Programmatuur-sectie

STAP : from STAndard Package to STatistical APpendix *)

B. Niemöller - C. van de Wijgaart

Standard Packages and Tailored Programs

Following the example of American universities, standard packages for data management and consecutive statistical processing, like BMD, OSIRIS, PSTAT and SPSS, came into usage in the Netherlands around 1970.
Among the packages SPSS is by far the most favorite one; it has been installed in the computing centers of all thirteen Dutch universities. SPSS is also available in many Government agencies and commercial institutions in the Netherlands. It is known from SPSS publications that the same situation exists in many other countries.

It is also a well-known fact that the more simple techniques of standard packages are most widely used, e.g. data manipulations, descriptive statistics and cross tabulations. More complex topics like analysis of variance and factor analyis appear to be used significantly less frequently. One of the reasons might be that the development of software packages is so time consuming that statistical methods in standard packages are almost outdated by definition. Anyhow, researchers are usually not much impressed by the methodological sophistication of standard packages, and, as a result, fall back on more up to date stand alone programs. These programs, as a rule, confront the user with other nasty problems, such as different I/O conventions, data structure, command language etc. Therefore we decided to look for a close relation between new programs and SPSS. Once this choice has been made there are two possibilities: addition of new programs to SPSS or developing your own package. The extension of SPSS with own programs would make the user even more dependent on SPSS then he already was: at any modification of SPSS the risk exists that he has to adapt his own appendices; at any new version of SPSS own additions have to be built in again.
These objections do not apply with a self written package, although one will have other problems: much more programming will have to be done, as a complete package requires for example a self designed, written and maintained "main'

---

*) This is part of a paper that has been delivered in the workshop "New Trends in Social Science Computing", ECPR Joined Sessions of Workshops; Florence, March 25-30, 1980.

frame" for internal control of the package, an input/output system, core allocation provisions and other relations with the operating system, some data management, error control, control card processing, etc. Amsterdam opted for the second solution: development of a new package, closely related to SPSS.

Statistical Appendix

The package in question is called STAP, an acronym for STatistical APpendix. STAP is meant to be an appendix to SPSS, which means: STAP tries to offer supplementary software to SPSS and does not try to be a complete, new package. This way the large overhead of developing (elsewhere available and reliable) data management and classical statistical routines can be avoided; doing so more energy can be spent on more up to date statistical methods.
Rapid obsolescence is unforeseeable for some STAP programs: some programs are rather predictable, other ones may have become obsolete already during the process. An example is the procedure FACAN, i.e. Jöreskog's 1963 factor analysis, which will have to be updated in the next STAP release.

The connection of STAP with SPSS has been realized in several respects:
- STAP uses the same control language as SPSS and, up to a certain limit, the same keywords. Examples are: definition of keywords, identifiers and operators, the control language notation, but also the definition of a "variable list", "subfiles", "missing data codes", etc.. An attempt was made to improve the language definition.
- STAP can read an SPSS system file.
- STAP can modify an SPSS system file; the result can (generally) be used as an input system file to SPSS.
- STAP contains a very limited number of data management functions, known from SPSS, namely the temporary data selections. This saves the STAP user the trouble of having to go back to SPSS too often between his STAP analyses.

STAP lacks the possibility of reading (raw) data; SPSS has to be used to put these data in a system file. However, STAP can modify the data matrix in the system file, e.g. in adding factor scores to the data. As other packages, like BMDP, are able to produce an SPSS system file, these packages can supply STAP data as well.

