Journal of Official Statistics, Vol. 14, No. 4, 1998, pp. 503–508

Comment

Peter Kooiman[1]

One person's noise is another person's signal
(Gary S. Brown, 1998)

1. Introduction

The study by Fienberg et al. contains two lines of thought. Sections 2, 5, and 6 deal with data swaps in cross tabulations of categorical variables, keeping certain margins intact. I consider the log linear modeling approach advocated by the authors promising; it could provide a sound statistical underpinning to such data swaps. However, in Sections 3 and 4 the authors extend their approach to a strategy for the release of survey microdata sets broadly. For this type of data release I am quite skeptical about the feasibility of the modeling strategy. Finally I draw a parallel with the *National Accounts process*.

2. Data Swapping in Cross Tabulations

The authors provide an interesting and innovative discussion of data swapping in cross tabulations of categorical variables. Cross tabulations published by statistical agencies typically involve only a few dimensions. Only when very detailed classifications are used, or populations are very skew, disclosure problems may occur in such tables. Then table cells have to be suppressed or data swaps have to be applied, moving table entries from one cell to the other. Hitherto such swapping procedures have been applied rather mechanically or deterministically. In my opinion the main virtue of the study is that it opens up a line of research which could provide sound statistical underpinnings for data swapping methodology. The idea is to first try and reduce the frequency table to be protected by searching for a more parsimonious representation through log-linear modeling. Assuming that a satisfactory model exists which is more economical than the fully saturated one, we can separate off some noise from the signal present in the frequency table. Keeping the signal intact, we can then concentrate our data swaps in the noisy part. From the point of view of subsequent analysis this is harmless, provided we apply the swaps in such a way that no artificial structure emerges where in the original table no structure existed. If the model of the frequency table can be represented as a set of marginal tables these tables contain all useful information there is in the original table, and

[1] Department of Statistical Methods, Statistics Netherlands, Voorburg, The Netherlands <pkmn@cbs.nl>.
Acknowledgments: The views expressed in this comment are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author thanks Jeroen Pannekoek and Leon Willenborg for stimulating discussions and useful remarks on an earlier version of this comment. They bear no responsibility for any of the views expressed or any of the remaining errors, though.

it is then quite natural to devise procedures which keep these tables intact. As an alternative the agency might conclude that it should revise its set of tables to be published: when all useful information is contained in a subset of marginal tables, why not publish these tables instead of the original higher-dimensional one, contaminated with uninformative noise?

In Section 4 and parts of Section 5 of their study the authors claim that the approach set out above can be extended and developed into a new strategy for the release of survey microdata files. Unfortunately it is not entirely clear to me how the two parts relate. The general strategy is phrased in terms of the conditional distribution of Y given X. Apart from this being very problematical for a statistical agency preparing a data file for general use (almost any variable can act as Y or as X, depending on the research question involved), it is at odds with the log linear modeling of frequency tables which concentrates on the full joint distribution of all table entries, i.e., $F_{Y,X}$ instead of $F_{Y/X}$. Also the general strategy nowhere mentions the problem of simulating from a data model *keeping certain margins intact*, which is at the core of the other part of the study. Indeed almost all of the technical problems arising in the data swapping part of the study are precisely attributable to the fact that we have to simulate conditionally on given margins. The authors implicitly admit the weak relationship between the two parts when they state, a few lines before their Equation (1) in Section 5, that the general strategy applied in the context of log linear modeling of categorical data sets ''seems to *suggest*, at least *heuristically*, that we should consider making draws from the exact distribution conditional on a fixed set of margins'' (italics mine). This is indeed not a very strong claim.

3. Releasing Microdata

In the remainder of this comment I concentrate on the claim that the general modeling strategy the authors present can provide a basis for the release of survey microdata files. My frame of reference is a statistical agency that purports to provide the academic community with general purpose microdata files for statistical research. The strategy consists of a modeling step, in which the agency develops a data model which is more parsimonious than the data set itself, and a simulation step in which a number of replicate pseudo microdata files are created by drawing from the exact distribution associated with the model. As I understand it, the authors have in mind a situation where a model can be obtained which on the one hand ''overfits'' the data, so that it does not distort the relevant data patterns, and, on the other hand, is economical enough to leave room for data swaps orthogonal to these data patterns.

To fix ideas let us think of a data set of 10,000 records and 6 categorical variables with 10 categories each. The fully saturated model has 10^6 cells, and clearly represents a considerable overfit. No analyst is likely to be interested in fourth or fifth order interactions; one would not even know how to interpret such effects. In practice almost all analysis concentrates on first order interactions, i.e., second moments of the data, and only incidentally on second order interactions. So, if we represent the data set by a log linear model leaving out all interactions of order three and higher, we will not lose much. This model involves 20 three-way tables with $10 \times 10 \times 10 = 1,000$ cells, accounting

for about 15,000 non-redundant restrictions on the data set. Representing each variable by 10 (0, 1)-dummies the data file contains $6 \times 10,000 = 60,000$ non-zero entries, which we can swap around a bit, provided we do not violate the 15,000 restrictions on the second order interactions. So there is some hope that we have sufficient degrees of freedom to make this a feasible exercise. At the risk of distorting the data for some subsequent analysis one might do a more thorough modeling, and throw out a number of the three-way tables, thereby increasing the degrees of freedom available for data swaps.

