The design and precision of data-fusion studies

Trevor Sharot
AGB Nielsen Media Research

Fusion is the linking of two survey datasets by pairing up similar respondents and joining their data records, in order to be able to cross-analyse outputs from one survey with those from the other. Invariably, the two surveys are pre-existing rather than being designed specifically for the fusion, and their samples of respondents differ both in design and size. Depending on the particular method of fusion used, the size of the fused dataset may be the same as one of the surveys or different to both. An unresolved issue is: what is the effective sample size of the fused dataset – that is, the size of a hypothetical single-source sample that would deliver equal variances and standard errors to the fusion? This paper addresses this question and provides three main findings.

First, it is shown that the assumption of conditional independence, crucial for good fusion, also facilitates analysis and comparison of effective sample sizes and variances. Second, across the range of fusion methods and outputs examined, the effective sample size is shown to be a weighted geometric mean of the two source sample sizes and therefore lies between them; and for designers of fusion the simple (unweighted) geometric mean may be taken as a representative figure. Third, while limited validation of the geometric mean result has been performed so far, the generality of the conditions under which it was derived implies that it should have wide validity across different fusion methodologies.

Knowledge of the effective sample size in turn provides several benefits: it is a tool for designers of fusion to deliver outputs of required precision, and a tool for users to compute the standard error of outputs; this in turn permits calculation of confidence intervals and significance tests.

Introduction

Fusion is the linking of two surveys, A and B, at respondent level. In its simplest form, each respondent in A is matched with a 'similar' respondent in B and their two data records are joined together; the resulting fused

Received: 4 October 2006

dataset may then be analysed as a 'quasi-single-source' survey. Frequently, to capture both A and B data in a true single-source survey is unachievable, either on grounds of cost, respondent burden, or conflicting client interests. Fusion, then, provides the only opportunity to create a joint dataset and cross-analyse the two surveys.

Invariably the two surveys have different sample sizes. In this case one-to-one linkage is not possible: some respondents may be unused while others may be used several times ('polygamy'), particularly in the quest for good matching.

Whether or not the samples sizes are equal, it is not obvious what the *effective* sample size of the fused dataset is – that is, the size that a true single-source survey would need to be to provide equal precision (i.e. equal variances and standard errors). Nor is it known how this might depend on the particular fusion methodology used; nor, therefore, whether some methodologies are preferable to others. It is these issues that this paper addresses.

This paper uses as an illustrative example one of the most common applications of fusion in market research: the fusion of a multimedia/product usage survey with a television viewing panel, in order to provide analysis of viewing by target groups such as users of various products. Such information allows advertising agencies to plan campaigns to reach targets more efficiently than the usual method of nominating some demographic subgroup as a proxy for a target group.

Two references provide background to this paper: Jephcott and Bock (1998) provide a wide-ranging overview of fusion methodology, validation methods and the challenges facing a designer; Soong (2006) provides a comprehensive online bibliography of the literature, including many references to the associated topic of imputation of missing data.

Sampling variation

The paper uses two terms relating to sampling variation, namely 'variance' and 'effective sample size'. The *variance* of a sample-based estimate is the measure of sampling variation most amenable to analysis. The variance of a proportion based on a simple random sample of size n is governed by the well-known '$p\ q$ over n' formula (where $q = 1 - p$):

$$\mathrm{var}(p) = p.(1 - p)/n \qquad (1)$$

As an example, a rating of 20% (p = 0·2) based on a simple random sample of 1000 has variance:

var(0·2) = 0·2.0·8/1000 = 0·00016

Equation (1) is a special case of the formula for the variance of a sample mean $\overline{X}$ from a distribution with variance σ_x^2:

$$\text{var}(\overline{X}) = \sigma_x^2/n \quad (2)$$

However, the variance is not expressed in the same units as the estimate it describes; this is provided by the square root of the variance, called the 'standard error'. The 20% rating has a standard error that is √0·00016 = 0·013, or 1·3 percentage points. The standard error can be used in the computation of confidence intervals and tests of hypotheses.

Effective sample size (*neff*) is most often used to describe the impact of real-world sample designs on variance. A typical stratified, multi-stage sample of the general population delivers higher variances than a simple random sample of equal size, but such designs are invariably employed because they are more practical and less expensive. Suppose the 20% rating from such a sample of size 1000 had variance 25% greater than from a simple random sample (i.e. the variance is 0·00016.1·25 = 0·00020). We can say that the effective sample size *neff* is 1000/1·25 = 800. Equation (1) still applies providing *neff* replaces n:

var(0·2) = 0·2.0·8/800 = 0·00020

The effective sample size can be thought of as the size of a hypothetical simple random sample that would deliver the same variances as the actual sample. However, this paper is concerned with the reduction in effective sample size resulting not from the sample designs but from the fusion process. In fact, it is assumed throughout that the source surveys employ simple random sampling. This is not a restrictive assumption: if the source surveys employ complex sample designs, their effective sample sizes can first be estimated, and then used in place of the actual sample sizes in all variance formulae presented below.