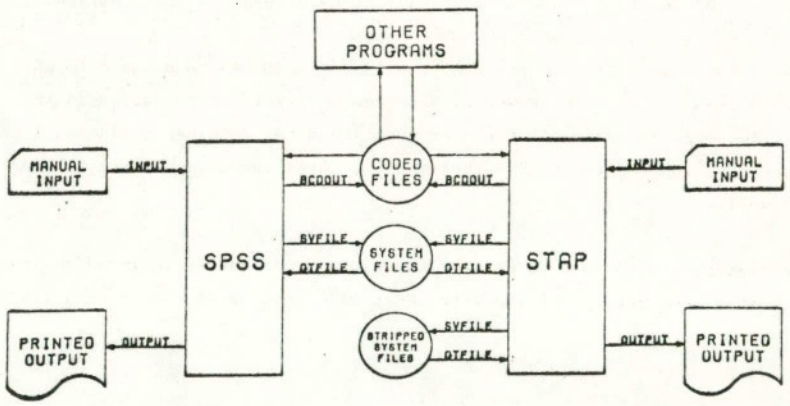
Another feature of STAP has to do with the kind of procedures included in the package, that are largely multivariate. Most STAP procedures have possibilities

for matrix input and output. In some cases the number of matrices involved is rather large, especially when using subfiles. In this respect STAP provides an extension of the system file with those matrices, including occasional additional vectors, like means and communalities. The addition of matrices to the system file does not disturb its usability as an SPSS system file: in fact the added matrices form one record that can be skipped easily by a skip-record command on the job control level, if this file is used again in SPSS.
The possibility, kwnown in SPSS, to read in and write out card image matrices has been retained in STAP. Besides this, STAP can store matrices in the system file and retrieve them. Some advantages are:

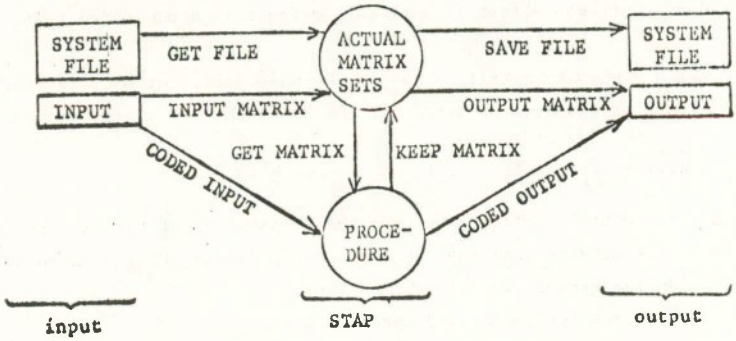
- It is possible to pass matrices from one procedure to another within one setup.
- Matrices are provided with user selected names. The function in which a matrix is used is indicated by a keyword. Also the subfile group will automatically be retained. This simplifies the user's administration and may prevent errors.
- Other information, like the size of a matrix, the row and column labels, will also automatically be stored and retrieved.
- A matrix can be used partially, selecting rows and/or columns.
- Full machine precision will be retained.

Different from SPSS, matrix transfer is obtained by means of control cards and not by means of specified "option numbers".
It is possible in STAP to delete the (raw) data in the system file so that only the matrices and dictionary remain. This type of system file can obviously not be used in SPSS; however it may speed up the use of STAP procedures.

The relation between STAP, SPSS and other programs is shown in the next diagram:



The transport of matrices within STAP is summarized in the following diagram:



In this diagram the "actual matrix sets" should be understood to be the collection of matrices "in memory" during a STAP execution. The required control cards are given alongside the arrows.

STAP procedures and documentation

The documentation is contained in seven volumes, each dedicated to more or less related subjects or procedures. In this section we will introduce each of these volumes by a brief description of the procedural contents.

Volume 1

1. Introduction to STAP. A description is given of the control language in general, similarities and differences between STAP and SPSS, the data definition and data management instructions and the use of BCD matrices.
2. STAP Matrix System Guide. The introduction into STAP of this new feature makes this chapter an indispensable one. The use of the Matrix System as an extension to the SPSS system file is discussed.
3. Summary. This chapter contains the recommended sequence of STAP control cards. Putting together all information needed for the creation of a STAP setup simplifies the use of STAP.
4. Appendix with a description of the parameters of the STAP execution command within the CDC operating system.

