Martin R Frankel,
Baruch College, CUNY/SMRB, USA
Pamela Baxter
Simmons Market Research Bureau, USA

8.3 FUSION, INTEGRATION, ASCRIPTION AND IMPUTATION

INTRODUCTION

The purpose of this paper is to describe the status and future directions of data fusion, integration, ascription and imputation as it applies to syndicated US magazine audience research.

IMPUTATION AND ASCRIPTION: STATUS

Let us begin with imputation and ascription. In this paper, the terms ascription and imputation will be used interchangeably to describe procedures which attempt to compensate for item level missing data. Typically, this compensation involves the process of explicitly ascribing or imputing a data value into the data base where one does not exist.

For example, suppose we have a questionnaire with 150 items (questions). Suppose that as we review the respondent data on an individual basis we find: Respondent 1 has all items complete; Respondent 2 has item 123 missing and all others complete; Respondent 3 has all items complete except for item 88; and so on.

When it becomes time to tabulate the survey, and in particular the items that have at least one respondent value missing, we have the following three options:

(1) Throw out the case for all tabulations.

(2) Retain the case, use the category 'missing' but then re-percentage based on valid categories.

(3) Impute or ascribe a value.

At the first International Readership Symposium held in New Orleans, we reported that imputation and ascription of missing data had received only scant attention by statisticians. This situation has changed considerably. There has been a remarkable amount of interest by both applied and theoretical statisticians on the subject. Currently the standard US statistical literature (journals and proceedings) contain more than 50 papers on the subject. We are aware of at least three books published by mainstream US statistical publishers that are devoted to the subject of imputation. It is now generally recognised among survey statisticians that any of the three options mentioned above are, in fact, a form of adjustment. Thus the non-use of imputation and ascription (ie the use of the other options) is a form of adjustment. It is now generally recognised that imputation or ascription is a valid and often a most preferable method of bias reduction and control in the case of missing data.

Some of this acceptance and endorsement of imputation and ascription by statisticians has been passed along to the general US media research community through some of the community's more technically sophisticated members. In addition, there have been several public forums in which users of media research have been able more closely to examine the impact of the three missing data alternatives.

In the United States, the Advertising Research Foundation (ARF) serves as the impartial auditor of the two syndicated magazine audience research services. In its last audit of the Simmons (SMRB) service, the ARF examined the impact of ascription on audience levels and turnover rates. The negligible impact of this ascription or imputation on these levels

and rates seemed to have convinced a number of researchers that ascription and imputation were legitimate tools for minimising the impact of Phase II non-responses.

In summary, as the result of various public examinations of the impacts of ascription and imputation in the US magazine research community, the controversy surrounding these methods appear to have subsided. Instead, the focus has been shifted to the methods of fusion and integration.

FUSION AND INTEGRATION: MOTIVATION

As we approach the end of 1980's there are at least two major factors that are motivating the search for scientifically sound and credible methods to fuse or integrate survey data.

One of the major factors motivating this search is the perceived usefulness of 'single source' research information. It is generally felt that in order best to serve the needs of manufacturers and producers it is necessary to work with data bases that include very specific information about media exposure as well as product purchase/consumption. Furthermore it is felt that these data bases should be available on a representative sample basis for the entire US as well for various subnational and local markets.

The second major motivating factor for the development of fusion and integration methods is the environment in which research is produced. In the United States research producers are faced with a growing reluctance of persons to cooperate and participate in market and media research studies. Successful data collection efforts must be based on lessening respondent burden and increasing respondent incentives and motivation. It is now more difficult and more costly to collect data. Furthermore, the quantity of data that may be reasonably collected from a single respondent is less, rather than more.

Before turning to some specifics about fusion and integration these terms are more explicitly defined. In this paper, the terms fusion and integration will be used interchangeably to describe the general process of bringing a body of data from one survey over to another survey. In fusion one data set will be considered the donor data set and the other data set will be considered the recipient or host. Fusion involves attaching to each member of the recipient or host data set, a set of information from the donor data set.

SINGLE SOURCE REVOLUTION AND DATA QUALITY TRADEOFFS

In order better to understand the potential for data fusion and integration in the United States, it is necessary to examine some of the research products now available to US advertisers and their agencies.

For the past decade the two major suppliers of syndicated magazine audience estimates in the United States have supplemented this audience information with ancillary information about product purchase/use and well as exposure to television, radio and newspapers.

Both of these services offer this supplementary information for a large scale sample of about 20,000 households and persons that is representative of the entire United States. However, since the primary products of both of these services are magazine audience estimates, the specificity of the ancillary information about product purchase/consumption as well as TV, radio and newspaper exposure is not at the same level that is offered by services specialising in these measures. Very recently, several companies in the United States have begun to offer data bases that include television

viewing obtained at the household level via TV set meters and grocery product information that is obtained via scanning devices at supermarket check out counters. This household level information is very specific in the sense that product purchase information is available at the individual package level and TV viewing is available on a very specific time basis. The shortcoming of this information is that it is available at the household level only (ie TV viewing is not available at the person level) and the data bases are restricted to what is best described as convenience-volunteer samples from non-major markets.

