The SAS system and Meta-Knowledge

Gérard HATABIAN - Hervé AUGENDRE

Abstract

As explained by Y.DODGE and D.J.HAND in a paper entitled What Should Future Statistical Software Look Like? published in the November 1991 issue of the Statistical Software Newsletter “Meta-data is information about data and can obviously be used to help the user of statistical software. More exciting is the prospect that the computer itself can access the meta-data to control the analysis in some extent.” Meta-knowledge are not only these meta-data but also the ways to use them.

These ideas have still been explored by SAS Institute in JMP®. Attributes are associated with each variable and the statistical platforms adapt themselves to the attributes. The question is Why not in the SAS System?

The existing possibilities are listed and discussed including those available in SAS/INSIGHT™ and SAS/LAB®. The SAS announcements concerning the data and SAS/EIS® will also be taken into account. This is to be balanced with what is available in other softwares including Statistical Expert Systems.

An example

Before defining Meta-knowledge, let’s have a small example to see what it is. The classical “Class” data table contains five variables: name sex age weight height for the pupils of a classroom. Suppose you want first to study the distribution of the variables and then look at the relation between two variables with the SAS system.

To study the distribution, you use the CHART (or GCHART) procedure to have an histogram. This is all right for sex and height. On the contrary, with age the six different values of the variable are ignored and some strange midpoints are used instead. You have to use the DISCRETE option to obtain a satisfactory result.

How to look at the relation between two variables depends of ... the variables themselves. You compute a frequency table (with proc FREQ) to cross sex and age. You fit a line (with REG, or GLM, or even GPLOT) to look after the effect of weight on height. One-way analysis of variance (with TTEST or GLM using a different syntax) enables to look after the effect of sex on weight. Finally the effect of age on height needs some logistic regression (with LOGISTIC) to be determined.

The two arising questions are: How to make SAS choose the good options by itself, How to make it choose the good procedure (or method).

Solving the problem

This is one of the examples of JMP®, the SAS Institute product for
the Macintosh. JMP is especially efficient to solve the problem.

In JMP, variables are associated with a type or measurement level: sex is nominal, age is ordinal, weight and height are interval.

<table>
<thead>
<tr>
<th>Nom</th>
<th>Ord</th>
<th>Nom</th>
<th>Int</th>
<th>Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>prénom</td>
<td>age</td>
<td>sexe</td>
<td>taille</td>
<td>poids</td>
</tr>
<tr>
<td>KATIE</td>
<td>12</td>
<td>F</td>
<td>59</td>
<td>95</td>
</tr>
<tr>
<td>LOUISE</td>
<td>12</td>
<td>F</td>
<td>61</td>
<td>123</td>
</tr>
<tr>
<td>JANE</td>
<td>12</td>
<td>F</td>
<td>55</td>
<td>74</td>
</tr>
<tr>
<td>JACQUES</td>
<td>12</td>
<td>F</td>
<td>66</td>
<td>145</td>
</tr>
</tbody>
</table>

You choose the analysis platform: "Distribution of Y", "Fit Y by X", ... and JMP automatically choose the good method depending on the measurement level of the variables.

**What are Statistical Meta-Knowledge**

The knowledge we used to solve the problem is meta-knowledge. A possible definition is "additional information to use the data and to be able to transform it into statistical information". This knowledge can be meta-data, strategies, or other knowledge.

**Meta-data**

Meta-data are data about data. They constitute information describing the numerical data and its properties. As explicit and accessible piece of information, they provide a guidance in choosing the meaningful statistical methods [see HAND].

Bounds on the values that a variable can take enable automatic validation. Measurement scale (nominal, ordinal, interval, ratio) determines the sort of analysis it is meaningful to apply. The knowledge of the links between variables enables cross-validation and eases imputation; this can go up to the relational data model (domains, integrity constraints, ...) [see CHURCH].

But meta-data can also be the measurement procedures or the research design. In fact, it is hard to give a limit to the scope of information that might be relevant.

**Strategies**

"A statistical strategy is a formal description of the choices, actions and decisions to be made whilst using statistical methods in the course of a study" (HAND). In fact, strategies are two-levels knowledge. They are knowledge on the generic organisation of statistical processing. And they also are knowledge of the specific statistical data processing tasks. That is to say, for a specific method, knowledge on conditions of use, assumptions, and rules for interpretation.

As an example explicit conditions of use permit either to verify them for the chosen method, or to select a meaningful method. In the
same way, assumptions have to be verified and might suggest transformation of data when they are violated.

Of course, as explained by HAND, issues of robustness to departure from the assumptions and of experience gained from practice arise here. It is putting these kinds of things into statistical expert systems which poses the real challenge.