A great many common outputs, including TV ratings, media reach, average issue readership and product usage or ownership, are expressible as percentages. However, for computational convenience, these will be

written as proportions (e.g. a 20% rating is equivalent to the proportion $p = 0{\cdot}2$).

An output of particular interest is the target group rating (TGR). A TGR denotes the rating of programme Y among a target group such as users of product X. When the TGR is indexed on the programme's rating among the entire population, the result is the media selectivity index (MSI). An MSI of greater than 100 means that the programme has greater appeal to the target group than to the overall population and is therefore an effective opportunity for placing advertising, especially if this comes at no cost premium.

Fusion methodologies

While a number of fusion methods have been developed, this paper examines only the two most commonly used techniques: a simple matching process, which will be called 'classical fusion', and a newer method based on the transportation algorithm borrowed from operations research (OR).

Classical fusion

One sample is nominated as the donors (D) and the other as the recipients (R). A similarity value is computed for each possible donor–recipient pair, typically based on their respective demographic and geographic characteristics and other variables such as some measures of media exposure or product usage. In a common implementation, the recipient with the closest available match is paired first, then the second best, and so on. Each recipient's data record is augmented by the donor's record to create an extended record, and these form the fused dataset.

If the donor sample is larger than the recipient sample, not all potential donors are required to complete the fusion. Conversely, if D is smaller than R, some respondents in D must be used more than once: 'polygamy'. And polygamy may be used even when D is greater than R, to allow better matching. Excessive multiple use of any donor for several recipients is clearly undesirable but it can be controlled. For example, each time a donor is used, its similarity to all remaining unpaired members of R may be artificially reduced by some amount, a 'polygamy penalty', which reduces its chance of being used again. However, there has been no guidance so far as to the effect of polygamy on estimates, nor therefore how much polygamy is acceptable.

Because recipients' data and weights are unaffected, this survey's outputs or 'currency' are preserved. However, analysis of the donated data from the fused dataset will not exactly match the original data, because the original weights are left behind, owing to unused donors or polygamy, and because the recipients' demographics replace those of the donors. This method is therefore also termed 'unconstrained' fusion.

Transportation fusion

In the transportation fusion method, both the original surveys are treated as donor surveys. The fused dataset is a construct. Each record consists of outputs donated by a member of A together with outputs from a (similar) member of B; and a new weight is calculated for each synthetic record, such that currencies from both A and B are preserved, except perhaps within demographic subgroups. This is therefore also called 'constrained fusion'.

Another difference compared to classical fusion is that all respondents from both surveys are used at least once, and many are used polygamously under the control of the algorithm. The resulting number of synthetic records is at least equal to the larger of the sample sizes. The parallel with the transportation problem in OR is described in Appendix 1.

Matching respondents

The matching process consists of:

- selecting a set of matching variables, including but not necessarily limited to demographic classifications
- computing a measure of similarity in terms of these variables between each possible A–B pair
- deciding a set of pairings exhibiting good similarity.

In both the classical and transportation methods, good matching is crucial, the key challenge facing the designer being the choice of matching variables. It is usual to define some as 'critical' variables that must match exactly – typically sex is one, so that men are not fused with women; in effect, the critical variables are interlaced to form cells and a separate fusion is performed within each cell. Other variables are matched as closely as possible; these will normally include demographics such as age and income, and some behavioural measures such as ownership or

consumption of selected products, lifestyle measures or media consumption variables such as daily hours of viewing television. The measures should be good predictors of both surveys' outputs and, ideally, they should be largely independent of each other to minimise redundancy. The minimum necessary number of matching variables should be used since it is easier to find good matches on a few measures than on a large set.

The definition of the similarity measure also matters, though perhaps not quite as much as the choice of matching variables. In practice, a dissimilarity or distance measure is usually constructed and a minimum sought. First, each matching variable's classes are assigned scores and each difference in score between the donor and recipient contributes to the measure for that variable. For example, being in the same age band contributes zero, being in adjacent classes contributes a difference of 1, and so on. The distance scores for each matching variable are then consolidated into a single measure. Options at this stage include the 'city-block' metric (using simple addition, analogous for two matching variables to the walking distance between two points in Manhattan), or distance 'as the crow flies', or its square. In all cases, some of the matching variables may be upweighted to raise their priority. A complex measure that does all this and also tries to take into account redundancy between the measures is the Mahalanobis distance (Mahalanobis 1936). However, current thinking is that such complexity probably adds little to the quality of the fusion, and this will be demonstrably true if it results in similar matches as a simpler method.

Conditional independence

The reason that good matching is important is that fusion preserves only that part of the relationship between an A–output X and a B–output Y that can be explained by the matching variables.

As an illustration, if the distributions of X and Y differ between males and females and there is no other predictor, then matching on sex is necessary and sufficient to preserve the relationship between X and Y in the fused dataset. A practical example would be where X is viewing of a football match and Y is ownership of football boots. Among adults, sex is by far the most powerful predictor of both. Matching on sex is necessary and sufficient to ensure that the fused database will exhibit high football match ratings among owners of football boots.

