STATISTICAL SOFTWARE FOR SAMPLE SURVEY DATA: DISCUSSION

Georgia Roberts, Statistics Canada
15 R.H. Coats Bldg., Ottawa, Canada, K1A 0T6. robertg@statcan.ca

OVERVIEW

I work in the Data Analysis Resource Centre in the methodology area of Statistics Canada, where one of our main objectives is the promotion of the use of acceptable methods for the analysis of survey data. Needless to say, our promotion efforts are much more likely to be effective if we can point an analyst to a software package that implements the methods that we are recommending. It was therefore a pleasure to be given the opportunity to read and discuss these four papers, thus better informing myself of what is recent and imminent in the area of software for survey data.

Statistical software is required for all steps in the survey process:

1. Pre-analysis for choosing survey design
   - how to stratify
   - how to allocate sample
2. Selection of the sample
3. Preparation of the data for use
   -edit and impute records and variables
   -adjust weights
4. "Use" of the survey data
   -descriptive population estimates
   -analytical purposes
5. Dissemination of the data and/or results

All but the first step is addressed by one or more of the papers, and the fourth step ("use" of the survey data) is a prominent feature of every paper.

Because of the diversity of topics and style of the four papers, a comprehensive discussion of them is impossible in 3 pages. Thus, the focus of this article will be a comparison of the software as described in the four papers. The following questions about each software product will be kept in mind when making the comparisons:

1. Does this software fill a gap in what is needed?
2. Is the software available now? If not, what is its state of development?
3. How wide are its applications?
4. Is it using leading-edge techniques (i.e. leading-edge for survey data)?
5. What are the possibilities for the software to expand in scope?

With these questions in mind, it is possible to categorize the software described in these papers on several different dimensions, thus coming up with some interesting dichotomies. One way to categorize the software is by its type of output - (a) what has been standard for survey software (e.g. traditional estimates of descriptive population statistics or of coefficients of common regression models); or (b) what has not been standard. Another possible categorization is by type of package - (a) commercial product; or (b) specialized product currently intended for the use of the people producing it. My cross-classification of the four software products by these two categorizations results in one product falling into each of the four cells, as shown below.

		OUTPUTS	
		Standard	Non-standard
TYPE OF PACKAGE	Commercial	SvySAS*	IntGraph
	Specialized	ExGES	SAGibbs

***SvySAS: "SAS Procedures for Analysis of Sample Survey Data"**
ExGES: "Extending GES's Capabilities via Estimating Equations"
SAGibbs: "Design consistent small area estimates using Gibbs algorithm for logistic models"
IntGraph: "Disseminating Survey Results with Interactive Graphics"

Other categorizations also come to mind. The technology and/or methods being used in each software product could be labelled as (a) standard for survey software; or (b) leading edge for survey software. On the other hand, the level of sophistication of the user (with respect to the knowledge of survey methods required to effectively make use the software) could be categorized as (a) any level; or (b) high / advanced. Again, my cross-classification of the four software products by these two categorizations results in one product per cell, as shown below.

		TECHNOLOGY / METHODS	
		Standard	Leading-edge
SURVEY SOPHISTICATION OF USER	Any level	SvySAS*	IntGraph
	High	ExGES	SAGibbs

* See table above

SPECIFIC COMMENTS AND QUESTIONS (AND BIASED OPINIONS)

"SAS Procedures for Analysis of Sample Survey Data"

My main observation is that the procedures for survey data that are described in this paper are superceded by what is available in several other commercial packages. However, since the developers of these procedures are starting from scratch and are planning to expand, they have the opportunity to fill a gap with a product that might surpass their competitors. Here are a just a few suggestions of features to consider:
(a) If the objective is to produce a commercial product that will satisfy the survey needs of a wide spectrum of users, it would be good to have integrated procedures for the full survey process. At the moment, the software contains procedures for sample selection and limited data analysis. There is nothing to assist in choice of survey design or in informative display of survey results.
(b) The software should have good variance estimation capabilities for the survey designs of its users. In the case of SAS, the users would be both those who have selected their sample through use of PROC SURVEYSELECT and those who are secondary users of data from a survey conducted by others. Of particular concern would be capabilities to handle variance estimation for WOR designs as well as WR designs, and to be able to account for weight adjustments in the variance estimation.
(c) While ease of use is a sought-after feature for any software package, there should be protection against easy abuse of accepted survey practices. Such an approach would influence such aspects of the software as the choice of default settings and the provision of warning messages.

"Extending GES's Capabilities via Estimating Equations"

The unified estimation approach described in this paper appears to be very good for computer implementation. The proposed software will certainly fill a gap in what is available since, as well as handling standard descriptive statistics and extending readily to more complex analytic uses, its strong point will be its ability to incorporate both complex nonresponse and calibration adjustments to weights in its variance estimation routines. While the software could also be useful just for producing calibrated weights for secondary data users, there is currently no commercial package that could "properly" make use of these weights.

One topic not addressed in the paper which could be useful in the software is alerting the users to the dangers of over-calibration.

Even though this is not intended as a commercial product, target dates for completion of this software were not given. Potential users will likely have to keep their ear to the ground for this.

"Design consistent small area estimates using Gibbs algorithm for logistic models"

This paper certainly describes leading-edge methodology being applied to the survey case, which is very exciting to see.

It appears that the software could be very useful for the non-survey case too, due to the speed and model-size capabilities demonstrated in the application described in the paper.

From the description provided in the paper, the software is presently limited in options when compared to such packages as BUGS or MlwiN. Do the producers have any plans for expansion?

The paper concentrated mostly on the methodology implemented in the software. It was therefore not possible to assess features such as accessibility or user-friendliness. There was also no discussion of whether there are plans to make it part of a commercial package. Potential users could find it helpful to contact the authors on these matters.

"Disseminating Survey Results with Interactive Graphics"

The software described by the author is certainly applying leading-edge technology for the dissemination of data and results from surveys. The ease of use and variety of features of the software are also very inviting.

However, it was not clear from the material that I was given to review whether design-based methods were offered to produce the statistics and analytics. Incorrect conclusions could be drawn from the outputs if various aspects of the survey design had to be ignored.

The development of informative methods for graphically displaying survey data and survey results is a current research topic. Incorporation of results of such research could be a future development for this software.

CONCLUSION

Exciting changes are taking place in the development of statistical software for survey data. The products described in the four papers reviewed will certainly contribute to the variety and quality of what is available to the producers and users of survey data..