Other meta-knowledge

Explanation ability (of the results, of the reasoning) is an important feature which rely mainly on textual meta-knowledge such as on-line documentation. Ability to interpret graphical output of many shapes and forms is also an important feature which rely on more technical and less explored knowledge.

Then, knowledge on how to use all these pieces of knowledge is required. A modelling of knowledge and an "engine" able to deal with this model are also needed. Engine can be an Expert System (Knowledge Based System), an AF Application or an other program. But to be usable, the system need some knowledge on the statistical software itself (syntax, input/output).

Existing Meta-Knowledge in the SAS System

The SAS System does contain some meta-knowledge. They are to be examined in the "classical" SAS System and in the "new" products SAS/ASSIST®, SAS/INSIGHT™, SAS/LAB® and JMP on the Macintosh.

Meta-data in the SAS System

The SAS System includes common databases meta-data. That is to say, it handles data tables described by a name, a label and a type (DATA, CORR,...). This type is under-used, and the label is not used.

A table contains variables and observations. Each variable has a name, a label and a type (NUMERIC or CHARACTER). It also may have a format, a label, ... Let's note that variables' type has been introduced only for computational reasons.

Thus, SAS contains a very limited set of metadata in which table type is the only real meta-data.

Meta-data in SAS/INSIGHT, JMP and SAS/LAB

In SAS/INSIGHT, the "Set Properties" function, available in the Data Window, enables us to define variable's default role (Weight, Frequency, Label or None), its measurement level (Nominal or Interval) and of course its name and label. Role and measurement level are real novelties in SAS but it seems that they are not stored.

These notions were already there in JMP, the Macintosh product from SAS. JMP is richer with two extra roles X and Y for independent and dependent variables and an extra measurement level Ordinal - Ratio is missing but is harder to take into account. These meta-data are stored with
the table and guide the analysis.

In SAS/LAB, detection of the measurement level - Nominal or Interval - for numerical variables is said to be automatic. A role is also attributed; in fact a nominal variable is to be an independent variable. These meta-data are stored in a specific structure (the journal), which is not a table, and guide the analysis.

So, SAS has tried to introduce some metadata, but up to now this not achieved in the SAS System.

**Strategies in the SAS System**

The SAS system does not contain real strategies as we defined them. However it does contain some interesting "behaviour". Thus, existence of default options balances the clumsiness of the language. Missing data are excluded from analysis. And these things are certainly key-features of the SAS System.

However, a same procedure may deal with numerous issues and the same issue may be dealt by numerous procedures. The defaults options are not enough. But, in its cautious position, SAS Institute just offers a few pieces of advice in the “Introduction to ...” chapters of the SAS/STAT® User’s guide. The rest of the documentation is, alas, useless in terms of strategy. On-line Helps are also useless for this purpose.

**Strategies in SAS/ASSIST, JMP and SAS/LAB**

As for meta-data, SAS Institute added some kind of strategies in its new products.

SAS/ASSIST includes knowledge of the SAS syntax but nothing more. User must choose the meaningful method and know how to read the SAS outputs and interpret them.

In JMP, automatic choice of the meaningful method has been introduced. So when one wants to “Fit Y by X”, JMP selects automatically between contingency table, one-way anova, simple linear regression or logistic regression depending of the measurement levels of the X and Y variables. Furthermore, several graphical tools have been added to help interpretation. But the clarity of the first version has been shadowed by the incoherent introduction of new methods and options.

In SAS/LAB (see TOBIAS), deep knowledge on a few methods has been introduced. In the regression and ANOVA domain, SAS/LAB checks assumptions, suggests transformation of data and provides tools to help interpretation.

So, there is some hope to find more real strategies in the SAS System in the near future.

**Explanation ability**

The SAS System does not provide explanation. It computes everything it can and lets users extract important pieces of information.
and interpret them. JMP and SAS/INSIGHT do not give more explanation but they have adapted computation and graphical aids.

On the contrary, SAS/LAB provides natural language explanation such as: "There is a strong statistical evidence that an increase in WEIGHT is associated with an increase in the expected value of HEIGHT". It also includes graphical summarization and gives suggestions for what to do next. So the user is guided even if he has to know the statistical and technical terms.

**What have been added in the SAS System by users**

Some users tried to use or even to enhance the meta-knowledge in the SAS System.

Systematic use of the available meta-data in SAS (using proc CONTENTS and FORMAT) has been studied by PODGURSKI. The new announcements of SAS during the SEUGI'91 will give new means. There are a data dictionary (SASHELP.SQL) and a meta-data manager included in SAS/EIS®. These solutions are to be tested when possible. The meta-data manager might be an excellent solution if it is available everywhere in the SAS System.