Formally, we require X and Y to be *conditionally independent* given the matching variables Z:

X|Z is independent of Y|Z (X given Z is independent of Y given Z)

Conditional independence is not achieved if, for example, owners of boots are even more likely to view matches than non-boot-owning men. This additional selectivity will not be exhibited by the fusion, only by single-source data. This bias[1] is called *regression to the mean*, and is further illustrated in the next section. The effect of lack of conditional independence is more fully explored by Barry (1988).

Achieving conditional independence actually requires two conditions to be met:

1. the donor and recipient data are independent given the matching variables, and
2. each donor and recipient matches perfectly on these matching variables.

Strictly, the first condition is that of conditional independence and the term will be used in this restricted sense; a fusion that meets both conditions will be called *ideal fusion*.

In an ideal fusion, the distinction between critical and matching variables disappears and all may be considered critical variables. Fusion is performed within each resulting cell and at this level all pairings are equivalent. The donated data are completely equivalent to their (unobserved) single-source counterparts except in one respect: donors are selected with differing probabilities as required to achieve good matching. Equivalence does not mean that the donated data would be exactly the same values as the unobserved single-source or recipient data, since both have a random component; merely that the two datasets are different realisations from the same distributions.

Ideal fusion not only provides unbiased outputs, free from regression to the mean, but also facilitates analysis of their variances for two reasons. First, variances resulting from a non-ideal fusion must depend on the particular fusion methodology used and the closeness of the matching

[1] Bias is an error of a different nature to variance; it is the fixed or systematic error common to repeated samples, arising from one or more imperfections in the survey methodology; whereas variance measures the difference between the surveys due mostly to selection of different respondents. Unlike variance, bias does not reduce with increasing sample size.

process. Such hard-to-quantify features stand in the way of developing any simple or general results. By contrast, in an ideal fusion, it is irrelevant how the fusion was performed, other than needing to know the probabilities of selection required for matching. It is possible to seek variance formulae that are simple in that they depend on these probabilities alone and on no other details of the fusion; and such results will apply to all fusion methodologies.

Second, freedom from bias enables straightforward comparisons of these variances, without the complication that regression to the mean itself reduces variances, as explained in the section on non-ideal fusion (see below).

Regression to the mean

Table 1 provides a simple illustration of regression to the mean when condition 1 above is met but 2 is not. There are six (monogamous) donors and six recipients. The distribution of the output variable depends only on socio-economic class (SEC). Matching was attempted on sex as well as SEC, with perfect matching on sex but not SEC.

Among the population and these donors the mean is 6 for ABC and 2 for DE. But in the fused dataset the mean is 4·67 for ABC and 3·33 for DE. Both values have regressed towards 4, the overall mean.

Table 1 may be reinterpreted also to illustrate meeting condition 2 but not condition 1. If it arose from matching on sex alone, then perfect matching was achieved but not conditional independence, again producing bias.

Regression to the mean has been detected in several actual fusions by means of 'folding' (see Jephcott & Bock 1998). In the case of media selectivity indices, its effect is to drive the indices towards 100. Provided it is not too severe, this may be considered a conservative and therefore acceptable error, since the true MSI will be at least as big as its fusion-based estimate.

Table 1 Regression to the mean

Donor	SEC	Value	Marriages	Recipient	SEC
1	ABC	5	→	i	ABC
2	ABC	8	→	ii	ABC
3	ABC	5	⤫	iii	ABC
4	DE	1	⤫	iv	DE
5	DE	1	→	v	DE
6	DE	4	→	vi	DE

Effective sample size

If classical fusion is used, the fused dataset is the same size (in terms of number of records) as the recipient survey. However, it is not obvious what its effective sample size *neff* is – that is, the sample size that should be used to calculate the standard errors of outputs. For example, if 2000 donors are fused with 1000 recipients, is the effective sample equal to, larger or smaller than 1000? What if there are 2000 recipients and 1000 donors used polygamously?

Using transportation, the number of virtual respondents in the synthetic dataset is at least as large as the larger sample, but *neff* is presumably not as large.

Knowledge of the effective sample size is valuable for many reasons. For fusion designers, it would help with decisions such as determining the required source sample sizes, the choice of fusion methodology and, in the case of classical fusion, to guide the choice of direction of fusion and the extent of polygamy. For users, it would permit calculation of standard errors, which in turn allows construction of confidence intervals and performing tests of significance.

Marginal outputs

We first analyse the variance of marginal outputs: those that, though computed from the fused dataset, depend on data from only one of the two surveys.

Denote the sample size of the donor survey D as n_D and similarly n_R for the recipient survey R. Note that some of the n_D potential donors may not be used.

Outputs from classical fusion that are based solely on recipients' data

These outputs are not affected by the fusion process and therefore *neff* = n_R, the size of the recipient sample.

Outputs from classical fusion that are based solely on donors' data

Outputs based on donated data in the fused dataset will in general differ from the original estimates, since each recipient's survey weight will differ from that of its donor; and donors may be used with differing frequencies. By restricting attention to simple random samples the donor and recipient weights all become unity, but the influence of the shape of the donor usage frequency distribution cannot be ignored.

