FOUNDATIONS OF SPLIT-SAMPLE FOLDOVER TESTS

Roland Soong and Michelle de Montigny, Kantar Media Research

1. ABSTRACT

The split-sample foldover test is an important tool for assessing the accuracy of techniques such as data fusion/integration, ascription and predictive modeling, as it is an empirical procedure that does not depend on unverifiable assumptions. When a split-sample foldover test is conducted, there are many criteria that can be evaluated, some of which are relevant and others not. In this paper, we present a typology of evaluative criteria and we will use a number of applications to illustrate how to choose the relevant criteria.

2. BACKGROUND

Data fusion is the practice by which two or more respondent-level databases are brought together to form a single respondent-level database that contains the previously separate information. There are many ways in which this can be achieved. One common approach is through the method of statistical matching. The Worldwide Readership Research Symposium has been the repository of many works on data fusion (e.g. Antoine (1985), Bennike (1985), Scheler and Wiegand (1985; 1987), Frankel and Baxter (1988), Bedwell (1991), Czaia (1993), Raimondi and Santini (1997; 2001), Soong and de Montigny (2001)).

There is no lack of ideas for fusing data, but eventually these methods will have to be evaluated for accuracy. The principal methodology for validation is the split-sample foldover test, wherein a single source database is split into two different portions which are fused together and then the fused data are compared with the original data.

The split-sample foldover test can also be used in other applications, such as data integration (Walsh (2001)), ascription (Frankel (1981), Frankel and Baxter (1988), Speetzen (1988), Baim and Frankel (1997), Mallet (1997)), predictive modeling (Weiss and Indurkhya (1998)) and split-sample surveys (Page (2001), Bals (2002), Rässler (2002b)). A good understanding of the foundations of this methodology is important, otherwise we will be talking at cross purposes about the validity and usefulness of these types of techniques.

3. DATA FUSION

For discussion purposes, it would be helpful to describe an example of data fusion in order to give some concreteness to the abstract ideas. This example is based upon statistical matching, and has been used extensively around the world (Baker, Harris and O'Brien (1989) and Soong and de Montigny (2001)).

The most prevalent form of syndicated data fusion is the (TAM+TGI)-like fusion. On one side, we have a television audience (TAM) people meter panel. On the other side, we have a Target Group Index (TGI) consumer survey of media and product usage behavior. The respondents from the TAM and TGI databases are matched to each other based upon the similarity on common variables (such as gender, age, geography, television viewing, etc). The fusion database is a static respondent-level database, where the 'respondents' now carry information from both databases.

If our goal is to study the accuracy of the data fusion, then the (TAM+TGI) fused database itself will be uninformative. Unless we know what the true values are, we cannot know if the fusion was done accurately. But if we know the true values, we would have no need to conduct data fusion.

Rather, the standard approach in assessing the accuracy of data fusion is through a split-sample foldover test. For this test, we require a 'single source' database that contains both TAM-like and TGI-like information. Usually, the TGI-like database contains some television variables that can be used as surrogate TAM variables. Sometimes, this database can even be a third independent database.

This 'single source' database is randomly split into two halves, which are then fused together using the designated method. At the end, there is a respondent-level database, where each 'respondent' carries both original and fused data which can be compared.

4. REPRESENTATIVENESS

If we plan to use a split-sample foldover test on a single source database to evaluate a data fusion, we must ensure that this is an accurate representation of the actual situation. Although this seems obvious, it is sometimes easy to forget. For illustrative purposes, we consider the 2003 NTI/MARS fusion. In the actual fusion for the first quarter of 2003, there were 11,723 adults in the NTI database and 21,106 in the MARS database.

The split-sample foldover test would have to be done on the MARS database, which contains the matching variables, target group information as well as a set of surrogate television variables. A straightforward split-half sample test would involve dividing the MARS database into two halves of 10,553 and 10,553 cases. There may be reason to wonder if a 10,553/10,553 test is a realistic representation of the true 11,723/21,106 configuration.

To get to sample sizes that are closer to those in the syndicated fusion product, it is necessary to use a MARS doublebase combining the 2002 and 2003 surveys. The total sample size is 43,203 persons. If we subsample 25% of this doublebase for an "NTI" sample and 50% for a "MARS" sample, we get a 10,801/21,602 distribution, which is a lot closer to the true situation.

The simplest set of characteristics for a statistical matching method is the success rate in the matching variables. In Table 1, we show the successful match rates in the NTI/MARS syndicated study, the 2003 MARS split-half samples and the 2002-2003 MARS 25%/50% samples. In addition, to see the effects of various combinations of sample sizes, we also ran one-half, one-quarter and one-eighth of the 25%/50% sample sizes.

From Table 1, the smaller the sample size, the harder it is to find the perfect match on everything. At some point, the characteristics of the split-samples can no longer be said to correspond to the original databases.

Table 1. Successful Match Rates by Fusion Methods

	NTI/ MARS Syndicated Study	MARS 2003 Split Half	MARS Doublebase 25%/ 50%	MARS Doublebase 12.5%/ 25.0%	MARS Doublebase 6.25%/ 12.50%	MARS Doublebase 3.175%/ 6.2500%
"NTI" Sample Size	11,723	10,553	10,801	5,400	2,700	1,350
MARS Sample Size	21,106	10,553	21,602	10,801	5,400	2,700
Sex*	100 %	100 %	100 %	100 %	100 %	100 %
Age*	100 %	100 %	100 %	100 %	100 %	100 %
TV Viewing*	100 %	100 %	100 %	100 %	100 %	100 %
Presence of 2<	95 %	91 %	92 %	91 %	89 %	91 %
Household size	92 %	90 %	91 %	91 %	88 %	85 %
Presence of 6-11	90 %	88 %	90 %	87 %	85 %	82 %
Presence of 2-5	89 %	89 %	90 %	88 %	89 %	88 %
Presence of 12-17	89 %	87 %	86 %	84 %	81 %	79 %
Working Women	88 %	86 %	87 %	85 %	84 %	85 %
Cable TV	85 %	85 %	87 %	84 %	79 %	72 %
Age of HOH	76 %	75 %	75 %	73 %	74 %	73 %
County Size	76 %	67 %	68 %	63 %	59 %	53 %
Race	75 %	73 %	71 %	70 %	71 %	65 %
Household Income	64 %	61 %	61 %	57 %	52 %	45 %
Educ of HOH	59 %	57 %	58 %	55 %	49 %	48 %
Occup of HOH	59 %	55 %	58 %	54 %	50 %	44 %

* Fusion strata achieve 100% success by definition.

