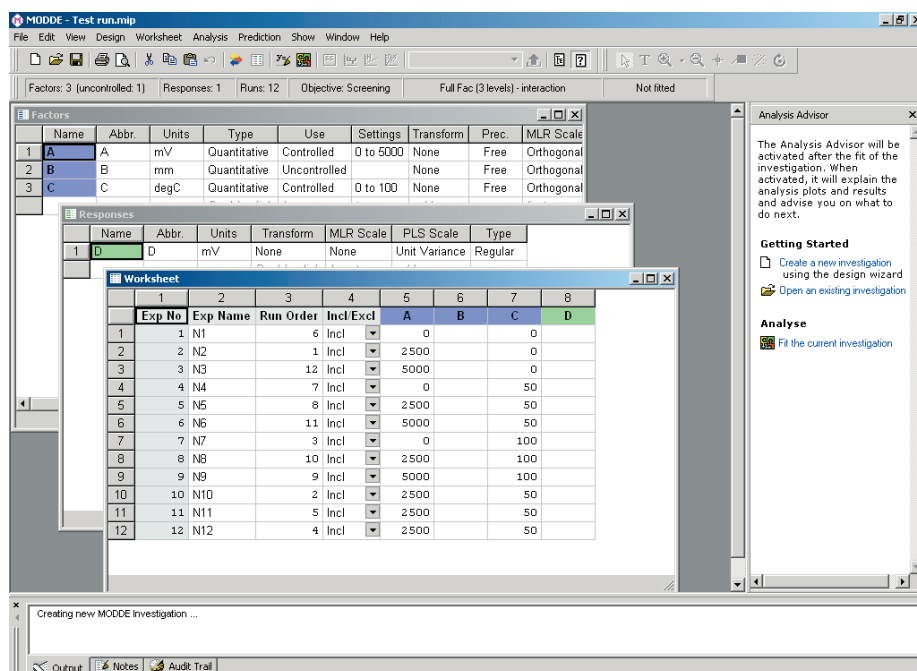
## A la MODDE

Alongside rationally planned use of multivariate analysis, in any environment within even partial control of data sampling, must inevitably come a preference for designed investigations. A second Umetrics package, MODDE (for MODelling and DEsign), covers this concern. Although a full, detailed review would be of limited value at the moment, with a new release due in September or October of this year, a quick overview seems in order.

MODDE is very much of the same family as SIMCA-P, with similar organisation and, if anything, even easier to use. That last is not something you can often say of experimental design software, but it's true in this case. Experimental plans are organised into investigations, which behave much like SIMCA-P projects, although the content reflects the different purpose; factors, responses, constraints, inclusions, candidate set, model, design, worksheet, analysis and predictions.

On opening, there is an analysis advisor, which remains on-call throughout the process. The MODDE window consists of a command menu bar, toolbars, and a model status line. A command menu handles factors, responses, objective specification, and generation of model, design, and worksheet. After entry of response values into a worksheet, a prediction menu covers analysis and display. As with SIMCA-P, easy and rapid click routes lead to plots of lists representing investigation components. A selection of 'fast buttons' handles everything, or as you get to know your way around, right-click context pop-ups are more convenient still. The investigation is summarised on the status line, with a display of fit method, objective, model, design, and the number of factors, responses, and runs.

It goes without saying that any program of this type that gets as far as being reviewed will be capable, effective, and reliable. As with SIMCA-P, I applied MODDE to a variety of predominantly medical case studies; it did everything well. Ergonomics and psychology, which permit transparent use, are less common, and they are what shape the distinguishing impression, which you carry away here.



A screen shot of MODDE, Umetrics' DOE software.

difference is a capacity for batch handling, the most striking feature remains its 'point and click' ease of use from the interface. Variables and cases are seen as entities and, while a conventional worksheet view of the data can be summoned up, it is neither necessary, nor is it used in most operations. By clicking menu options (or those in a right-click context pop-up, or a 'fast button'), most of what you want can be done from a highlighted object – fitting models or generating lists, plots and/or analyses. Case and variable names can contain up to 256 case sensitive characters. More than one model can be fitted at the same time, though only the currently active one is available to work on – all displayed options and statistics, at any given time, refer to this active model.

Plots and lists have their own context – sensitive right-click pop-ups, allowing seamless working, with return to the top menu rare once you've got to know the layout and philosophy. Plots can be generated from lists, and vice versa.

SIMCA-P supports import from a worksheet within an Excel file, generic formats such as DIF, and a variety of specialist ones including Brimrose, Galactic SPC, and so on. Export is more restricted, while plots can be saved in several graphic formats and lists go out as text. MultiVariate Analysis Common Data Form (MVACDF) is supported for both import and export; this is an open standard format specifically designed to hold raw data tables for multivariate classification or calibration. MVACDF (based on NetCDF) holds naming conventions and associated attributes, each observation being stored as a vector or as a 2D matrix; only one other product so far supports the format, as far as I know, but it is an idea which deserves success.

I applied SIMCA-P+ to a number of known problems, using historical data from laboratory, industrial and field contexts. Since a quick skim search of recent literature revealed more explicit applications of multivariate analysis to medicine than to anything else, the test data sets were selected with the same skew. Despite differences from the original application hosts, results were just as rigorous and successful and the quality of usability was a definite bonus. The stricter data-sampling requirements made it more at home in the more controllable environments, but field data were handled with equal success, given reasonable methodological forethought.

Batch handling (available only in SIMCA-P+, not in the base version) is a whole extra area of power and capability and worth an article in itself. A batch project comprises two or more linked SIMCA-P projects – one at observation level, holding observations for each batch, and one or more at batch level, with one batch per row. Observation-level variables are measured during batch evolution; those at batch level are scores or lateral unfolding of observation-level equivalents. Phased projects are handled by a structure of multiple worksheets, one of which holds all the PLS class models for the phases.

Once again, though, all of this is simply handled – click on the appropriate autofit and it all comes together nicely, allowing quick exploration without getting distracted by operational detail.

If you are concerned with analytical fields where the repeated extraction of major influences, not general statistical work, is of particular concern, then a dedicated multivariate tool makes serious sense. Having made that decision, productivity is closely linked to usability; and they don't come much smoother or easier than this.