By assumption, the donor values start out as a representative, unweighted, sample from the population distribution. But if there are insufficient donors for some of the matching classes, some must be reused. When these are analysed from the fused dataset, it is as though each has acquired a weight, this being the donor's frequency of usage. In terms of effect on variances, the process is essentially the same as weighting a single survey to the population profile (except that in classical fusion only integer-valued weights are used and it is the profile of R not the population that drives the weights). Thus the resulting effect on variances may be calculated using the 'weighting effect' expression:

$$n_D.\sum w_i^2 / \left(\sum w_i\right)^2$$

where n_D is the donor sample size and w_i is the frequency of usage of the ith donor (see Kish 1965, equation 11.7.6′; Conway 1982; Sharot 1986).

The weighting effect is usually termed *weff* and is understood to mean the effect on variances of weighting associated with departures from simple random sampling; to avoid confusion, we will call this polygamy effect *peff*.

Suppose, for example, that there are 1000 potential donors for 1200 recipients and that the usage distribution is as shown in Table 2.

The number of actual donors is 800, but since there are 1200 fused records, this might naively be thought to be the value of *neff*. In fact, the effective sample size is not even 800. The polygamy effect *peff* is:

$$1000.(500 + 2^2.200 + 3^2.100)/1200^2 = 1{\cdot}53$$

Thus the effective number of donors $neff_D$ will be 1000/1·53 = 655, necessarily less than the 800 actual donors and much less than the 1000

Table 2 Frequency of usage of potential donors

Frequency of usage	Number of donors
0	200
1	500
2	200
3	100
4+	–
Total	1000

potential donors or 1200 recipients. It is the latter value that most commercially available analysis software would assume, leading to understatement of standard errors by a factor of √(655/1200) or 0·74 (–26%).

Since no recipient data are involved, this result is not dependent on whether the donated and recipient data are conditionally independent (condition 1 above). However, mismatches do impact (condition 2), since the donated data become tagged with different recipient demographics. If the donated data are dependent on such demographics, regression to the mean results, as in Table 1.

Outputs based on one survey in a transportation fusion

With transportation, all members of both surveys are used at least once and in many cases multiple times; but their original survey weight is split between the resulting records, so there is no polygamy effect. Transportation thereby preserves the marginal outputs from both source surveys. It follows that their variances are unaffected and therefore the effective sample size is the same as for the corresponding source survey.

Outputs that use both donor and recipient data

These are the focus of this paper; for brevity they will be termed 'fusion outputs'. In this section we assume ideal fusion; the next section discusses the effect of failure to achieve this. Consider first fusion outputs that can be written as a function $f(X, Y)$, where X is an output from the donor survey and Y is an output from the recipient survey. Let the underlying distributions of X and Y have means μ_x and μ_y and variances σ_x^2 and σ_y^2. The variance of f will depend on its particular form, but some simple examples provide insight.

The sum of two means

Consider:

$$f(X, Y) = \overline{X} + \overline{Y}$$

Fusion is not required to compute this, but we assume fusion has been performed and the fused database is being used for analysis. If using

classical fusion without polygamy, or transportation, equation (2) gives:

$$\text{var}(\overline{X} + \overline{Y}) = \sigma_x^2/n_{aD} + \sigma_y^2/n_R \quad (3)$$

where n_{aD} is the actual number of donors. Conditional independence means that there is no covariance term.

If using classical fusion with polygamy,

$$\text{var}(\overline{X} + \overline{Y}) = \sigma_x^2/neff_D + \sigma_y^2/n_R \quad (4)$$

where $neff_D$ is the effective number of donors as derived in the previous section. Equation (4) is a general result that includes equation (3) as a special case.

The variance from a single-source survey of size n_R would be:

$$\text{var}(\overline{X} + \overline{Y}) = \sigma_x^2/n_R + \sigma_y^2/n_R \quad (5)$$

Comparing equations (4) and (5), it follows that:

$$neff = n_R.(\sigma_x^2/n_R + \sigma_y^2/n_R)/(\sigma_x^2/neff_D + \sigma_y^2/n_R)$$
$$\therefore neff = (\sigma_x^2 + \sigma_y^2)/(\sigma_x^2/neff_D + \sigma_y^2/n_R) \quad (6)$$

Thus *neff* is a weighted geometric mean of $neff_D$ and n_R. (See Appendix 2 for notes on the geometric mean.) In the simplest case where the variances of the two distributions are equal:

$$neff = 2/(1/neff_D + 1/n_R) \quad (7)$$

which is the simple geometric mean of $neff_D$ and n_R.

The same result holds for the difference of two means and for the sum of the projected totals:

$$f(X, Y) = N.(\overline{X} + \overline{Y})$$

where N is the population size.

In the previous example:

$neff_D = 655$ and $n_R = 1200$, so the fusion *neff* = 847

If instead the 1000 sample were chosen as the recipients and the 1200 as the donors, the extent of polygamy and hence *peff* could probably be

reduced. Apart from any such benefit, we now have $neff_D$ = 1200/1·53 = 786; n_R = 1000; and fusion *neff* = 879 compared to 847. With a reduction in *peff*, the increase will be even greater.