Therefore, we remind people who wish to use the split-sample foldover test to check that the split samples are reasonable representations of the original databases.

5. REPORTING STANDARDS

In 1974, the physicist Richard Feynman (1997) gave a memorable commencement address at Caltech:

> … we all hope you have learned in studying science in school --- we never say explicitly what this is, but just hope that you catch on by all the examples of scientific investigation. It is interesting, therefore, to bring it out now and speak of it explicitly. It is a kind of scientific integrity, a principal of scientific thought that corresponds to a kind of utter honesty --- a kind of leaning over backwards. For example, if you're doing an experiment, you should report everything that you think might make it invalid --- not only what you think is right about it: other causes that could possibly explain your results; and things you thought of that you've eliminated by some other experiment, and how they worked --- to make sure the other fellow can tell they have been eliminated.
>
> Details that could throw doubt on your interpretation must be given, if you know them. You must do the best you can --- if you know anything at all wrong, or possibly wrong --- to explain it. If you make a theory, for example, and advertise it, or put it out, then you must also put down all the facts that disagree with it, together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that give you the idea for the theory; but that the finished theory makes something else come out right, in addition.
>
> In summary, the idea is to give all of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another.

So what is the full information set that can be made available after performing a split-sample foldover test? A systematic approach can be found in the monograph of Suzanne Rässler (2002), section 2.5 which we will present below. In doing so, we have reversed her numbering scheme because our experience was that the Nintendo generation prefers to think of Level 1 as the basic entry level and then moving up the level numbers.

Level 1: Accuracy of marginal distributions

This is the basic requirement that the marginal distributions for the variables within each of the separate samples should be accurate.

Level 2: Accuracy of correlations between variables

When we bring two variables together, one from each database, the simplest measure of their association is their correlation. This is the requirement that such correlations should be accurate.

Level 3: Accuracy of joint distributions among variables

When the joint distribution of several variables is accurate, a complex estimates derived from this set of variables will be accurate as well. The preservation of the joint distribution between a pair of variables implies that the correlation between them is preserved, but not necessarily vice versa. Furthermore, a joint distribution involves two or more variables. Therefore, Level 3 is more demanding than Level 2.

Level 4: Accuracy of individual variables

For each individual respondent, we require that the value of each individual variable be accurate. Whereas Levels 1 through 3 are aggregate-level measures, Level 4 is an individual-level measure that is examined on a case-by-case, variable-by-variable basis. This makes it the most challenging level to achieve.

These fairly abstract descriptions are now made concrete in the context of the data fusion of two databases through statistical matching to create a single respondent-level 'single source' database. For simplicity, we will assume that one database contains television program ratings and the other database contains magazine ratings.

Level 1: Accuracy of marginal distributions

This is the requirement to preserve the media currencies in the two separate databases. The television ratings are the currency values for television planning, while the magazine ratings are the currency values for magazine planning. It would be undesirable for these ratings to become distorted somehow in the fused database to affect analyses and decisions.

Level 2: Accuracy of correlations between variables

The correlation coefficient of two variables is related by formula to the duplication between the variables. In substantive terms, this is the requirement that the duplication between any television rating and any magazine rating ought to be accurate.

Level 3: Accuracy of joint distributions among variables

The joint distribution of a group of variables (such as a set of magazines and television programs) is too unwieldy. In practice, we are interested in the accuracy of complex estimates derived from the joint distributions, such as the reach/frequency distributions of a mixed media television-print schedule of insertions in a collection of television and magazine vehicles.

Some reach/frequency estimation services require only ratings and pairwise duplications to feed into mathematical models, so that Level 1/2 accuracy is sufficient.

Level 4: Accuracy of individual variables

This is the requirement that each outcome variable for each respondent should be accurate. Thus, a television viewer should be classified as such and vice versa; and a magazine reader should be classified as such and vice versa.

In practice, total accuracy is unlikely to be attained. So accuracy is not a choice of "Yes, everything is exactly the same and therefore it is perfectly accurate" versus "No, something is not perfect somewhere and therefore this method is not accurate." Rather, the accuracy may be captured with a measure of error that can be put in a statement such as "the average distortion in the magazine rating is of the order of magnitude of about 2 parts in 1,000, and so the magazine ratings can be considered accurate for practical purposes" or "the average magazine receives an audience level that is 15% higher and this is a significant distortion of the currency values."

Levels 1 through 3 are aggregate-level measures, so that the accuracy will most typically be reported as the difference between a true average versus the estimated average. Level 4 is an individual-level measure. When the variable is discrete (e.g. reader vs. non-reader), the result is usually presented in the form of a 2x2 confusion matrix with the four cells identified as true positive, true negative, false positive and false negative. When the variable is continuous (e.g. number of minutes spent watching a television program), the result is usually presented in the form of an error measure (e.g. mean absolute difference between true and estimated number of minutes spent averaged across all respondents).

If we follow Feynman's suggestion, we would be reporting on all these possible statistics and perhaps some more. But is it necessary? We don't think so. The situation is analogous to media planning. Users typically have massive amounts of information on hand, of which some but not all may have relevance or connection to a specific problem. But, at some point, the user has to reduce the information set and find a solution. For example, the user defines a target group, selects a set of media vehicles and then obtains an optimized plan (e.g. maximum reach within fixed budget, or minimum cost within fixed reach) with specialized software. It is nice to know all that other information, but they have no direct relevance at that moment.

