

Building a Statistical Expert System with Knowledge Bases of Different Levels of Abstraction

K. M. Wittkowski, Tübingen

INTRODUCTION

Statistical analysis is predetermined by the way a (prospective) experiment is planned or data are collected in a (retrospective) study. The a-priori knowledge of observable, theoretical and hypothetical relations (WITTKOWSKI 1987) determines the semantically meaningful database activities and statistical analyses. For instance, relations between variables and types of observational units may be used to determine whether or not the meaning of a value depends on values of other variables. The models underlying the statistical methods are determined by theoretical knowledge on the sampling strategy of factors, scales, and constraints (WITTKOWSKI 1985). For confirmatory analyses, the primary goal (hypothesis) needs to be specified at the time the sample size is computed.

METHODS

Based on this classification of relevant knowledge, a statistical expert system can be logically divided into six layers according to different levels of abstraction in the knowledge: EXECUTION, ACCESS, SEMANTICS, STRATEGY, DOMAIN, DIALOG (ELLIMAN, WITTKOWSKI 1987). Each layer contains knowledge of a different area of expertise implemented using different techniques of knowledge representation.

The bottom EXECUTION layer contains data and programs, i.e. observed knowledge from observational units and algorithmic

U S E R		
dialog KB	INTERACTION	user KB
domain KB (old)	DOMAIN	domain KB (new)
model KB	STRATEGY	statistics KB
design KB	SEMANTICS	methods KB
data KB	ACCESS	programs KB
<i>data</i>	EXECUTION	programs

Fig.1 Layers of a statistical expert system (KB = knowledge base)

knowledge implemented in the code of the programs. Because we do not want to invent the wheel a second time, we use commercially available data and program management systems.

The ACCESS layer contains knowledge of how to access programs and data. In the data and program knowledge bases we have the tree of links between original and derived attributes (c.f. e.g. BLUM 1982) and the parameters and data structures required by the programs, resp.. Because there is no uniform language for calling statistical programs, the latter knowledge base consists of small transformation procedures that generate the appropriate command sequences.

The abstractions then move away to experimental designs and statistical methods. The SEMANTICS layer contains knowledge of the (lattice) structure of observational units, i.e. whether or not these units are nested or crossed both in the study design and (implicitly) in the statistical methods. The lattice structure is represented using the method of WITKOWSKI (1985).

Functional models and statistical concepts are represented in the STRATEGY layer. Functional models link the design to the reality. SI-units, level and type of scales, ranges, the distinction of factors, strata and observed variables, and the sampling strategy are represented in an object oriented approach (c.f. OLDFORD, PETERS 1986). The knowledge on statistical concepts is represented in form of production rules.

At the DOMAIN level, domain knowledge (e.g. in medicine or agriculture) is either entered or generated. Because domain knowledge is relatively independent on the other knowledge

bases, this level will not be discussed in the present paper.

In the top DIALOG layer knowledge of the user is handled and all activities concerning a given experiment are monitored. Some approaches to implement this layer have been discussed in THISTED (1986).

RESULTS

In the development of PANOS-ES, we have started with a three-layer knowledge based front end (ACCESS, SEMANTICS, STRATEGY) to commercially available statistical analysis systems (EXECUTION layer). Within each layer different activities of human experts are to some degree overtaken by the expert system.

Because the inference in the higher layers is based only on theoretical, hypothetical, and observable relations, all knowledge based on the data (e.g. homoscedasticity, Gaussian distribution of residuals, symmetry of covariance matrices) can be handled by the analysis systems at the EXECUTION level.

Commands and parameters for underlying database management and analysis systems are generated in the ACCESS layer. A knowledge based spread-sheet editor has been developed, which guarantees that only observed data may be edited and that corrections are propagated to all related observational units and to all derived attributes. Because both programm calls and all requests for updating, aggregating, and retrieving data are handled at this level (WITTKOWSKI 1988b), data structures are automatically generated according to the programs requirements.

At the SEMANTICS level appropriate data structures and methods are planned and selected, respectively. Knowledge on the lattice structure of observational units is utilized to distinguish between observational and non-observational relations, between dependent and independent replications, and between identical and replicated data (WITTKOWSKI 1988b). Knowledge on statistical methods is used to find a method (statistical model) that conforms both to the lattice structure of observational units and the hypothesis formalized in terms of influence types (WITTKOWSKI 1986). Thus it is guaranteed

that the methods chosen reflect the problem under consideration and that these methods are applied to sufficient data subsets.

In the STRATEGY layer, knowledge of statistical concepts is utilized to check consistency of the model, and to help the user formalizing primary or secondary hypotheses. Only those hypotheses will be accepted that conform to the experimental design. If the desired level of significance and size for a given difference between groups are also specified, the system can help the user in determining the sample size. At this level, the user interacts with the system through the same type of structured interface that is used in the knowledge based spread-sheet editor during input of data.