Survey data sets associated with the large surveys that statistical agencies conduct are much more detailed than in the example above. A typical data file may contain over 200 variables. These are recorded using very detailed classifications with hundreds or even thousands of categories: location by ZIP-code, industrial activity, profession, educational level, illnesses, causes of death in four of five digits, age in years, and so on. So, as a more typical situation to cope with, we now consider a data file with 50,000 records, and 200 variables with 25 categories per variable. Representing each variable by a set of dummies again, we now have $25 \times 200 = 5,000$ dummies. If we restrict ourselves to first order interactions only we have approximately $0.5 \times 5,000^2 = 1.25 \times 10^7$ cells, representing 1.15×10^7 non-redundant restrictions. There is no hope of keeping all of these intact with only $50,000 \times 200 = 10^7$ non-zero entries to swap around. Things are even worse when we consider a number of very detailed variables. If the file consists of 50,000 records and 10 variables with 500 categories each we have approximately 1.12×10^7 restrictions and 5×10^5 non-zero entries. If we were to include second order interactions, doing justice to the idea of some overfitting of the data, the number of restrictions would explode. With probability close to one, the only data configuration satisfying all these restrictions is the original data set and nothing else. With typical survey data files the number of variables, and the amount of detail about these variables, is such that non-distortive modeling is entirely out of scope.

Researchers are eager to obtain as much detail as they can. They consistently express their discomfort with reductions in detail statistical agencies impose in view of disclosure protection. One of the puzzles here is why researchers want so much detail. Even enormous amounts of records will not provide enough degrees of freedom to support valid statistical inference at the very fine level of detail researchers require. Once they restrict themselves to data patterns that can sensibly be investigated statistically they necessarily resort to far lower dimensional spaces using subsets of variables at far more aggregated levels. This seems to support the modeling approach sketched by the authors. Details beyond a certain level of aggregation will never contribute to valid inference, so what are we going to lose when this is replaced by noise in the sampling process of the pseudo microdata files? The answer is that researchers want to construct their own aggregates, tailor-made for the specific research questions they want to investigate. For certain studies they need age groups from 12–18, for others 17–21 is more appropriate. Having a model based on 5-year classes, 10–15, 16–20, ..., *or a pseudo microdata file representing such a model*, is not helpful to them. Similarly, they want to be able to construct their own derived variables, such as travelling distance between place of living and place of work. When we aggregate such locational variables into relatively crude indicators, researchers can no longer make such derivations. If we want to support all of these research needs, without knowing beforehand the future use of the released microdata, the only solution is to provide

as much detail as possible. I simply do not see how this could ever be accommodated within the framework of the modeling approach advocated by the authors.

It is the task of official statistics to provide society with impartial and trustworthy data reflecting the true state of society as closely as possible. These data constitute the basis for social and scientific debate and subsequent decision making. Survey data collected by statistical agencies constitute an extremely valuable resource for scientific and policy research. The number of questions that can be addressed is enormous. An evolving scientific and policy debate continuously generates new parameters of interest. It is hardly conceivable how such a rich data mass could ever be summarized in a single statistical model in an impartial way. Degrees of freedom considerations necessarily lead to a very restrictive specification. Model selection is an art, and certainly proceeds in crude ways when such masses or variables have to be analysed. Higher order interactions, representing several hundreds or thousands of individual dummy variables, are included or excluded all at once, neglecting underlying subtleties. Detailed classifications can be aggregated in numerous ways, none being uniformly superior to the others. Without a specific research question in mind there is no guidance as to which data patterns are relevant or not. The probability that two equally qualified analysts end up with the same model is close to zero. As long as this is true, a considerable amount of subjectivity cannot be avoided. As a consequence multivariate statistical modeling of large survey data sets cannot provide a foundation for the dissemination of general purpose survey data sets by a statistical agency, *by principle.*

Now, thinking the unthinkable, suppose we have obtained an unambiguously satisfactory model, i.e., one that properly represents all ''significant'' relationships in the survey data set. When we generate pseudo microdata sets by sampling from this model the information in the samples cannot be more than what was already contained in the model. Otherwise stated: an analyst will at best be able to reconstruct the model underlying the data generation process (or some reduction thereof). If the analyst does not retrieve the true model he or she errs, he or she will end up with invalid conclusions. If the analyst does, he or she might ask why the statistical agency did not simply publish the model instead of disguising it in the form of pseudo microdata files. If the agency does publish the model, or the equivalent set of marginal tables, the knowledgeable analyst will not start analysing the pseudo microdata files at all. It is like cross-word puzzles: nice for entertainment, but not really of interest when the solution is on the back of the envelope. Following this line of thought *ad absurdum* we clearly see the enormous difficulties of the modeling approach: if really successful it would make superfluous *any* subsequent statistical analysis of the pseudo microdata sets. Thus, it necessarily assumes that statistical offices are able and qualified to extract *any useful information there is* from their survey data files. Needless to say, they are not.