Here, it is no different. We can report on everything possible, but at some point someone has to take the responsibility of identifying those key measures that bear directly on the validity and accuracy of the data fusion for the intended applications. This is our position.

Our position is grounded heavily upon our experience in presenting our results. For example, in Soong and de Montigny (2003a), we presented the comparative results from several data fusion techniques and we reported on various evaluative criteria *à la* Feynman. Although we made sure to point out those criteria that we regard as being directly relevant, we found the discussion in the Q&A session included much that we did not regard as relevant. By contrast, in Soong and de Montigny (2003b), we have streamlined the presentation to focus just on those criteria that we regard to be immediately relevant, with a clear explanation as to our reasons. We can produce the other information, but there ought to be some rationale beyond showing more information for its own sake.

In the following, we will offer some illustrative examples. In each case, we describe a very common application of data integration/fusion. Based upon the way in which that application will be used, the relevant set of evaluative criteria is identified and the results are then presented.

6. APPLICATION A. TARGET GROUP TELEVISION RATINGS

This is best known application of data fusion/integration. On one side, there is a Television Audience Measurement (TAM) system based upon a people meter panel. On the other side, there is a Target Group Index (TGI)-like survey in which product usage data are collected. The data fusion/integration brings together the television ratings and product usage to produce target group television ratings.

There are a number of data fusion/integration techniques that can be used to produce target group television ratings. A partial list includes unconstrained statistical matching (Baker, Harris and O'Brien (1989); Carpenter and Wilcox (1995)), constrained statistical matching (Soong and de Montigny (2001)), predictive isotonic fusion (Soong and de Montigny (2003a; 2003b)), just-in-time modeling (Raimondi and Santini (1997)) and multi-basing (Walsh (2001)).

The users will make use of television program rankers, reach/frequency estimators and schedule optimizers on these data. The inputs into these systems consist of the following statistics, which we have marked with the corresponding level in the Rässler scheme:

- Target group incidence (Level 1)
- Total television ratings (Level 1)
- Target group television ratings (Level 2)
- Target group television schedule reach & frequency characteristics (Level 3)

For illustration, we will consider the example based upon the 2002 MARS database as reported in Soong and de Montigny (2003a). Our interest is in comparing the accuracy of four different fusion methods: a constrained statistical matching using only demographics, and three different predictive isotonic fusions using various combinations of predictor variables.

The 2002 MARS database was randomly split into two halves, and then the split-samples were fused together. Since the MARS database contains television viewing as well as target group information, we are able to calculate true target group ratings as well as fused target group ratings.

Since all the data fusion methods here are of the constrained variety that retains full sample size and case weights, the target group incidences and the total television ratings are automatically preserved. The MARS television data are generic variables (e.g. program type viewing, weekly cumes for cable networks and total daypart viewing) which do not correspond to actual television schedules. Therefore, realistically speaking, we are left with the evaluation of the target group ratings themselves. In Table 2, we show an example of target group ratings.

Table 2. Comparison of original and fused target group ratings for program types
Fusion method: Constrained statistical matching with demographics
Target group: Adults with acid reflux

TV Program Type	Original TGR	Fused TGR	Difference
Audience participation	40.6	39.1	1.5
Awards/pageants	19.8	17.4	2.4
Day animation	9.3	9.7	-0.4
Daytime drama	22.5	21.6	0.9
Religious programs	11.9	10.1	1.8
Drama	40.9	39.9	1.0
Evening animation	17.7	18.4	-0.7
Movies	56.5	53.9	2.6
Music	20.0	19.6	0.4
News	62.2	65.9	-3.7
News magazines	34.3	34.6	-0.3
Reality shows	20.7	20.2	0.5
Science fiction	18.6	19.2	-0.6
Situation comedy	47.8	48.5	-0.7
Sports anthology	7.0	7.0	0.0
Sports events	40.0	40.8	-0.8
Talk	34.0	30.0	4.0

In Table 2, the mean difference is 0.46 and the mean absolute difference is 1.31. The entire exercise covers 40 target group and 63 television variables, and the overall results of those 40 x 63 = 2,520 combinations are shown in Table 3.

Table 3. Target group rating differences for four data fusion methods

Method	Mean Difference	Mean Absolute Difference
Constrained statistical matching	1.78	2.50
Predictive isotonic fusion: demographics	0.97	2.71
Predictive isotonic fusion: demographics+tv	0.48	2.02
Predictive isotonic fusion: demographics+tv+print	-0.04	1.90

Table 3 allowed us to see how the choice of method and matching/predictor variables can lead to improvements in accuracy. These arithmetic means are not the only possible measures from the results, as we can also look at other measures such as medians, quantiles, standard deviations, interquartile range, range, etc.

So far, we have simply reported some summary statistics. The same set of original and fused target group ratings can be evaluated by more complicated approaches. For example, we can check if the fusion leads us to select the same set of television vehicles for a target group; or to see how different original and fused optimized schedules look. In all cases, we would be imitating the actions and decision-making processes in real-life media planning.

We note that we have identified the relevant measures in this application as being Level 1 through 3 aggregate-level measures. We do not believe that Level 4 individual-level measures have direct relevance here and we will explain why.

On one hand, for example, suppose that we are told that 85% of the target group information is correctly identified at the individual level. This tells us nothing about the accuracy of the target group ratings, which was our primary interest. We would still be obliged to go through the same steps to obtain the direct answer.

On the other hand, for example, if we believe that the target group ratings are sufficiently accurate in the sense that we will up making the right media decisions, then Level 4 measures are irrelevant. Conversely, if we believe that the target group ratings are inaccurate in that we are making inappropriate media decisions, then Level 4 measures are irrelevant as well.

