THE ANATOMY OF DATA FUSION

Roland Soong, Kantar Media Research
Michelle de Montigny, Kantar Media Research and TGI Latina

1. ABSTRACT

This paper is about the fusion of a television people meter panel and a magazine readership/product usage survey in Mexico. During the process, two different methods of data fusion --- unconstrained and constrained statistical matching --- were evaluated on a number of criteria: computational complexity, donation frequency, sample sizes, matching success rates, preservation of audience and incidence levels, together with a number of individual-level and aggregate-level diagnostic tools. Collectively, these results provide a detailed look into the inner workings of data fusion.

We do not advocate the inherent superiority of either constrained or unconstrained matching. So our discussion will focus on how to make decisions based upon the trade-offs among various factors in a given situation, and this will therefore be of interest to researchers everywhere.

2. DESCRIPTION OF THE DATABASES

We report here the results of the fusion of two databases in Mexico.

The first database is the television people panel operated by IBOPE AGB Mexico. We will refer to this database by the acronym TAM (for Television Audience Measurement). This TAM database contains household tuning and viewing by persons age 4 or older on a continuous basis. On each day, the intab sample is weighted for reporting purposes. The fused database contains the viewing for persons between the ages of 12 to 64 years old inclusive. For the fusion study, we used the TAM panel during the one-week period of July 10-16, 2000.

The second database is the TGI (for Target Group Index) study conducted by Kantar Media Research and Moctezuma Asociados. This is a survey of persons 12-64 years old, and covers newspaper, magazines, television, radio, internet, cinema, demographics, product usage, lifestyle and psychographics. The data are collected in a personal interview for the media and demographics, and a self-administered booklet for the rest of the information. The intab sample is weighted for reporting purposes. The fused database contains magazine readership and a select list of product usage variables for those persons 12 to 64 years old who live in television homes.

The TAM and TGI studies have the same geographical coverage (namely, the 27 Mexican cities with populations of more than 400,000). However, the two studies have different sample allocations, including these important differences: (1) the TAM sample is allocated with broad national coverage, but with sufficient sample to report the three major cities of Mexico City, Guadalajara and Monterrey separately; (2) the TAM sample contains two booster pay-TV samples inside Mexico City financed by the two operators Cablevisión and Multivisión; (3) the TGI sample has disproportionately higher allocations in Mexico City, Guadalajara and Monterrey in order to report on the local media (especially newspapers and radio); (4) the TGI sample has disproportionately higher allocations in upper-class areas and lower allocations in lower-class areas. The TAM and TGI have been weighted appropriately to account for these disproportionate allocations, but their weighting structures and universe estimates are not identical.

The results here are based upon 7,385 TAM respondents and 10,954 TGI respondents, all of whom are between 12 and 64 years old inclusive and living in television homes.

3. DATA FUSION ALGORITHMS

The number of algorithms that can be used for data fusion is limited only by human imagination. Within each algorithm, there may also be infinitely many variations in the details. For the purpose of this paper, we will be comparing two major algorithms. These two algorithms fall into the class of statistical matching algorithms. To emphasize the difference between them, we will refer to them as unconstrained statistical matching and constrained statistical matching. These names are not commonly used in the data fusion literature, but they are standard terms in the statistical matching literature (see Rodgers (1984)).

Unconstrained Statistical Matching

The TAM database is designated as the recipient database, and the TGI database is the donor database. For each person in the recipient database, we examine the TGI database for the best donor. The definition of 'best' is defined in terms of a distance metric based upon a weighted sum of deviations over a list of variables (age, sex, socio-economic level, etc). The details on the variables and the distance metric are given in the next section.

The distance between a recipient and a potential donor has a value of zero when they match each other on every variable. When there is no perfect match in the TGI database, the best donor is the TGI respondent who matches on the highest weighted number of variables, where the weights of the variables represent their importance towards the accuracy of the fusion. This is a weighted nearest-neighbor technique (see Dasarathy (1991)).

In addition, a penalty function is imposed by subtracting the number of previous donations by the TGI respondents from the distance value. This has the effect of distributing the donations among TGI respondents with similar characteristics instead of using one of them much more often than others.

We point out some characteristics of unconstrained statistical matching. First of all, the recipient database assumes primacy in the fusion. For each recipient, one (and only one) donor is identified, and that donor's information is then transferred over to the recipient. For example, if we start off with 10,000 recipients and 30,000 donors, then the fused database will contain information from only 10,000 donors and less if multiple donations occur. This has the drawback of losing some sample size and data.

Secondly, there is no guarantee that the donated information from the 10,000 people will look identical to that for the full set of 30,000 people. Here, we mean information such as product usage incidences as well as audiences, profiles, duplications, reach and frequency for magazines.

Thirdly, this implementation of unconstrained statistical matching does not make use of the case weights in the TAM and TGI databases. Those case weights were present in the original studies to account for certain sample design features such as disproportionate sample allocation and differential response rates. The omission may be a cause for concern and we will discuss this point later.

The best published reference on unconstrained statistical matching is the paper by Baker, Harris and O'Brien (1989). The best full-fledged commercially deployed instance is the one developed by RSMB originally for Granada Television and refined for the BARB/TGI fusion. We note that there are some important differences between our implementation here versus the RSMB version (especially with respect to their use of the weighted Mahalanobis distance, their use of ANOVA to derive importance weights and their choice of penalty function).