The main problem with the approach, which it shares with data swapping, is that it tries to restrict disclosure protection measures to the noise in the data, thereby keeping the signal intact. Swapping noise is harmless for statistical analysis, but can help to protect individual records from re-identification by a data intruder. However, without a specific model noise is hardly defined. Aiming at a general purpose microdata file we must recognize that the only sufficient statistic for all the information that is present in a typical rich survey data set is the data set itself. Adding noise to protect such data against disclosure

necessarily distorts potentially relevant data patterns. For some analyses this may be innocent, since these do not exploit the distorted part of the data patterns. Others are inevitably affected. The alternative approach of Gouweleeuw et al. (1998) recognizes this and therefore no longer tries to keep data patterns intact. Instead it employs the known statistical distribution of the data swaps (i.e., misclassifications) to estimate the latent unperturbed frequency table. Only when we know beforehand which data patterns to concentrate upon, such as when a limited set of low dimensional tables is published from e.g., a census, is it possible to control properly for the distortion due to data swaps. It is for such limited applications that the modeling approach advocated by the authors may be appropriate, especially when it is impractical to publish the set of marginal tables equivalent to the data model employed.

An important remaining question, on which the study touches only briefly, is whether the modeling approach provides sufficient protection against disclosure. The implicit assumption seems to be that the log linear data reduction employed is sufficient to disguise the identities of the subjects underlying the whole exercise. In practice it is difficult to verify such an assumption. Indeed, it is not sufficient to check whether the marginal tables representing the model employed are safe one by one. These tables are linked through their common source, and it is the combination of the tables which matters. Jointly they define a set of admissible solutions for the underlying microdata file. When degrees of freedom are insufficient, as in one of the examples above, this set must degenerate locally (e.g., the General Motors record) or perhaps even globally into a single point, i.e., the original micro data set. So, apart from being a sufficiently rich data representation, we should add the requirement that the log linear model employed entails enough degrees of freedom to support a sufficiently broad set of admissible solutions, especially with respect to all potential identification keys. Verifying this requirement involves very hard combinatorial computations that are unfeasible given the size and the amount of detail of typical survey data sets.

This is further complicated by the release of *replicates* of the data file. By matching replicates an intruder can find clues as to which data fields in which records have been swapped or not, especially when the set of admissible solutions for a specific record is narrow. Using modern matching technology, and modest quantities of noise, almost perfect matches can be obtained, given the large numbers of variables involved (see e.g., Winkler 1998). Perhaps, such matching exercises could be used by the agency to check the safeness of the pseudo micro data files to be released.

4. National Accounts Process

The prescription, by the authors, to include all information the agency has about errors in the data in the modeling exercise, reminds me of the data integration process typically performed by National Accounts people. They try to reconcile conflicting information from several surveys, using their accounting framework as a data model. Correcting for differences in definitions of variables, and supplementing for missing subpopulations, they exploit accounting restrictions, physical demand-supply equalities, and sampling variances to construct a consistent picture of the national or regional economy. Similar accounting systems have been worked out for other phenomena: labour accounts, tourism

accounts, socio-economic and demographic accounts, environmental accounts (see e.g., Van Tuinen 1995). Typically these accounts are both prepared and published in the form of tables at an intermediate level of aggregation. Although in many cases no formal statistical procedures are applied, the resulting figures can nevertheless be conceived of as *full information (gu)estimates* based on all available evidence.

Within the general framework presented by the authors the National Accounts tables can perhaps be identified with the model from which pseudo microdata files could be generated. At Statistics Netherlands a similar idea has been discussed in a quite different context. Due to the corrections made to the primary survey data inputs in the course of the National Accounts process, National Accounts tables are not numerically consistent with tables the agency publishes from the primary survey data sets themselves. To solve this problem it has been contemplated to reweight the surveys *ex post*, taking the National Accounts outcomes as given. The formal underpinning of such a procedure was developed by Renssen and Nieuwenbroek (1997). In following this line of thought we would end up with microdata files consistent with a given set of tables, i.e., the National Accounts tables, or any other applicable accounting framework used to reconcile conflicting survey outcomes. Since we stick to the survey data itself, only adjusting the individual record weights, this obviously would not contribute to the solution of the disclosure protection problem, though.

5. References

Brown, G.S. (1998). Guest Editorial, IEEE Transactions on Antennas and Propagation, 46, 1.

Gouweleeuw, J., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.-P. (1998). Post Randomization for Statistical Disclosure Control: Theory and Implementation. Journal of Official Statistics, 14, 463–478.

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys. Journal of the American Statistical Association, 92, 368–374.

Van Tuinen, H.K. (1995). Social Indicators, Social Surveys and Integration of Social Statistics. Statistical Journal of the United Nations, 12, 379–394.

Winkler, W.E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. Paper presented at the SDP'98 conference, March 25–27, Lisbon.