7. APPLICATION B. MULTIMEDIA TELEVISION/PRINT ANALYSIS

This is a common application that links the data from two large advertising sectors together for planning purposes. On one side, there is a Television Audience Measurement (TAM) system based upon a people meter panel. On the other side, there is a Target Group Index (TGI)-like survey in which product usage and print readership data are collected. The data fusion/integration brings everything together to analyze multimedia television-print schedules for target groups.

The same type of data fusion/integration techniques in Application A can be used here. A partial list of examples of this application include Wilcox and Johnson (1997), de Montigny and Lima (2001), Collins, Mallett and Traub (2002) and Soong and de Montigny (2002).

The users will make use of reach/frequency estimation services and optimizers. The inputs into these systems consist of the following statistics, where we have marked the corresponding level in the Rässler scheme:

- Target group incidence (Level 1)
- Total television ratings (Level 1)
- Total magazine ratings (Level 1)
- Target group television ratings (Level 2)
- Target group magazine ratings (Level 1)
- Target group pairwise duplications (Level 2)
- Target group television-print schedule reach & frequency characteristics (Level 3)

For illustration, we will consider the example based upon the 2002 MARS database as reported in Soong and de Montigny (2003b). Our interest is in comparing the accuracy of four different fusion methods: a constrained statistical matching using only demographics, and three different predictive isotonic fusions using various combinations of predictor variables.

The MARS database was randomly split into two halves, and then the split-samples were fused together. Since the MARS database contains television viewing, magazine reading as well as target group information, we are able to calculate true target group estimates as well as fused target group estimates.

Since all the data fusion methods here are of the constrained variety that retains full sample size and case weights, the target group incidences, the total television ratings, the total magazine ratings and the target group magazine ratings are automatically preserved. The accuracy of the target group television ratings is in fact addressed in Application A. The MARS television data are generic variables (e.g. program type viewing, weekly cumes and daypart total viewing) which do not correspond to actual television schedules. So realistically, the only additional task here is to check the accuracy of the target group pairwise duplications. In Table 4, we show an example of target group intermedia duplications.

Table 4. Comparison of original and fused target group duplications
between television news and selected magazine titles
Fusion method: Constrained statistical matching with demographics
Target group: Adults with acid reflux

Magazine Title	Original Duplication	Fused Duplication	Difference
Martha Stewart Living	1.27	1.39	-0.15
Maxim	3.07	3.08	-0.01
Men's Fitness	3.83	4.21	-0.38
Men's Health	3.64	3.74	-0.10
Men's Journal	1.48	1.65	-0.17
Midwest Living	1.11	1.05	0.06
Money	2.03	1.78	0.25
National Geographic	10.00	10.01	-0.01

In this example, the mean difference is -0.06 and the mean absolute difference is 0.14.

Soong and de Montigny (2003b) considered 40 target groups, 63 television measures and 96 magazine ratings. There are 40 x 63 x 96 = 241,920 combinations. Not all are suitable, since there is no reasonable chance that anyone would include opposite media vehicles like MTV and AARP Magazine within the same schedule. If we restrict for each target group to just those media vehicles that have target group indices greater than 105, then we are left with 76,093 television-print pairs. The summary numbers for these pairs are shown in the Table 5.

Table 5. Target group television-print duplication differences for four data fusion methods

Method	Mean Difference	Mean Absolute Difference
Constrained statistical matching	0.40	0.57
Predictive isotonic fusion: demographics	0.25	0.65
Predictive isotonic fusion: demographics+tv	0.17	0.56
Predictive isotonic fusion: demographics+tv+print	0.11	0.53

Table 3 allowed us to see how the choice of method and matching/predictor variables can lead to improvements in accuracy.

Once again, as in Application A, we have identified the relevant measures in this application as being Level 1 through 3 aggregate-level measures in the Rässler scheme.

So far, we have presented only data fusion methods here. Data fusion is characterized by the physical presence of a 'respondent'-level database, as opposed to data integration techniques that combine information without producing such a database (e.g. Cannon (1988), Cannon and Seamons (1995), Danaher & Rust (1992), Walsh (2001)). The accuracy of all these methods is measured in terms of the target group ratings and pairwise duplications, which are fed into the end results. Therefore, the evaluative criteria that we have just described for Applications A and B are also appropriate for data integration techniques. It then comes as no surprise that Level 4 individual-level measures do not matter for data fusion, since the concept does not even exist naturally for data integration methods.

8. APPLICATION C. PREDICTIVE MODELING

Predictive modeling is used extensively in database marketing, data mining, direct marketing, credit card solicitation, credit scoring, insurance prospecting, loan approval, etc (see Weiss and Indurkhya (1998)). For an example in the print industry, we consider the case of a magazine publishing group that has compiled a large database of its subscribers, for whom a small number of demographics and their subscription history are known.