Volume 2

1. ASSOCIATION

   The subprogram ASSOCIATION computes a number of measures of association. These measures are

   - covariance and pearson product moment correlation r,
   - point biserial and phi-coefficient,
   - biserial correlation, the biserial phi and the tetrachoric correlation coefficient,

- multiple correlation coefficient R,
- Spearman's rank correlation coefficient.

A two-sided significance test is optional; default is a one-sided test.

2. DISTANCE

The so-called Minkowsky metric is one of the many different metrics that can be defined as measures of distance. The Minkowsky metric is defined as

$$d(\vec{x},\vec{y}) = \left[\sum_{i=1}^{r} | x_i - y_i |^p\right]^{1/p} \qquad (r \geq 1,\ p \geq 1)$$

The subprogram DISTANCE offers the user the possibility of calculating three different metrics from the large Minkowsky family. Depending on the value of p in the general formula above, they are

city block (or manhattan) metric : $p = 1$
euclidean metric : $p = 2$
dominance (or chebychev) metric : $p = \infty$

Each of these measures can be normalized in a number of ways and/or transformed into the corresponding similarity measure.

3. CLUSTER

The subprogram CLUSTER performs hierarchical cluster analysis according to one of two different methods.
The first one is the well-known method developed by Johnson (Johnson, 1967). The user can choose between two versions of the Johnson model, the minimum (or nearest neighbour or single linkage) method and the maximum (or furthest neighbour or complete linkage) method.
Starting point of these methods is a matrix with (dis)similarities. If one prefers to run the subprogram on raw data, the program can optionally calculate one of the following (dis)similarities: Pearson product moment correlations, (co)variances or euclidean distances.
The second method is the cluster analysis according to Elshout:
the cluster criterion here is reciprocity as defined by McQuitty. As this criterion is based on certain properties of the Pearson correlation coefficient, the similarity measures that are appropriate are product moment correlations and covariances.

Volume 3

1. FACAN

This subprogram for the performance of factor analysis is based on Jöreskog's dissertation (Jöreskog, 1963).

2. FACAN SCORES

As it is not possible to estimate factor scores with the subprogram FACAN, a separate procedure has been added to STAP. FACAN SCORES executes

(optionally) a complete factor analysis according to Jöreskog (see FACAN) then calculates the factor scores according to one of three methods and adds them as new variables to the system file. The options of the actual factor analysis, however, are somewhat more limited than in the FACAN procedure, as is the case with the printed output. The program is based on three well-known criteria for estimates:

- minimal differences between "true" scores and estimate scores
- the estimator has to be unbiased
- the estimator has to be structure-preserving, that is the relation between the unmeasured variables has to be preserved.

It turns out to be impossible to arrive at factor score estimates that meet all three requirements, so three different methods are offered which differ in the criteria that are met.

Two of these are well-known in literature, the regression method and the method of Bartlett. These methods, however, yield estimates which, as a rule, are not structure preserving.

Therefore a generalisation of the Anderson-Rubin estimator which is structure preserving, is part of the program as well. (See Saris c.s., 1978).

As mentioned before, the program is able to perform a complete factor analysis before the estimation of scores and therefore input of raw data is sufficient. In that case the factors are, as in FACAN, orthogonal. If the user wants correlated factors the relevant matrix with (rotated) loadings must, in addition to the raw data, be part of the input. This may result in quite a number of matrices getting involved but using the matrix system file will be of great help.

3. PROCRUS ROTATE

The subprogram PROCRUS ROTATE performs a procrustean rotation of an orthogonal (factor) matrix towards a target matrix (oblique or orthogonal).

With this program unrestricted as well as restricted procrustean transformations are possible.

As literature on this subject is quite scattered, the methodological introduction is rather extensive.

Volume 4

This manual gives a complete theoretical introduction to Mokken's model for stochastic cumulative scaling together with the documentation for the two subprograms MOKKEN SCALE and MOKKEN TEST.

Mokken's model is stochastic and more general than the deterministic Guttman model for cumulative scaling.

