Paradata or Process Modeling for Inference

Arthur Kennickell, Edward Mulrow and Fritz Scheuren
Modernization of Statistics Production Conference
Stockholm, Sweden
November 2-4, 2009

In this paper, we look at how paradata, which we define below, might be used after the completion of a survey and how such use may be constrained or enhanced by the nature of that paradata. In a companion paper at this conference[1] Francois Laflamme looked at how to better manage a survey by using paradata to intervene or at least judge the survey's operational success. There was also a presentation[2] by Mick Couper of the University of Michigan (who coined the word 'paradata,' some years ago). From Mick we heard how to use paradata adaptively to better achieve a survey's key inference objectives during execution. Such so-called adaptive approaches may be powerful, when the relationship between the available paradata and the survey goals is well understood.

Paradata are of two types: (1) process data recorded as a by-product of the work to conduct a survey (e.g., call records, interview length, missing data codes, etc.) and (2) context data that are obtained separately or with a specifically targeted effort. The first type is what LaFlamme and Couper and most other practitioners consider paradata. The second type of "paradata" is an extension of the terminology to include data not collected in the original survey but which can be linked to the sampled units and used after the fact during survey inference.[3] Although the latter type of paradata is independent of the measurement process, the former may be influenced in a variety of ways by the decisions in the measurement process that cause action to be taken. We explore a general approach for using both types, individually and in combination.

---

[1] Laflammme, F. (2009). "Data collection research using paradata at Statistics Canada." Presented at the ***Modernization of Statistics Production*** in Stockholm Sweden, November 2-4.

[2] Couper, M. and Groves, R. (2009). "Moving from prespecified to adaptive survey design." Presented at the ***Modernization of Statistics Production*** in Stockholm Sweden, November 2-4.

[3] Examples might be characteristics of sampled addresses obtained from county real estate records and the backgrounds against which they are seen, the characteristics of the agents for data collection, and other aspects of the procedure by within which measurement takes place. This can also be called extended frame information, a very important idea in the data dense we now live in.

Basic paradata inference expressions

From the inception of modern probability sampling there has been a division between sampling and nonsampling errors.[4] The early emphasis was on sampling errors and thus sample design characteristics were what received most attention.[5] Until recently this emphasis remained; but, no more; considerations of sample design are giving way to a fuller use of paradata in our sense.[6]

Ideally, paradata could be brought to bear on estimation or inference problems, in a way to accounts for the effects of the measurement process on the estimation. We know how to do this analytically already for sampling variables (e.g., weights). We still lack, though, the practice of regularly doing this for paradata more generally; hence the reason for this paper. Much of what we do with paradata remains at the descriptive stage where we use words not deeds at inference.

One way of conceptualizing the survey inference setting is in terms of the following iconic estimation equation of the random variables **Y** that we wish to make inferences about:

$$\mathbf{EY = f(X, Z) + Є}, \text{ where}$$

"EY" is, in standard notation, the expected value of the vector **Y** over repeated surveys,

"X" is a vector of subject matter variables commissioned by the client for the data collection,

"Z" is a vector of paradata which are usually under the control of the survey practitioner, and

"Є" is a measure of model error, i.e., the difference between **EY** and **f(X, Z)**.

The term "e," which does not appear in the above expected value equation, is, in the in usual notation, a statistical error (e.g., due to sampling), and such that **Ee = 0**. In general **Є** will not be such that **Є** equals 0. Unless **Є** is small and orthogonal to **f(.)** any resulting estimates of **f(X, Z)**, say by least squares, may be so biased as to be misleading

---

[4] Cochran, W. (1977). ***Sampling Techniques.*** Wiley: New York. Hansen, M., Hurwitz, W. and Madow, W. (1953) Sample Survey Methods and Theory.

[5] Scheuren, F. (2005). *"Paradata from Concept to* Completion," (2005). ***Proceedings of Statistics Canada Symposium 2005, Methodological Challenges for Future Information Needs.***

[6] Sample design variables are part of paradata, of course. And paradata are part of the still larger set of context variables that when, combined with a system of retrieval, goes now by the name "metadata."