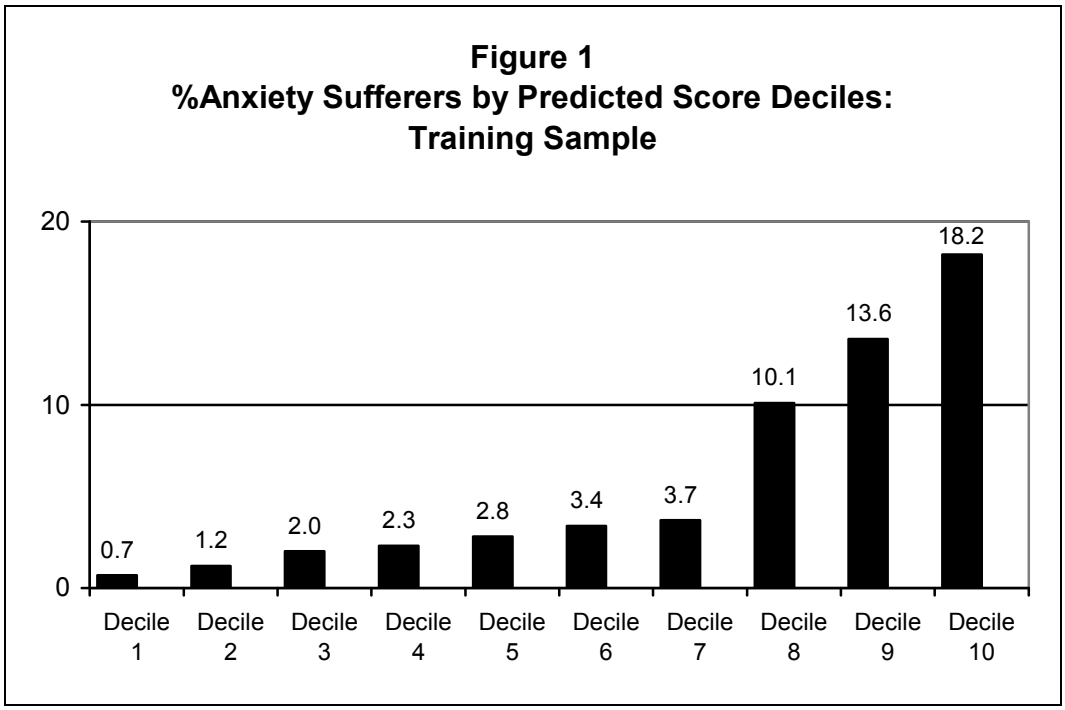
For a database marketing project, a random sample of about 1,000 of the subscribers was surveyed with respect to their likelihood of purchasing certain products/services (such as luxury import cars, travel packages, laptop computers). With this sample, a predictive model can be built to score the respondents in terms of their propensity to purchase. This predictive model can be applied to all the people in the large subscriber database. Those who score above a certain cutoff score will be included in a direct mail campaign. Such a campaign can be quite expensive because millions of names may be selected. Therefore, the user needs to have a robust and accurate predictive model of the performance.

For illustrative purposes, we show an example from the MARS database. The target product is a pharmaceutical product for relieving anxiety/depression symptoms. We split the 2002 MARS database into two halves. Following the technical jargon of predictive modeling, we designate them as the training sample and the validation sample respectively.

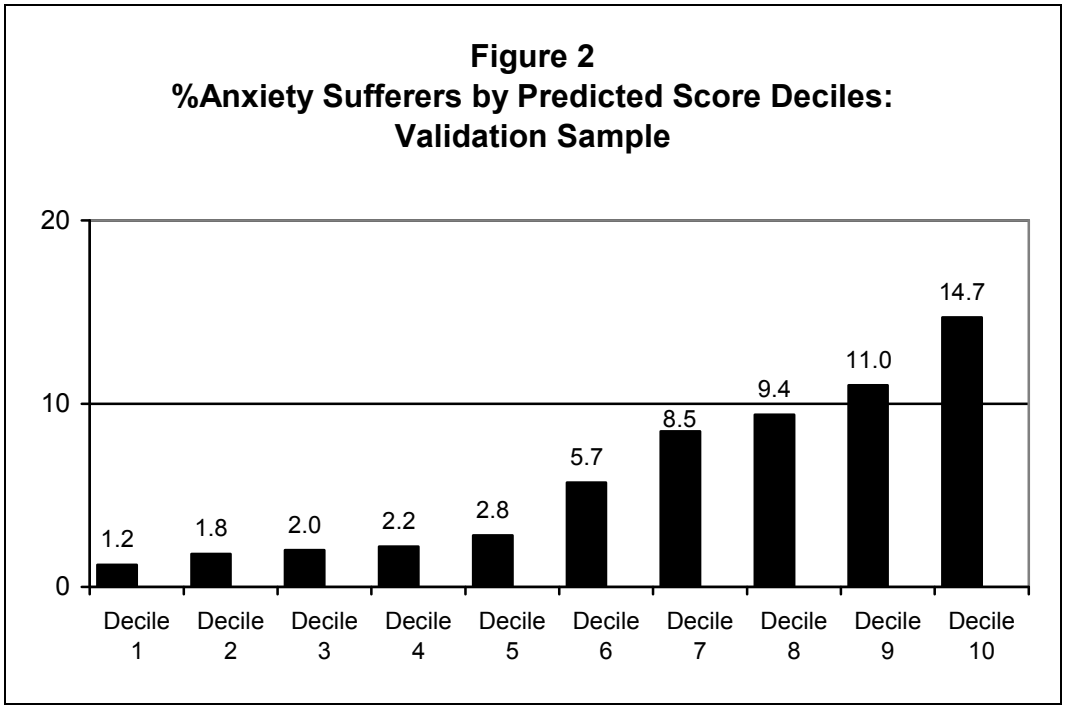
Within the training sample, we built a regression model with the presence of anxiety/depression as the dependent (outcome) variable and a set of demographic variables (e.g. age, gender, education, household income, occupation, household composition, geography, etc) as predictors. There are many other techniques (such as discriminant analysis, neural networks, CHAID, etc) that could have been used. This allows us to assign a score to every person in the training sample.

The goodness-of-fit for the model could be summarized in terms of measures such as the correlation coefficient, R^2, likelihood ratio and so on, but they do not provide directly relevant information about the business aspects. Predictive modelers have a more appealing visual approach. Here, the training sample is sorted into deciles (10%-tiles) based upon the predicted scores, and then the target group incidences are calculated by decile. This is shown in Graph 1.

Overall, the incidence of people who suffer from anxiety/depression is 5.8%. If the predictive model were totally ineffective, the incidences would have around 5.8% everywhere. If the predictive model was effective, then the top deciles would have considerably higher incidences, with a declining trend down the deciles. This is indeed the observed situation in Chart 1.



The use of a many-parameter predictive model will result in the overfitting of the data. This means that the performance measures from training samples may be inflated. The predictive modeler will apply the predictive model onto a validation sample, which has been held out from the analysis so far solely for this purpose. The validation sample is then sorted into deciles of these predicted scores, and then the target group incidences are calculated by decile. The results are shown in Chart 2.