Future implementations will use DOMAIN level knowledge for further assistance of the user both for building models (selecting appropriate subsets of relevant variables) and in building designs (determining the required sample size). Multiple tests and explanation of confirmatory or exploratory results can then be handled at the DIALOG level. We expect that the interaction at these higher levels will require natural language procedures (WITTKOWSKI 1988a).

DISCUSSION

Statistical expert systems differ from other expert systems (e.g. for medical diagnosis) in that different types of knowledge have to be considered. First we have knowledge of data management and statistics. Even this knowledge has more (three) levels of abstraction than domain knowledge in many other applications. Then we have knowledge of the domain, where models are build and where results of statistical analyses are to be applied (e.g. medicine or agriculture). Thus single monolithic systems for different domains would not only be difficult to construct but it would also be inefficient to maintain and expand these systems.

The proposed approach differs from other approaches (see GALE 1986; HAUX 1986; DE ANTONI 1986; PHELBS 1987 for references) in

that decisions are primarily based on hypothetical, theoretical and observable relations. Empirical distribution of the data (observed relations) affects the inference only at the lowest level. It has been shown (WITTKOWSKI 1987) that this distinction between formal and actual relations is a necessary attribute of expert systems for testing statistical hypotheses. We have now demonstrated that this approach also allows for an effective bottom-up approach in the design of expert systems:

- Knowledge from higher levels of abstractions and formalized dialog procedures may be used to reduce the expenditure of time during knowledge acquisition.
- The implementation is facilitated, since the techniques of knowledge representation can be tuned to the special considerations of different types of knowledge.
- The path of actions in such a multi-layered system reflects the iterative process in statistical problem solving (Hand 1985). The relation among knowledge bases and stages in the problem solving process reduces the amount of rules that have to be simultaneously considered by the expert system.
- A well-defined control strategy leads not only to improved response time but also to transparent decisions. Thus the behavior of the system will become better understandable to the user and it can be verified, whether it conforms to appropriate statistical strategies.

The expert system approach presented leads to considerable reduction in the amount of information to be entered during analysis, the probability of selecting semantically meaningless data sets or inappropriate methods, and the probability of misinterpreting output of statistical analysis systems. Finally, statistical expertise is explicitly defined so that it can be discussed and misleading or wrong "heuristics" can be corrected. As a consequence the statistical strategy and thus the result of an analysis will become less dependent on the subjective opinion of a single expert. Thus application of AI-techniques may also lead to a better understanding of the concepts underlying the area of application (e.g. statistics).

REFERENCES

- BOARDMAN TJ (ed. 1986) *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*. Washington, DC: ASA
- BLUM RL (1982) *Discovery and representation of causal relationships from a large time-oriented clinical database: The RX project*. Berlin: Springer
- DEANTONI R, LAURO N, RIZZI A (eds. 1986) *COMPSTAT 86*. Heidelberg, FRG: Physica
- ELLIMAN AD, WITTKOWSKI KM (1987) The impact of expert systems on statistical database management. *Statistical Software Newsletter* **13**:14-18
- FAULBAUM F, UEHLINGER HM (eds. 1988) *Fortschritte der Statistik-Software 1*. Stuttgart, FRG: Fischer
- GALE WA (ed. 1986) *Artificial intelligence and statistics*. Reading, Mass.: Addison-Wesley
- HAND DJ (1985) More intelligent statistical software and statistical expert systems. *The American Statistician* **29**, 1-16.
- HAUX R (ed. 1986) *Statistical expert systems*. Stuttgart, FRG: Fischer
- OLDFORD RW, PETERS SC (1986) Object-oriented data representations for statistical data analysis.
In: DEANTONI R, LAURO N, RIZZI A (eds. 1986) 301-306
- PHELPS B (ed. 1987) *Interactions in artificial intelligence and statistical methods*. Aldershot, GB: Gower
- RAFANELLI M, KLENSIN J, SVENSSON P (eds. 1988) *Proceedings of the fourth international workshop on statistical and scientific database management*. (in press)
- THISTED RA (1986) Tools for data analysis management.
In: BOARDMAN TJ (ed. 1986) 152-159
- WITTKOWSKI KM (1985) *Ein Expertensystem zur Datenhaltung und Methodenauswahl für statistische Anwendungen*. Stuttgart, FRG: Dissertation (in German); Internat. Bookseller Hans Hartinger Nachf., Xantener Str. 14, D-1000 Berlin 15
- WITTKOWSKI KM (1986) An expert system for testing statistical hypotheses. In: BOARDMAN TJ (ed. 1986) 438-443
- WITTKOWSKI KM (1987) An expert System Approach for generating and testing statistical hypotheses. In: PHELPS B (ed. 1987) 45-59
- WITTKOWSKI KM (1988a) Intelligente Benutzerschnittstellen für statistische Auswertungen.
In: FAULBAUM F, UEHLINGER HM (eds. 1988)
- WITTKOWSKI KM (1988b) Knowledge based support for the management of statistical databases.
In: RAFANELLI M, KLENSIN J, SVENSSON P (eds. 1988) (in press)