At present, users desiring information that combines both media exposure information and product purchase/consumption must make a choice. This choice is between extensive and very specific purchase and media information on a non representative basis on the one hand, and less specific but nationally representative media and marketing information on the other. The methods of fusion and ascription may lead to products that are somewhere in a middle ground.

ALGORITHMS FOR FUSION AND INTEGRATION

Historical basis

Many of the techniques and methods that have been developed for use in imputation and ascription provide a basis for the techniques and methods of data fusion and integration. Indeed, when we examine some of the literature associated with the 'hot deck' and 'cold deck' imputation of missing data from the 1940's, we find the precursors to what is now described as data fusion or integration.

Features of the algorithms

Most algorithms for fusion and integration are proprietary. However, all systems must come to grips with the following problem.

Given two data sets D (donor) and R (recipient), each consisting of elements $(d_1, d_2, ...)$ and $(r_1, r_2, ...)$, data fusion or integration of data set D into data set R is a function F(R,D) which assigns to each element in set R an element from set D*.

Most algorithms for fusion and integration develop these functions by using some form of a distance function minimisation. More specifically, the function that links elements in data set D to elements in data set R is specifically based on the examination of Minkowski p-metrics of the form $d_{ij}(p)$ defined as:

$$d_{ij}(p) = [\ |x_{ik} - x'_{jk}|^p\ ]^{1/p}\ , p >= 1$$

In this characterisation, x_{ik} and x'_{jk} are elements of vectors of k variable values where, x_{ik} is the k^{th} variable value associated with the i^{th} element or member of the set D and x'_{jk} is the k^{th} variable value associated with the j^{th} element or member of the set R. In general, a fusion function F(D,R) will involve finding pairwise linkages between members from set D to members in set R that minimise $d_{ij}(p)$. The specifics of a particular fusion of two data sets will involve a number of decisions. Included in these decisions are:

(1) *Choice of variables on which to compute distance values* $d_{ij}(p)$. This may be the most

* For this formulation it is assumed that the linkage is not one for one, but rather directional from donor to recipient. Without loss of the ability to reverse the roles we assume that each element of the R (recipient) set has appropriate linkages from the D (donor) set while the inverse need not necessarily be defined.

critical decision that determines the efficacy and validity of a particular fusion attempt. Most fusions involve the use of demographic variables. In general these variables should be as strongly related to the set of values that are linked to the ultimate data use as the various data sets will allow. Analytic techniques that have been suggested for use in variable selection involve discriminate analysis, canonical correlation analysis, AID (Aided Interaction Detection) and Multiple Classification Analysis.

(2) *Choice of p.* There exists some literature which recommends either p = 1, the 'city block model' or p = 2 'euclidian distance'

(3) *Choice of stratification or mutually exclusive classes within which distances will be formed.* For example, in most fusion efforts males and females will be fused on a completely separate basis. In fusions involving media it is not uncommon to find fusion accomplished separately for mutually exclusive classes defined on the basis of sex, race and some form of geographic classifications.

(4) *Re-use of elements in set D.* One feature that has received some attention in both theoretical and applied literature involves control over the number of times an individual or element in one data set may be linked or fused to an individual or element in another data set. This particular issue may take on increased importance if the establishment of linkages is carried out in sequential fashion. Some of the discussion (eg marriage at first sight, with a childhood friend, by adultery, etc) is related to this issue. It is the authors' belief that global rather than local methods must be used in developing linkage search algorithms. To the extent that re-use is to be limited, the question arises whether this should be carried out on a weighted or unweighted basis. If some potential donors have relative weights of one and others have relative weights of two, should we count re-use on a weighted or unweighted basis?

(5) *Level of measurement.* Standard metrics of the type $d_{ij}(p)$ typically involve variables that have at least an interval level of measurement. However, it is often the case that surveys involve nominal or ordinal levels. For example suppose we are considering two members of set D. These two individuals are exactly alike with respect to all variables measured except for age. One of these individuals is 23 years old and the other is 33 years old.

Suppose further, that a 26 year old individual in data set R is identical to these two individuals in set D, on all variables except for age. If age information is collected and available to the nearest year, then the 23 year old and the 26 year old will be judged a closer pair than the 33 year old and the 26 year old. However, if age is collected on a categorical basis the standard age breaks of 18-24 and 25-34 would produce the opposite result. Even if age were collected using the categories 18-24, 25-29 and 30-34, the two pairs would be judged equal in terms of their distance.