By comparing Figures 1 and 2, we can see that there is a general pullback (known as regression-to-the-mean) in the top deciles. These are now realistic reflections of the performance of the predictive model.

At this point, the predictive modeler would go through a series of "What If?" analyses based upon profit/cost considerations. For persons who are not in the target group, the estimated cost will be a loss of $2 per person for postage/materials/handling because he/she has no need for the product. For persons who suffer from anxiety/depression, about 10% will respond and each responder will yield an average lifetime value of $200 per person. In other words, each person in the target group generates an average profit of $20 per person.

Suppose we mailed to the entire database, which contains 10,000,000 names. Since the overall incidence is 5.8%, there are 580,000 persons in the target group who will generate 580,000 x $20 = $11,600,000 in profits. But there are also (10,000,000 – 580,000) = 9,420,000 persons not in the target group who will cause 9,420,000 x $2 = $18,840,000 in losses. This project would result in a net loss of $11,600,000 - $18,840,000 = -$7,240,000. This is a money-losing proposition.

But suppose instead we mailed only to the 3 million names in the top three deciles. According to Figure 2, the average target group incidence is (14.7+11.0+9.4)/3 = 11.7%, which is about twice the overall average. There are 3,000,000 x 11.7% = 351,000 persons in the target group who will generate 351,000 x $20 = $7,020,000 in profits. But there are also (3,000,000 – 351,000) = 2,649,000 persons not in the target group who will cause 2,649,000 x $2 = $5,298,000 in losses. The project would result in a net profit of $7,020,000 - $5,298,000 = $1,722,000. Now we have a money-making proposition by being more selective.

The profit/loss analyses are highly sensitive to the robustness and accuracy of the predictive model, where erroneous assumptions can result in huge losses. That is why predictive modelers go through a lot of trouble to get it right. Although they use their own set of technical jargon, it is clear from our description that this is the split-sample foldover test.

With respect to the Rässler scheme, this is a Level 4 analysis. We are interested in how individual persons were classified correctly or incorrectly as members of the target group, which lead to profit/loss implications.

9. APPLICATION D. SPLIT-SAMPLE MAGAZINE SURVEY

The media-rich environment is such that more and more magazines are being published. Whereas in the 1950's, it was enough to talk about the measurement of just 10 major magazines, the number of published titles in any country will be in the hundreds, even thousands. More and more magazines need to be measured in order to compete for advertising dollars. However, it is a humanly impossible task for people to respond to questions on so many magazines.

One alternative is to ask each respondent just a subset of the possible magazines, and then use some form of fusion/ascription to impute the readership of the magazines that were not asked. Examples of this approach can be found in Page (2001), Bals (2002) and Rässler (2002b).

For illustrative purposes, we will use an artificial example taken from the MARS 2002 database. We divide the sample into two equal halves. We assume that one half-sample has demographics, target group information and readership information for 24 magazine titles. We assume that the other half-sample has demographics, target group information, and readership information for 48 magazine titles, of which 24 are those that appear in the first half-sample. The goal of the data fusion is to impute the readership of the 24 missing magazines in the first half-sample.

The key application here is the target group magazine schedule analysis. For a specified target group and a schedule consisting of various insertions in a set of magazines, we are interested in the gross rating points, reach and frequency of the schedule. The required input data consists of:

- Target group magazine ratings
- Target group pairwise inter-magazine duplications

These input data are fed into mathematical models to generate gross rating points, reach and frequency.

For illustration, we will be comparing two different methods of data fusion. In the first instance, this is a constrained statistical matching based upon demographics only. In the second instance, it is a constrained statistical matching based upon demographics as well as readership to the 24 magazines that are measured in both half-samples. We wanted to know if the readership information is useful in improving the accuracy of the fusion.

In the table below, we show the results of the target group magazine ratings for 24 magazines by 40 target groups. We compare the fused target group rating (based upon the actual target group information and the fused readership information) against the original target group rating (based upon the actual target group information and the actual readership information). We see from this table that the readership variables had the expected effect of reducing the differences.

Table 6. Target group magazine rating summary for two data fusion methods

Method	Mean Difference	Mean Absolute Difference
Constrained Statistical Matching: Demographics only	-0.85	1.41
Constrained Statistical Matching: Demographics + Magazine Readership	-0.30	1.11

In the table below, we show the results of the target group inter-media duplications for the 24 measured magazines against the 24 other magazines that were fused. The summary numbers were obtained for 40 x 24 x 24 = 23,040 combinations. Again, the readership variables yielded closer numbers.

Table 7. Target group pairwise inter-magazine duplication summary for two data fusion methods

Method	Mean Difference	Mean Absolute Difference
Constrained Statistical Matching: Demographics only	-0.23	0.28
Constrained Statistical Matching: Demographics + Magazine Readership	-0.13	0.22

10. REPEATED SAMPLING

So far, we have acted as if one single split-sample was adequate. The split-sample assignment is usually done randomly (such as by an odd-even assignment). As such, it will incur sampling error in the sense that we cannot expect all the variables to have identical joint distributions in the two split-samples.

Consider an example in which a target group has an incidence of 10% in the total sample. After randomly splitting the sample into two halves, one half-sample has an incidence of 9% and the other half-sample has an incidence of 11%. Under a constrained fusion, the original and fused incidences must have a 2% difference. Any variable measured based upon the comparison of fused versus original values will reflect this difference, which is due to the randomness of the split sampling and not necessarily because of any deficiency in the fusion methodology.

These random fluctuations can be evened out through repeated sampling. As an illustrative example, we ran 10 different random split-samples on the MARS database for the target group television ratings in Application A. Here is one case example for illustrative purposes.