Thus this simple and indeed unlikely example in fact contains far-reaching implications for the design of fusions. In a practical fusion, there is a wide range of outputs from both surveys and it is likely that their variances span a similar range; then equation (7) may be used as a design principle rather than equation (6). The conclusions are then as follows:

- In classical fusion, use if possible the larger survey as the donors, as this will require less polygamy. In practice, another consideration is that classical fusion only preserves the currency of the recipient survey; if it is important to preserve the currency of the larger survey, then it might be chosen as the recipient.
- Transportation will be more efficient still because there is no polygamy effect.

The product of two means

The second example is the product:

$$f(X, Y) = \overline{X}.\overline{Y}$$

For single-source data with independent X and Y, the standard result for the variance is:

$$\text{var}(\overline{X}.\overline{Y}) = \mu_y^2.\sigma_x^2/n + \mu_x^2.\sigma_y^2/n + (\sigma_x^2/n).(\sigma_y^2/n)$$

For fusion this becomes:

$$\text{var}(\overline{X}.\overline{Y}) = \mu_y^2.\sigma_x^2/neff_D + \mu_x^2.\sigma_y^2/n_R + (\sigma_x^2/neff_D).(\sigma_y^2/n_R)$$

In large samples the last term can be ignored, thus:

$$\text{var}(\overline{X}.\overline{Y}) = \mu_y^2.\sigma_x^2/neff_D + \mu_x^2.\sigma_y^2/n_R$$

$$\therefore neff = (\mu_y^2.\sigma_x^2 + \mu_x^2.\sigma_y^2)(\mu_y^2.\sigma_x^2/neff_D + \mu_x^2.\sigma_y^2/n_R)$$

A weighted geometric mean again results, and with the same design implications as before.

The ratio of two means

$$f(X, Y) = \overline{X}/\overline{Y}$$

The standard approximation for the variance of this function is:

$$\text{var}(\overline{X}/\overline{Y}) = \sigma_x^2/(neff_D.\mu_y^2) + \mu_x^2.\sigma_y^2/(n_R.\mu_y^4)$$

where again conditional independence means there is no covariance term. This clearly leads again to a weighted geometric mean for *neff*.

Target group ratings

A TGR is calculated using:

$$TGR = \sum x_i.y_i / \sum y_i$$

where x_i = 0 or 1 indicates whether the programme was viewed by respondent i; and y_i = 0 or 1 indicates whether the respondent is a member of the target group. The product in the numerator is 1 only when the respondent views and is a member of the target group; the sum therefore represents the number of targeted viewers. The denominator is the total number in the target group. This is a more complex function to analyse than those above because it is not expressible as *f(X, Y)*; there is no corresponding output *X*.

Let p_x = prob(x_i =1) be the programme rating and p_y = prob(y_i =1) be the proportion in the target group. The variance of a TGR from a single-source study with sample size n can be shown to be:

$$\text{var}(TGR) = p_x.(1 - p_x)/(p_y.n) \quad (8)$$

The equivalent expression:

$$\text{var}(TGR) = p_x.(1 - p_x)/n + [p_x.(1 - p_x)/p_y^2].p_y.(1 - p_y)/n$$

indicates that $p_x.(1 - p_x)/n$ represents the variance component arising from the viewing data and $p_y.(1 - p_y)/n$ is the variance component from the product data.

This suggests the corresponding formula for a fusion where the viewing data are donated:

$$\text{var}(TGR) = p_x.(1 - p_x)/neff_D + [p_x.(1 - p_x)/p_y^2].p_y.(1 - p_y)/n_R$$

or

$$\text{var}(TGR) = p_x.(1 - p_x)/neff_D + p_x.(1 - p_x).(1 - p_y)/(p_y.n_R) \quad (9)$$

The accuracy of equation (9) was tested by a Monte Carlo simulation, in which var(TGR) was computed for 144 combinations of values of p_x, p_y, $neff_D$ and n_R; a full description of this is given in Appendix 3. Using ordinary least squares regression, the fit to the test data was:

$$\text{var}(TGR) = 0{\cdot}936.p_x.(1 - p_x)/neff_D + 1{\cdot}140.p_x.(1 - p_x).(1 - p_y)/(p_y.n_R) \quad (10)$$

Note that *t*-values etc. are inappropriate because the 144 combinations do not represent a random sample from the population of all possible parameter values. However, the fit is excellent with $R^2 = 0{\cdot}998$; the goodness-of-fit scattergram is shown in Appendix 3. The effective sample size implied by equation (9) is:

$$neff = 1/[p_y/neff_D + (1 - p_y)/n_R]$$

Thus a weighted geometric mean of the source sample sizes is again indicated, with a simple geometric mean if $p_y = 0{\cdot}5$. Note that *neff* does not depend on p_x, a result also supported by the simulation; such a finding is a valuable simplification in practice because it means that there is a single *neff* associated with a target group, valid across all the ratings or reach of the programmes or media used to target it. As an example, if $p_y = 0{\cdot}5$ and $n_D = n_R/2$ with all donors used exactly twice so that $neff_D = n_R/2$:

$$neff = 2.n_R/3$$

The ratio from the simulation was 0·676. Similarly, if $neff_D = n_R/4$ with all donors used four times,

$$neff = 0{\cdot}4.n_R$$

and the simulation gave 0·397.