Deming's name for **Є** is model failure. He would argue, as do we, that in the words of Box all models are flawed, but some are useful. What is new here? Formerly statisticians set out to solve an estimation problem where in place of **f(X, Z)** we used **g(X).** The expression **g(X)** assumes that the data producer is able to "throw over the wall" a dataset for which the measurement effects, including nonresponse, have been integrated out.[7] We prefer the humbler formulation **f(X, Z)** in part because it reflects our realization that, in a given analysis problem, the client researcher may have special knowledge that can be brought to bear, if he or she were given the variables **Z.** Implicitly in this approach we are envisioning an ongoing interaction between the data producer and data analyzer, where the traditional walls between roles fall away. We would go further and do so later in urging this of the community, not just its individual members

We argue that the formulation **g(X)** is not rich enough to allow us to bring our knowledge of the total process to bear, to use our "systems thinking," as Deming has advocated. If **Z** is associated with **Y** and **X**, then the inclusion of **Z** in the estimation process may add value through bias reduction (smaller absolute **Є)** or increased sampling efficiency or both. Note if **Є** remains large after modeling (with **Z** included), then we still have not understood the process very well and rethinking, including experimentation, would be warranted. The choice, too, of what to observe and measure, what **Z** looks like, is in many settings an open issue.

A very important set of questions concerns how the functional "**f**" is specified and how the components of **Z** are defined. A "non-theory" approach to **f**, such as regression[8] has value both as an exploratory device and as a test of robustness.[9] A theory based approach to **f** and the choice of **Z**, perhaps using principles of cognitive psychology or economics, could be much

---

[7] The idea that data producers are implicitly integrating out measurement concerns can be first found on the paper by Scheuren (2001) on data quality referenced in his 2006 Springer book with Herzog and Winkler, entitled ***Data Quality and Record Linkage Techniques.*** This integrating out attempt by data producers may even be a good first approximation in many settings, especially in the days when researchers were less sophisticated statistically than data producers, unlike today when the reverse is sometimes true.

[8] Scheuren, F. (2007)."Paradata Inference Applications," ***International Statistical Institute, 56th Biennial Session.*** Also "Eight Rules of Thumb for Understanding Survey Error," ***RTI International, Gertrude M. Cox Seminar Series***.

[9] Regression can be considered as an implicit Taylor-series like expression of **f**, where we do not state the function explicitly but let the data determine the number of terms in the series and their importance. While helpful for estimation and in some cases prediction the regression link between manipulation and causation is weak. This leads to a loss in credibility for data users and the risk of a new factor entering in and destabilizing understandings.

more powerful. But much work remains to be done to define such approaches. In addition, the collection of paradata is still at a "handicraft" level. Work is needed to define robust measures that are theoretically appropriate.

It is important that such **Z** measures be maximally compatible across time for a given survey and comparable across different surveys and survey organizations. Underlying this observation is the notation that we will engage ultimately in a meta-analysis across practice as one way to speed up the rate of improvement. Most likely, this evolution will have an iterative quality, to begin with -- we use available paradata to understand the inference limitation of that paradata in order to develop still better paradata or its better use(a mouthful).

The paradata now common are often generated automatically as a consequence of the technological tools needed to support the operation of a survey -- for example, the computerized contact-attempt records that are commonly used to communicate what interviewers are doing to their managers.[10] In the effort to move beyond this level to a more considered scientific approach we mention three important factors, not all of which are obvious:

- First, if the vector of paradata **Z** is to be collected successfully, it needs to be obtained in a way that only minimally burdens the process of data collection for **X**. Otherwise, the content of the **X** data is likely to be subverted in a variety of ways, subtle and not-so-subtle.
- Second, it is particularly important to consider at a deep level what potential elements of **Z** actually get recorded about what "happened." Events have many aspects, and those that are most salient (or conventional) for administrative purposes may not be the most useful for research or inference purposes.
- Finally, it is important to recognize the ways in which the process itself shapes when measurements are made and to consider how that process might be incorporated into the information measured; for example, the belief that one respondent is more "difficult" than another may increase the amount of effort (and thus the amount of process data recorded), making it seem that such respondents are the most difficult ones, when they may not be.[11]

In addition to such traditional paradata, there is the possibility of paradata sources that can be linked to sample cases, like characteristics of neighborhoods or specific sample addresses

[10] See the upcoming session on paradata to be given at the 2010 Joint Statistical Meetings in Vancouver Canada. Organized by Chun and Scheuren it is entitled "Innovative Use of Paradata across Continents in Large-scale Complex Surveys: Nonresponse Adjustment, Data Quality Control, and Theoretical Underpinnings"

[11] See, for example, Kennickell, A. B. [2000] "Asymmetric Information, Interviewer Behavior, and Unit Nonresponse," working paper http://www.federalreserve.gov/pubs/oss/oss2/papers/asa00.pdf.