**Statistical Expert Systems (S.E.S.) using the SAS System**

Statistical Expert Systems (S.E.S.) using the SAS System have been developed by many users (see HATABIAN). Here under are listed four SES which use SAS as statistical package:

- **WAMASTEX** (cf DORDA)
  Developed in Austria for the statistical needs of clinical statisticians, it includes the strategy for a study.

- **DEXPERT** (cf LORENZEN)
  Developed at General Motors, DEXPERT is specialized in the design of experiments. It helps to create a design and then to process it.

- **DAEDALUS** (cf DARIUS)
  Developed in Belgium, it includes the strategy for a study. It is oriented on the ANOVA models. The most interesting fact is that DAEDELUS is entirely in SAS. It uses TAXSY, an inference engine written in SAS/AF's SCL.

- **ADELE/ESIA** (cf AUGENDRE)
  Developed in France for surveys' analysis, it includes a modelling of meta-data and strategies.

Note that the two first systems are said to be used.

**Existing Meta-Knowledge in other statistical softwares**

What about meta-knowledge in the statistical softwares of the PC and UNIX world? A small selection of four representative softwares
follows.

In the PC world

Statgraphics® (Uniware) offers an interface very similar to SAS/ASSIST. Concerning meta-knowledge, there is nothing more than SAS. However it includes a very simple data dictionary.

SPAD® (CISIA - France) is a software specialised in multivariate analysis. It contains a data dictionary used for validation and coding of data. Explicit measurement level is required to perform analysis and some grouping of methods are described in the documentation.

In the UNIX world

RS/EXPLORE® and RS/DISCOVER® (BBN Software) [LANE] might be seen as an interesting alternative to SAS/LAB. With them, you have a menu-driven interface, with highlighting to indicate recommended choices, an on-line glossary for statistical and technical terms, help messages that describe not only how to use menu choices but also what to look for in the output from those choices, interpretation of graphs, tables of statistics, and other types of output, recommendations on the proper analysis method, a diary to organise a record of the analysis.

S (Bell Labs) is a pure UNIX product. This statistical language is very interactive (functions) and contains an object-oriented data-modelling. In S everything is data, even the programs themselves. S is the basis of many SES and especially the most well-known REX (see GALE).

What should be added in the SAS System by SAS Institute

Many knowledge may be added to the data to help the analyst in its day-to-day work. Some already exist in the SAS System and others are announced but upgrades still need to be made.

First, the data structure is to be modified or an open data dictionary is to be provided. Are the SQL Data dictionary and the EIS meta-data first steps in this direction?

Then, the existing procedures are to be enhanced to use these new information. By the way, more comprehensive outputs would be welcome. The new OUTPUT procedure is a tool but once again, it will rely entirely on the user decisions.

An important work is also needed on the on-line documentation. The version 6 help facility is really not satisfactory. Both a more context sensitive and a more general help are needed.

Finally, the most important is to ensure the consistency of the different products and especially SAS/LAB et SAS/INSIGHT and SAS/STAT with Basic SAS.

References

AUGENDRE, H., and al., 1991 "How to achieve SES: Proposals for more general systems", 594
Preliminary Papers of the Third International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FLO

CHURCH, Lewis, 1991, “Starting SAS Data Sets for Use with the SQL Procedure”, SUGI’91


DORDA W., and al., 1990, “WAMASTEX - Heuristic guidance for statistical analysis”, Short communication of the 9th COMPSTAT, Dubrovnik, YU


LORENZEN, Thomas J. and al., 1991, “DEXPERT -- An expert system for the design of experiment”, Preliminary Papers of the Third International Workshop on Artificial Intelligence and Statistics, Fort Lauderdale, FLO

PODGURSKI, John, 1989, “A meta-database implementation strategy for documentation and retrieval using the SAS System”, SUGI’89

TOBIAS, Dr. Randall D., 1991, "Guided Data Analysis Using SAS/LAB Software", SEUGI’91

Hervé AUGENDRE & Gérard HATABIAN
EDF - Etudes et Recherches
1, Avenue du Général de Gaulle
F-92141 Clamart Cedex
Email: augendre@cli23da.edf.fr

SAS software, SAS/AF, SAS/ASSIST, SAS/EIS, SAS/LAB, SAS/STAT, JMP are registered trademarks and SAS/INSIGHT is trademark of SAS Institute Inc., Cary, NC, USA.

SPAD.N is registered trademark of CISIA, FRANCE

Statgraphics is registered trademark of STSC Inc., USA

RS/Explore and RS/Discover are registered trademarks of BBN Software Products Corporation, USA

595