Because of the low methodological status of the latter method, the availa-

bility of MOKKEN SCALE and MOKKEN TEST outdates the Guttman scaling method completely. Owing to the fact that Mokken's model is not very well known so far, the theoretical introduction is self-contained and written on undergraduate level. For an enunciation at a much higher level the reader is referred to Mokken's thesis (Mokken, 1971).

Volume 5

ITEM ANALYSIS

In subprogram ITEM ANALYSIS the usual test statistics for multiple choice or attitude and personality inventory tests can be computed:
i.e. mean, standard deviation and standard error of the scores, Cronbach's alpha and the split-half reliability coefficient and the signal-noise ratio F. A histogram of the standardized test scores is provided.
The information per item consists of: the variance, item-test correlation (better: item-rest correlation) and a signal-noise ratio;
per alternative are given the frequencey, percentage and the weight of that alternative.
Optionally information per subject is supplied, consisting of a score profile, a total-test-score, labels of the items the subject missed and a subject identification.
For attitude scales the method of Lawshe and Harris is available, where in an iterative process the weights of the alternatives are estimated.

Volume 6

1. LINEAR COMB

   Although it is not impossible in SPSS to compute a sumscore over a number of variables (or factors), the COMPUTE statement in SPSS is for larger numbers of variables or standardized sumscores, a cumbersome enterprise. The subprogram LINEAR COMB computes (standardized) linear combinations of scores and adds these new variables to the system file.

2. SCATTER PLOT

   The subprogram SCATTER PLOT produces scatterdiagrams of datapoints on a plotter or a graphic display. As such the subprogram, just like SCATTERGRAM in SPSS, produces a two-dimensional graph of data points where the coordinates of the points are the values of the two variables being considered. The plots are produced on a plotter of graphic display which makes the scales, contrary to SPSS, practically continuous. The program does not calculate any statistics.

Volume 7

1. NONPAR T-TEST

   This program offers:

   - Wilcoxon's rank sum test and the test of Ansari-Bradley for indepen-

dent samples,
- Wilcoxon's symmetry test and Spearman's rank correlation coefficient for paired samples.

2. NONPAR ONEWAY

This program offers:
- the rank sum test for k independent samples according to Kruskal and Wallis,
- contrast estimates between k samples according to Spjøtvoll,
- the range test according to Steel,
- the trend test according to Jonckheere and Terpstra,
- the test for outliers according to Doornbos.

Postscript

STAP was developed and documented at the Technisch Centrum of the Social Science Department of the University of Amsterdam, as committed by the Amsterdam Academic Computing Center, and incorporates programs submitted by the Free University and the Mathematical Centre in Amsterdam.

The now developed version of STAP is suited for usage on installations of the Control Data Cyber series, and works under the operating system NOS/BE. A NOS version is being prepared. Also a conversion of STAP to IBM/370 is being carried out. It is not yet clear when this conversion will be completed.
STAP has been written in standard FORTRAN IV, augmented with CDC-assembler on places where this was inevitable, like communication with the operating system, dynamic core allocation etc. Mainly for contractual reasons the software will not be available for distribution before fall 1980.
The STAP User's Manual, that will be provided in seven volumes of coherent subjects, amounts to over 600 pages and is written in English.
In a couple of months this manual will be available.
Plans for future extensions of STAP are not yet concrete, although several proposals have been formulated. Suggestions in this respect as well as software to be built in are welcome. The main intention will be to keep STAP methodologically up to date.
Further information, like distribution conditions, will be available from the

Technisch Centrum FSW
University of Amsterdam
Roetersstraat 15
1018 WB AMSTERDAM

Literature

Johnson, S.C. (1967). Hierarchical Clustering Schemes.
Psychometrika, 32, 241-254.

Jöreskog, K.G. (1963). Statistical Estimation in Factoranalysis.
Stockholm, Almquist & Wiksell.

Saris, W.E., Pijper, M. de and Mulder, J. (1978)
Optimal Procedures for Estimation of Factor Scores
Sociological Methods & Research, 7, 85-106.

Mokken, R.J. (1971). A Theory and Procedure of Scale Analysis.
Paris/The Hague; Mouton.