Non-ideal fusion

In the real world, neither of the conditions for ideal fusion is completely achievable and indeed there is a trade-off between them. Perfect matching may be possible by making the matching criteria very simple, such as matching on sex alone, but this would not provide conditional independence for most outputs. By adding further matching variables it is possible to move closer to achieving complete predictive power but it would never be possible to fully achieve it, and in the process it would prove increasingly hard to make good matches.

It would therefore seem appropriate to investigate the impact on the effective sample size of failure to satisfy these two conditions – but in fact this line of inquiry is a cul-de-sac, because the resulting bias becomes a complicating factor.

Regression to the mean biases each TGR towards the overall programme rating and its variance reduces from that given in equation (8) to that of equation (1), i.e. from

$$\text{var}(TGR) = p_x.(1 - p_x)/(p_y.n) \quad \text{to} \quad \text{var}(TGR) = p_x.(1 - p_x)/n$$

Similarly, the MSI is biased towards 100 and therefore has reduced variance; in the theoretical limit when a fusion fails to detect any selectivity, all MSIs would equal 100 and would therefore have zero variance. Comparisons of design strategies for minimising variance are therefore confounded by the influence of bias on variance.

But conditional independence and perfect matching remain proper goals for fusion; a successful fusion must approach both conditions to a reasonable extent. Provided it does so, the results for effective sample size presented above will be valid approximations. They will also be safe (conservative) estimates, in that the actual standard errors will be somewhat smaller.

To improve estimates of *neff*, it is good practice to apply the geometric mean result separately for each critical matching cell, so as to compute the effective sample size for each, and then to sum these to obtain the overall *neff*. The result will always be smaller than the overall GM, though the difference may be modest if the two samples have similar profiles on the critical variables. As an additional refinement, the weights for the weighted GM may be computed within cell as well.

Outstanding issues

This research is still in progress and currently outstanding issues include:

- the deviation of the coefficients in equation (10) from unity as in (9)
- the variance of TGRs in classical fusion when it is the target group data that are donated rather then the viewing; this seems more complex
- variances for media selectivity indices
- whether or to what extent the weighted geometric mean result has general applicability across outputs and fusion methods.

Summary of results

Conditional independence and perfect matching are aims for any fusion, as they provide protection against regression to the mean; but they also enable derivation and comparison of the variances of outputs.

The effective sample size for many fusion outputs is a weighted geometric mean of the effective number of donors and the effective number of recipients, taking into account in each case the sample designs and the effect of polygamy in classical fusion.

The simple geometric mean may be used as a representative figure in designing fusions. Table 3 shows the geometric mean for a range of sample sizes. Use the effective sample sizes taking into account sample design effects and the polygamy effect, if any.

Polygamous use of donors in classical fusion is a mixed blessing. The improvement it produces in the overall quality of the matches must be weighed against the reduction in effective number of donors. The design guideline is therefore to entertain only modest polygamy (i.e. that which provides an acceptable *peff*) in the search for good matching.

Table 3 Geometric mean for different source sample sizes

	Size of survey A					
Size of survey B	**1,000**	**2,000**	**4,000**	**8,000**	**16,000**	**32,000**
1,000	1,000	1,333	1,600	1,778	1,882	1,939
2,000	1,333	2,000	2,667	3,200	3,556	3,765
4,000	1,600	2,667	4,000	5,333	6,400	7,111
8,000	1,778	3,200	5,333	8,000	10,667	12,800
16,000	1,882	3,556	6,400	10,667	16,000	21,333
32,000	1,939	3,765	7,111	12,800	21,333	32,000

For classical fusion, the effective sample will depend on the direction of fusion. Using the larger sample as donors reduces the amount of polygamy required for good matching, leading to lower variances. It also offers potentially better matching and hence reduced bias.

Transportation does not suffer from the polygamy effect and therefore offers larger effective sample sizes as well as preserving both currencies; but these advantages do not come free: the implicit constraints act against achieving such good matching as in classical fusion.

Appendix 1: The transportation algorithm

It is instructive to draw the parallel with the original transportation problem. Given a set of sources (such as coalmines) with given production levels and a set of destinations (such as power stations) with given input requirements, how much coal should be transported along each possible route so as to minimise the overall transportation cost? This problem can be solved algebraically, using a standard linear programming algorithm (the simplex method) or by more computationally efficient dedicated methods. Fusion can be performed in identical fashion by making the correspondences shown in Figure 1. The respondent weights are denoted by w.