(6) *Weighting of variables within the distance function.* In addition to taking respondent level weighting into account for re-use control and resolution of 'ties' weighting may be introduced in order to increase the degree to which certain variables impact on the distance function. This provides a middle ground between the use of variables to establish strata or classes across which fusion is not permitted and the zone of indifference that is implicitly created when two variables contribute the same to the distance function.

For example, suppose that two variables to be used in evaluating distance are total household income and individual employment income. Under the standard distance model described above, a difference of $1,000 in total household income would have the same impact on total distance as a difference of $1,000 in individual employment income. In some instances it may be desirable to view a difference of $1,000

in individual income to be greater than a difference in $1,000 in household income. In these cases, the use of specific weights that are variable specific can accomplish this desired result.

(7) *Control of marginals in linked data sets.* Suppose data set A is fused to data set B by finding donors from A to link recipients in B. When tabulations are carried out with data set B and its augmented information from A, it is most likely that we will find that certain marginal values (eg audience levels) that existed when data set A was tabulated on its own will not be 'exactly' preserved when they are tabulated from data set B.

In some instances this difference may be both appropriate and understandable. In other instances, these difference, which may be simple random noise, may diminish the usefulness of the final fused data set. In this latter case, it may be possible to make use of ascription and imputation methods to adjust data set B values so they match given marginals in data set A.

DESCRIPTION OF A PILOT PROJECT

During the past year Simmons Market Research Bureau and the Arbitron Ratings Company have been involved in a pilot project involving the fusion of three existing syndicated media survey products to a baseline survey of shopping habits and behaviour. In the Phoenix Arizona market a probability sample of individuals living in telephone households was asked a series of questions involving:

- Basic demographics.

- Banking services and specific banks used.

- Malls visited.

- Visits to specific stores in the following categories: drug, hardware-paint, appliance-electronics, furniture-carpet, grocery, general department.

In addition very general questions were asked to determine general levels of TV viewing and radio listening and newspaper reading.

This sample served as the R set for the data fusion of the following D sets.

- Arbitron Local TV report (1 week diary)

- Arbitron Local Radio report (1 week diary)

- Simmons Local Newspaper (2 interviews – yesterday reading).

As each D set was fused it was taken through an additional step to assure the conformation to published reports at marginal levels and within certain key cells. Since the basic demographic composition of each sample was very similar, this marginal conformation was minimal.

Evaluation of fusion

As the process of developing fusion and integration methodology progresses from its infancy stage, it is important that both potential producers and users of fused data sets recognise and agree upon the criteria against which such data sets should be judged. It is clearly inappropriate to expect that the fusion of two properly representative samples containing information sets A and B can result in exactly the same data that would result if a single set of respondents provided both data sets A and B under identical conditions. If this level of validation were demanded of fusion techniques, it would call into question all modelling that is now used to obtain reach and frequency estimates beyond either one or two interviews.

It is, however, clearly appropriate to expect that a successful fusion of data sets A and B will produce the same relevant summary estimates, statistics and projections that would result if both data sets A and B were obtained from the same respondents. Further, it is even more appropriate to expect that the decisions reached and actions taken (media selected) on the basis of the successful fusion of data sets A and B be the same as the decisions reached and actions taken on the basis of a survey which collected both A and B from the same representative sample of respondents.

ALTERNATIVE METHODS OF FUSION AND INTEGRATION

At the present time, all the complexity in data fusion involves the choice of singular best strategy of linkage. Once linkage has been established, data are attached to individual respondents as if they provided the data and existing survey software may be used for all tabulations. This simplicity in data tabulation is presently viewed as a positive feature of fusion; however the one time only nature of linkage represents a limitation as well.

If x and y represent two variables of interest in data set D and z represents variables of interest in data set R, it may be the case that the particular function F(D, R) that represents the optimal fusion for variable x to variable z may not be the same as the function F'(D,R) that represents the optimal fusion for variable y to z.

Given present technology, the appropriate strategy for fusing data sets D and R would most likely involve attempting to make use of some function F'(D,R) that incorporates features of both F(D,R) and F'(D,R). However, this function will not produce the same results as either F(D,R) or F'(D,R) by themselves.

As computer hardware and software become even more economical (on a relative basis) the most appropriate strategy for integrating two or more data bases may involve some form of 'real time' fusion. This might involve multiple fusion functions which are selected on an application specific basis or this might involve using aggregated data from both data sets. In either case, the expert system would ask the user a series of questions and help to selected the most optimal fusion based on the answers to these questions.

References

Frankel, M R (1981). Ascription in magazine audience research. New Orleans Proceedings.

Kalton, G (1983). *Compensating for missing survey data.* Ann Arbor: Institute for Social Research.

Little, R J A (1987). *Statistical analysis with missing data.* New York: John Wiley and Sons.

Madow, W (Ed) (1983). *Incomplete data in sample surveys*, volumes 1, and 2. New York: Academic Press.

Rubin, D (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley and Sons.