Table 8. Example: Acid reflux sufferers who viewed ESPN2

Repeated Sample	Original TGR	Fused TGR	Index (=Fused/Original)
1	20.7	18.9	88
2	20.6	19.4	94
3	20.9	20.1	96
4	20.8	20.4	98
5	20.7	20.9	101
6	20.7	20.9	101
7	20.9	21.3	102
8	20.8	21.4	103
9	20.7	22.1	107
10	20.5	22.2	108
Pooled average	20.7	20.7	100

Suppose that just one split-sample was executed. One could draw the 20.7/18.9 pair for an index of 88 for one conclusion of under-estimation, or the 20.5/22.2 pair for an index of 108 for the opposite conclusion of over-estimation. The sole difference is due to sampling error during the construction of the split-samples. When we pooled the results from the ten repeated samples, there was in fact no difference in the pooled averages. Overall, based upon our experience, our recommendation is that the result for an individual entity (such as a target group rating) on the basis of a single split-sample foldover test should not be trusted.

However, a single split-sample is adequate when we are evaluating averages taken over a large number of entities. For example, the averages in Table 3 in Application A , which are each taken over 2,520 different target group/television measures, are robust across repeated sampling with the same conclusions. Indeed, in Table 3 in Application A, when we showed the mean difference for constrained statistical matching to be 1.78, the range of the ten repeated samples is within 0.05 of this value. Thus, any of the 10 repeated samples would have led us to the same conclusions about the relative effectiveness of those data fusion methods. This applies to all numbers shown in Applications A through D in this paper.

When repeated samples of split-sample foldover test results are available, we can perform another type of analysis --- the mean-squared-error analysis.

For example, in the context of Application A (target group ratings by data fusion), the difference between the pooled original target group rating and the pooled fused target group rating is a measure of the bias of the fused target group rating. The repeated sample values of the fused target group rating can be used to estimate the variance of the fused target group rating.

The Mean-Squared-Error (MSE) is defined as $(Bias)^2$ + Variance. More commonly, it is reported in square-root form as the Root-Mean-Squared-Error (RMSE).

This is a seldom reported aspect in the suite of validation criteria, probably because it involves a lot more work. A synthetic estimate, such as that obtained via data fusion or integration, is a complex estimation method. Although the end result is a percentage (e.g. a target group rating), the usual simple random sample formula cannot be used to compute a standard error for the estimate. There is in fact no analytical formula that can be used. Instead, the standard error must be determined by some sort of empirical repeated sampling technique.

For illustrative purposes, we consider the comparison of the root-mean-squared-error (RMSE) for two methods. The database is the 2002 MARS database and the domain of application is the target group television ratings. Our first method is the constrained statistical matching based upon 21 common variables.

Our second method is the simulation method (see Papazian (1980), Cannon (1988) and Cannon and Seamons (1995). This was a method that was fairly common at a time when no obvious alternatives were available. In this method, we assumed that the distribution of product usage and media usage is random within 12 age/sex groups, such that the number of product users within each media audience can be obtained by the direct multiplication of the product incidence into the media audience in that age/sex group. This is not a classical data fusion method because no respondent-level database is physically produced. It may be described as a data integration technique based upon ratio estimation. There is some evidence that this method results in severe biases in some cases (Cannon (1988) and Cannon and Seamons (1995)) but it seems to be acceptable in other cases (Danaher and Rust (1992)).

Our interest in this particular comparison is motivated by what Cannon and Seamons (1995) wrote: "Data fusion has become increasingly popular among major media data suppliers, but relatively few scholarly papers have been written on the subject. These were all written by practitioners. Significantly, none of them referred to alternative methods of linking data sets or to the theoretical limitations of the method. None of them relate data fusion to the simulation approach, much less reference to its potentially fatal flaws."

We ran 10 repeated split-samples on the MARS 2002 database and obtained target group ratings under these two methods. Across 40 target groups and 63 television measures, we obtained the average results shown in the next table. Of the 40 x 63 = 2,520 cases, data fusion has the lower RMSE in 69% of the cases.

Table 9. Mean-Squared-Error summary results for two data fusion/integration methods

Method	Mean bias	Mean variance	Mean RMSE*
Data fusion	2.50	1.24	3.32
Simulation method	3.88	0.89	4.90

* The individual Mean-Squared-Errors are averaged first before applying the square root operation at the last step.

By target group, data fusion has a smaller RMSE in 27 out of the 40 target groups. The simulation method is better when the target group is strongly driven by age/sex (e.g. erectile difficulty/dysfunction, hangover, menopause/hormone replacement, urinary tract infection, yeast infection).

Data fusion is a more complex procedure that leverages many more matching variables beyond just 12 age/sex groups. As expected, data fusion reduces the biases in the estimates. In so doing, the complexity of the procedure has increased the sampling variance of the estimate somewhat. But the net tradeoff is that data fusion incurs lower RMSE.

The Rässler typology refers to the accuracy in terms of the bias of the appropriate measures. The total error of an estimate is decomposed into bias and variance. To estimate the variance of those measures, we can run split-sample foldover tests repeatedly.

This leads to the obvious question, "How many repeated samples do we need to run?" The answer turns out to depend on the specific measures, as some measures are very stable whereas others are highly volatile. For example, the individual-level measures in Application C (Predictive Modeling) were found to be very stable, whereas the target group television-magazine duplications in Application B (Target Group Multimedia Television/Print Analysis) are more volatile. The users will need to run as many repeated samples as they need to make sure that they have acceptable variance estimates.

11. DISCUSSION

The split-sample foldover test is a major tool for assessing the accuracy of techniques such as data fusion/integration, ascription and predictive modeling, as this is an empirical procedure that does not depend on unverifiable assumptions.

If we begin with a realistically representative split-sample setup, there are many different evaluative criteria that can be applied. The Rässler (2002a) classification scheme presents four different levels of evaluative criteria. However, the totality of all these evaluative criteria would result in information overload. Therefore, it is important to focus only on those evaluative criteria that are relevant to the application at hand.

We used four common applications to illustrate how the relevant criteria are selected. In each case, we looked carefully at how the end users use the data in practice and our evaluative criteria mirror those needs exactly. As for any other criteria, we believe that information should not be produced just because it can be. Rather information should be produced because it informs.