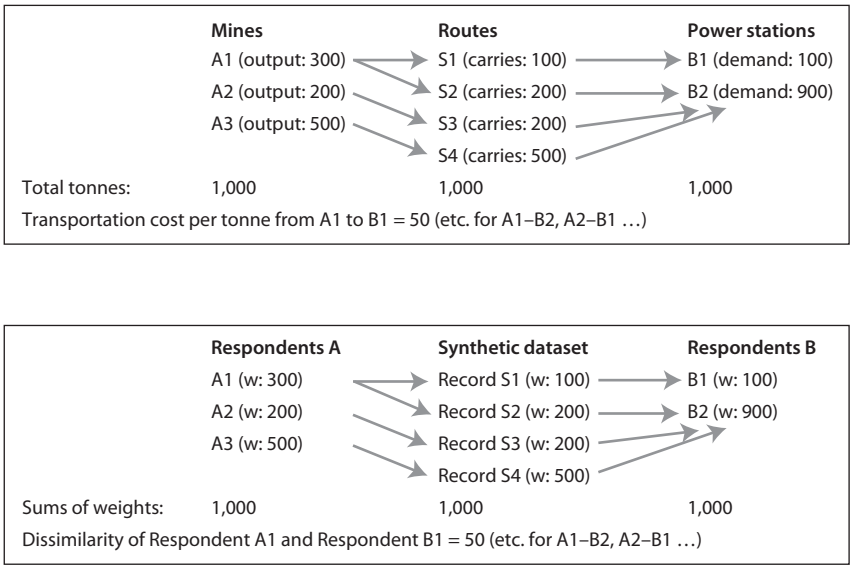

Figure 1 The original and fusion applications of transportation

Unlike classical fusion, transportation is symmetric – that is, either survey may be regarded as *A* and the other as *B*. The synthetic sample is given both sets of unique outputs and may take common outputs such as demographics from either survey.

So using transportation, every member of *A* (A_i) is used for at least one synthetic record; their weights sum to the A_i's weight so A_i's weight is preserved when totalling. Similarly, every member B_j of *B* is used for at least one synthetic record; their weights sum to B_j's weight.

Transportation differs from classical fusion in one other key respect: implementations of the latter typically seek out good matches in sequential fashion, best to worst, while the former finds the overall optimum arrangement that maximises the average similarity over all the matches, subject to the constraints. It is possible to show by example that these constraints may result in inferior matching (lesser average similarity) than classical fusion; the numerical effect of this may or may not be significant.

Appendix 2: The geometric mean

The geometric mean is encountered less commonly than the familiar arithmetic mean. The simple or unweighted geometric mean of *n* positive numbers x_i is:

$$GM = n/(1/x_1 + \ldots + 1/x_n)$$

This is the reciprocal of: the arithmetic mean of: the reciprocals of the observations. It has certain natural applications – for example, if an outward journey is made at an average speed of 50 kph and the return at 40 kph, the overall average is the geometric mean 44·4 kph.

The weighted geometric mean is defined as:

$$WGM = \left(\sum w_i\right)/\left(\sum w_i / x_i\right)$$

Appendix 3: Monte Carlo simulation

Suppose that viewing data *x* are derived from a donor sample of size n_D; and that product usage *y* is from a recipient sample of size n_R. Let the viewing probability be p_x; for simplicity this was assumed to be the same for all donors; similarly let the product usage probability be p_y for all recipients. Product users form the target group.

Each donor views or not according to whether a random number in the range (0,1) is less than p_x; a similar process identifies product users.

The TGR is:

$$\sum x_i.y_i / \sum y_i$$

The above calculations were replicated 10,000 times and the variance of the 10,000 TGR estimates was calculated (as was the mean, as a check on the process). This number of replications reduces random error to about the 4th or 5th significant digit in the variance.

The above was repeated for different values of p_x and p_y, taking all 16 combinations of (0·125, 0·25, 0·5, 0·75) × (0·125, 0·25, 0·5, 0·75).

All the above was repeated for the sample sizes (n_D, n_R) = (64, 64), (32, 64), (16, 64).

In the first case, $n_D = n_R$ and donors were paired with recipients in sequence, since perfect matching is assumed. In the second situation, each donor was paired twice. In the third situation, each donor is used four times. The choice of n_R is arbitrary and no variation is necessary since it is clear that (say) doubling both n_R and n_D will halve all variances. Nevertheless, for a richer dataset, the following sample sizes were also used:

(128, 128), (64, 128), (32, 128); (256, 256), (128, 256), (64, 256)

In total this provides 16 × 9 = 144 parameter combinations.

The 10,000 replications for each combination take in total about one minute on a Pentium 4 PC.

Figure 2 shows the scatter diagram of the 144 variances, according to the simulation on the x-axis and from formula (10) on the y-axis.

In order to show the fit more clearly for the smaller variances, the graph is repeated as Figure 3, using a log scale on both axes.