(utility payments, real estate taxes, etc). In the process of experimenting to develop more informative paradata, and we need to experiment in a big way here, many approaches should be tried.[12] Sometimes simply capturing (and reading) interviewer comments, to describe context, will be needed to develop a more refined sense of the narrower measures that are appropriate, but which we may have to develop later.

Some Examples and Concerns about Paradata

It seems worthwhile to show at least a few examples of how the use of paradata can operate in an inference, rather than just an operational setting. And we have done this here.

Analysis of nonresponse using paradata has been increasingly common in light of recent emphasis on identifying and minimizing nonresponse bias in surveys. If measurable variables can be found that separately identify causal factors in nonresponse, there is the possibility of reducing any nonignorable aspect of nonresponse. Sadly, so far in our practice we have not found a strong working paradigm[13] and so cannot, despite high hopes, offer more than the usual injunction to "try harder and smarter."

As noted earlier, process data may be contaminated by the choices that lead to the generation of such data, unless the choice process can also be captured as a part of the paradata. Context data may be available from the original frame or from variables matched to the sample of eligible respondents or the context of the potential respond. Be warned , however, that while variations in nonresponse across neighborhoods is important, it certainly seems true that often intra-neighborhood variation is more important than is inter-neighborhood, suggesting that address level rather than tract level, say, might be more needed.

Characteristics of the respondent and nonrespondents can be powerful explanatory factors, but there are serious limitations in the availability of such data. Setting aside timing, cost and feasibility issues such data collection is still governed by the rules for the treatment of human

[12] The idea is to add a chapter on paradata to a good survey data capture text, like that of Dillman, *et al.* (2009), ***Internet, Mail, and Mode Surveys.*** Wiley: New York. In fact this seems overdue. An attempt to benchmark practice is underway now (Scheuren and Kupperman 2010) and supports the beginnings of such a new chapter, especially if paradata benchmarking can be made a community project for a large part of the profession.

[13] Nearly all practitioners, including each of us, have many nonresponse experiences but not yet enough to categorize and summarize them into a thorough meta-analysis study. Such a study is badly needed, however. With a community-wide effort, such a meta-analysis may be fruitful.

subjects. For example, privacy concerns may limit the amount of matching that could take place without the explicit consent of sample members, which would normally be difficult to obtain, especially from nonrespondents.

Interviewers may be able to code specific context data, as was attempted by NORC interviewers in earlier years of the Survey of Consumer Finances (SCF). However, in the SCF, such information collection was terminated because the inter-interviewer consistency of coding was too low and the act of capturing such information distracted many interviewers from their principal task, persuading respondents to participate and administering interviews to them. Despite this negative experience, there remain many possibilities for obtaining context data that have not yet been attempted.

Interviewer-level effects appear to be important in shaping the information collected in surveys. However, one obvious source of information that might be brought to bear in gauging those effects has largely been ignored—that is, the information contained in the management information systems about the personal characteristics of interviewers. Concerns about interviewer privacy are correctly raised, but it seems likely that a means can be found to address these concerns. Some work has been done using information obtained from surveys of interviewers for particular surveys.[14]

In countries where substitution is allowed in the implementation of a sample, it seems essential to develop paradata sufficient to address what would otherwise be potentially nonignorable factors at play in the field decision to reject an original sample member and instead interview a substitute. The substitute is "easier" in some sense than the original sample element; do they also differ in terms of characteristics that affect the observed distribution of the data? An additional reason to collect substantial paradata in surveys that allow substitution is to ensure that discipline can be imposed on field staff in deciding how and who to pursue as a substitute. Substitution is frequently used in some European and Asian countries, but our experience is

[14] Groves, R, M. and M. P. Couper [1996] "Contact-Level Influences in Face-to-Face Surveys," *Journal of Official Statistics*, Vol. 12, No. 1, pp. 63-83. Also Analysis of Nonresponse Effects in the 1995 Survey of Consumer Finances," ***Journal of Official Statistics***, v. 15 no. 2, 1999, pp. 283-304.

mostly limited to the experiences in Armenia and Spain.[15] A conference on this single topic might be worthwhile.

An extremely important determinant of the quality of a survey is the degree of consistent engagement there is of a person or small group of persons throughout the various stages of design and measurement. Management structures can have enormous effects on measurement. To our knowledge, the closest the profession has come to recognizing the importance of such information is the earlier literature on "house effects." For us to improve as a community of practitioners, efforts should be made to identify important aspects of management that can be examined in meta-analysis later.

[15] Mushtaq, A. and Scheuren, F. (2009) "Weight adjustment for substitution of nonrespondents in household surveys." Unpublished draft.