BIBILOGRAPHY

Antoine, J. (1985) A case study illustrating the objectives and perspectives of fusion techniques. *Proceedings of the Salzburg Readership Symposium*, Salzburg (Austria).

Baim, J. and Frankel, M.R. (1997) Enhanced ascription. *Proceedings of the Magazine Audience Measurement Research*. Advertising Research Foundation: New York City, USA.

Baker, K., Harris, P. and O'Brien, J. (1989) Data fusion: an appraisal and experimental evaluation. *Journal of the Market Research Society*, 31(2), 153-212.

Bals, W. (2002) Controlled split survey in media practice. *IMPUTE: Symposium on Nonresponse, Questionnaire Split and Multiple Imputation*, Nüremberg (Germany), September 25.

Baynton, P. (2003) Data integration or fusion? *ARF/ESOMAR Week of Audience Measurement (Mixed Media Session)*, Los Angeles (USA).

Bedwell, R. (1991) Fusion – Britain's latest experience. *Fifth Worldwide Readership Research Symposium*, Hong Kong.

Bennike, S (1985) Fusion – an overview by an outside observer. *Proceedings of the Salzburg Readership Symposium*, Salzburg (Austria).

Cannon, H. M. (1988) Evaluating the 'simulation' approach to media selection. *Journal of Advertising Research*, 28(1), 57-63.

Cannon, H.M. and Seamons B.L. (1995) Simulating single-source data: how it fails us just when we need it most. *Journal of Advertising Research*, 35(6), 53-62.

Carpenter, R. and Wilcox, S. (1995) Data fusion in the British National Readership Survey – an experiment. *Seventh Worldwide Readership Research Symposium*, Berlin (Germany).

Collins, J.H., Mallett, D.T. and Mulligan Traub, J. (2002) Multi-media reach/frequency and optimization: Questions, answers and consequences for print. *ESOMAR/ARF Week of Audience Measurement*, Cannes (France).

Czaia, U (1993) Interactive fusion: step two. *Sixth Worldwide Readership Research Symposium*, San Francisco CA (USA), 489-493.

Danaher, P. J. and Rust, R. T. (1992) Linking segmentation studies. *Journal of Advertising Research*, 32(3), 18-23.

de Montigny, M. and Lima, A.L. (2001) Brazil fusion and multi-media duplication. *Tenth Worldwide Readership Research Symposium*, Venice (Italy), 555-558.

Feynman, R. (1997) *"Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character*. W.W. Norton & Company: San Francisco (USA).

Frankel, M. R. (1981) Ascription in magazine audience research. *Readership Research: Theory and Practice. Proceedings of the First International Symposium*. H. Henry (ed.). New Orleans (USA).

Frankel, M.R. and Baxter, P. (1988) Fusion, integration, ascription and imputation. *Readership Research: Theory and Practice. Proceedings of the Fourth International Symposium*. H. Henry (ed.). Barcelona (Spain).

Mallett, D. (1997) Ascription: there's still no such thing as a free lunch. *Proceedings of the Magazine Audience Measurement Research*. Advertising Research Foundation: New York City (USA).

Page, K. (2001) Personalised media lists. *Tenth Worldwide Readership Research Symposium*, Venice (Italy).

Papazian, E. (1980) Using product usage data in media selection. *Marketing and Media Decisions*, July issue.

Rässler, S. (2002a) *Statistical matching : a frequentist theory, practical applications, and alternative Bayesian approaches*. Springer-Verlag New York: New York (USA).

Rässler, S. (2002b) Split questionnaire survey sampling. *IMPUTE: Symposium on Nonresponse, Questionnaire Split and Multiple Imputation*, Nüremberg, (Germany), September 25.

Raimondi, D. and Santini, G. (1997) Just-in-time data modeling. *Eighth Worldwide Readership Research Symposium*, Vancouver (Canada).

Raimondi, D. and Santini, G. (2001) Fusion quality assurance. *Tenth Worldwide Readership Research Symposium*, Venice (Italy), 111-118.

Scheler, H.-E. and Wiegand, J. (1985) A report on experiments in fusion in the official German Media Research (AG:MA). *Proceedings of the Salzburg International Readership Symposium*, Salzburg (Austria).

Scheler, H.-E. and Wiegand, J. (1987) A report on experiments in fusion in the "official" German media research (AG.MA). In Henry, H. (ed), *Readership Research: Theory and Practice*, 352-360, Elsevier Science: Amsterdam (The Netherlands).

Soong, R. and de Montigny, M. (2001) An anatomy of data fusion. *Tenth Worldwide Readership Research Symposium*, Venice (Italy), 87-109.

Soong, R. and de Montigny, M. (2002**)** The contribution of magazines in mixed TV-print schedules. *ESOMAR/ARF Week of Audience Measurement*, Cannes, France. Also reprinted in *Excellence 2003 International Research* (2003). ESOMAR: Amsterdam, (The Netherlands).

Soong, R. and de Montigny, M. (2003a) Does fusion-on-the-fly really fly? *ARF/ESOMAR Week of Audience Measurement (Mixed Media Session)*, Los Angeles (USA), 183-204.

Soong, R. and de Montigny, M (2003b) Fusion-on-the-fly for multimedia applications. *Eleventh Worldwide Readership Research Symposium*, Boston (USA).

Speetzen, R. (1988) The art of models – ascription in Germany. *Readership Research: Theory and Practice. Proceedings of the Fourth International Symposium*. H. Henry (ed.). Barcelona (Spain).

Walsh, P. (2001) Multibasing: Data integration without regression to the mean. *Tenth Worldwide Readership Research Symposium*, Venice (Italy), 57-67.

Weiss, S.M. and Indurkhya, N. (1998) *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers Inc.: San Francisco (USA).

Wilcox, S. and Johnson, H. (1997) Multi-media reach and frequency analysis. *Eighth Worldwide Readership Research Symposium*, Vancouver (Canada).