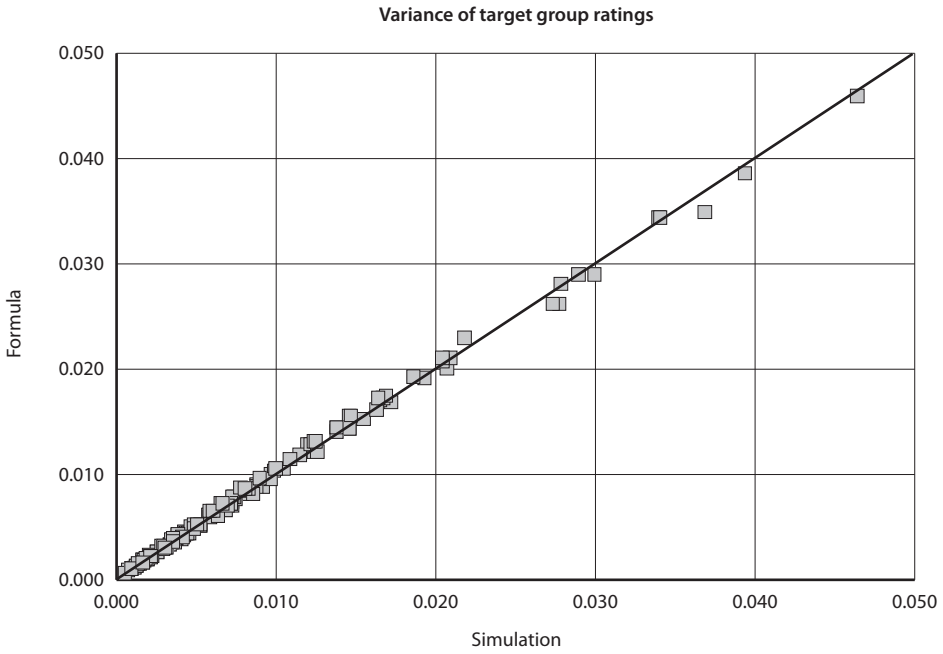

Figure 2 Scatter diagram of the 144 variances

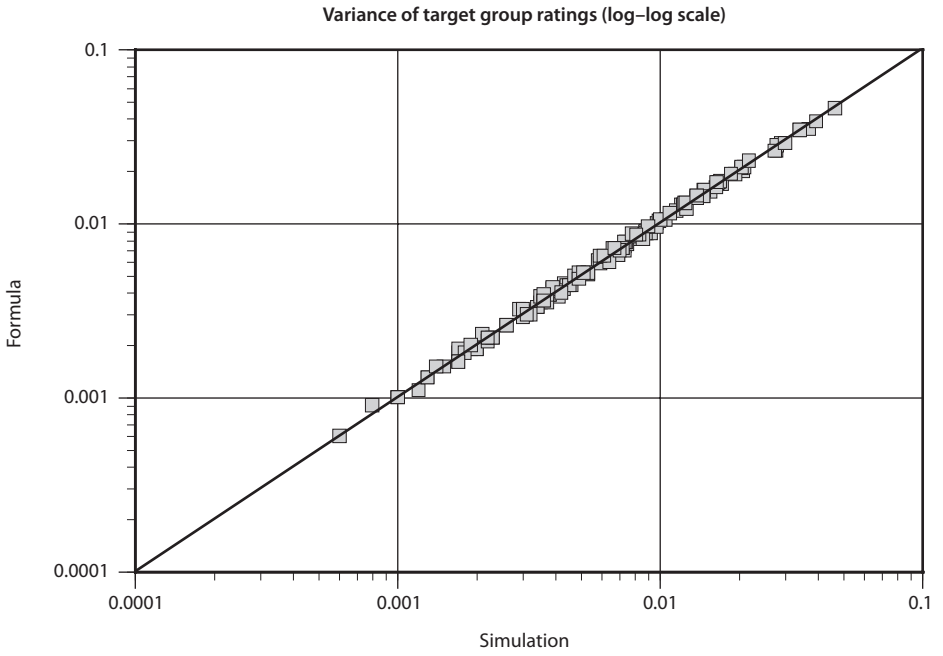

Figure 3 Scatter diagram of the 144 variances, using a log scale on both axes

Acknowledgements

The author is grateful to Pete Doe, Dr Jacky Galpin, Ian Garland and Andrew Whitney for their comments on drafts of this paper.

References

Barry, J.T. (1988) An investigation of statistical matching. *Journal of Applied Statistics*, **15**, 3, pp. 273–283.

Conway, S. (1982) The weighting game. *Market Research Society Conference Papers*, pp. 193–207.

Jephcott, J. & Bock, T. (1998) The application and validation of data fusion. *Journal of the Market Research Society*, **40**, 3, pp. 185–205.

Kish, L. (1965) *Survey Sampling*. New York, NY: Wiley.

Mahalanobis, P.C. (1936) On the generalised distance in statistics. *Proceedings of the National Institute of Science of India,* **12**, pp. 49–55.

Sharot, T. (1986) Weighting survey results. *Journal of the Market Research Society*, **28**, 3, pp. 269–284.

Soong, R. (2006) *The Data Fusion Bibliography*, www.zonalatina.com/datafusion.doc, current update 6 October 2006.

About the author

Trevor Sharot has been a professional survey statistician for a very long time. Based in Singapore, he is Director of Measurement Science for AGB Nielsen Media Research, a company specialising in television audience measurement; and is responsible for survey methodology, sample design and data quality for seven countries across Asia. His other pursuits are the arts, Hash House Harriers, climbing volcanoes and motorcycle racing, though not simultaneously.

Address correspondence to: Trevor Sharot, AGB Nielsen Media Research, 55 Newton Road, #13-02/03 Revenue House, Singapore 307987.

Email: trevsh1@hotmail.com