Constrained Statistical Matching

Our second algorithm is the constrained statistical matching. The first appearance of this type of approach in statistical matching is the work of Barr and Stewart (1979) at the U.S. Department of the Treasury. It was also reviewed in detail by Rodgers (1984). We are aware of only one application of this algorithm to the data fusion of media research databases, namely the A.C. Nielsen New Zealand Panorama service, of which the most detailed published description is in James Reilly's Master of Science thesis.

The backbone of constrained statistical matching is in the transportation problem that belongs to the area of operations research. The problem was originally formulated by Hitchcock (1941) as follows: how to ship the supply of homogeneous units of a product stored in m points of supply (sources) to meet the demand from n points of demand (destinations) with minimum total shipping cost.

The standard algorithm for solving the transportation problem is the stepping stone algorithm, which was developed by Charnes and Cooper (1954) as a special case of the primal simplex algorithm for solving linear programming problems. This stepping stone algorithm is described in classics like Dantzig (1963) and Hillier and Lieberman (2000). We implemented our own computer code for the data fusion. Commercial software for solving the transportation problem is available in the SAS/OR product from SAS Institute Inc., and as C/FORTRAN subroutines from NAG (Numerical Algorithms Group), among others.

In the context of data fusion, the transportation problem is cast as follows: find the minimum cost solution to shipping m TGI respondents to n TAM respondents, with their case weights representing supply and demand respectively and with the shipping cost between a TAM respondent and a TGI respondent being equal to the 'distance' between them.

The description of the stepping algorithm is too long for us to provide in this paper. The interested reader can consult Dantzig (1963) and Hillier and Lieberman (2000) for the details. We will present a simple illustrated example:

For the TAM database, we assume that there are only two persons, with IDs TAM1 and TAM2. The first person has a case weight of 3,000 and the second person has a case of 3,000 as well. The total weight is 6,000. On the matching variable MV (e.g. head of household status), the first person is a YES and the second person is a NO. On the television variable TV1 (e.g. watch the Oscar awards show), the first person is a YES and the second is a NO.

For the TGI database, we assume that there are only three persons with IDs TGI1, TGI2 and TGI3. These persons each have case weights of 2,000. The total weight is 6,000, same as in the TAM database. On the matching variable MV, the first two persons are YES and the third person is NO. On the product usage variable P1 (e.g. own credit card), the first person is YES and the other two persons are NO.

TAM Database

TAMCase	Weight	MV	TV1
TAM1	3000	Y	Y
TAM2	3000	N	N

TGI Database

TGICase	Weight	MV	P1
TGI1	2000	Y	Y
TGI2	2000	Y	N
TGI3	2000	N	N

Fused Database

TAMCase	TGICase	Weight	MV	TV1	P1
TAM1	TGI1	2000	Y	Y	Y
TAM1	TGI2	1000	Y	Y	N
TAM2	TGI2	1000	N/Y	N	N
TAM2	TGI3	2000	N	N	N

Explanation: MV = Matching variable; TV1 = TAM TV variable; P1 = TGI product variable

For the constrained statistical matching, we can start with person TAM1. We look for a matching person on the TGI side and we find that TGI1 is a perfect match on MV (note: TGI2 is also a match and such ties can be broken arbitrarily). However, TAM1 has a case weight of 3,000 while TGI1 has a case weight of 2,000. So it is possible only to 'ship' 2000 for now. In the fused database, the first fused record came from TAM1 and TGI1 with case weight of 2,000 with their respective contributed data.

At this point, TGI1 is completely accounted for but there are still 1,000 units of TAM1 left. If we look for a match again, we find that TGI2 is a perfect match. In the fused database, the second fused record came from TAM1 and TGI2 with case weight of 1,000 with their respective contributed data.

At this point, TAM1 is completely accounted for. As for TAM2, the fusion is necessarily done with TGI3 and the remainder of TGI2. This completes the fused database.

We will list the most important properties of the resulting fused database:

Firstly, the transportation problem is symmetrical in nature. The notions of donor versus recipient do not exist, as both databases are equal partners.

Secondly, the algorithm results in each TAM respondent being matched to one or more TGI respondents, and each TGI respondent being matched with one or more TAM respondents. The fused database consists of fractional records of the original people. By virtue of the property known as unimodularity, when *m* TGI respondents are fused with *n* TAM

respondents, the fused database will consist of exactly $m+n-1$ fractional records, containing the fractional weight, the TAM data from the TAM respondent and the TGI data from the TGI respondent.

Thirdly, within this fused database of $m+n-1$ fractional records, every original TAM respondent and every original TGI respondent are represented here in their original proportions in the sense that the sum of the weights of their fractional records add up to the original weight. There is no loss in sample size or data involved. It is for this reason that this method of statistical matching is characterized as being 'constrained' whereas our other algorithm is 'unconstrained' in the sense that not all persons in the donor database are included and not necessarily in their original proportion even if they are included.

Fourthly, within this fused database, all TAM information is completely preserved. By this, we mean all television ratings, duplications, gross rating points, reach and frequency for all stations and programs and for the total population as well as by standard TAM subgroups. This is guaranteed because the fusion has only fractionated the original TAM records, without dropping cases, changing case weights or modifying any data.

Fifthly, within this fused database, all TGI information is completely preserved. By this, we mean all product usage incidences as well as their duplications as well as all the total audiences, compositions, duplications, gross ratings points, reach and frequency for all magazine titles and for the total population and for any standard TGI characteristics. This occurs because the fusion has only fractionated the original TGI records, without dropping cases or modifying any data.

Having said this, we will now modify the last assertion from 'completely preserved' to 'almost completely preserved.' There is a couple of minor issues that prevents the preservation of TGI data, as a consequence of the fact that the TAM and TGI databases are not completely aligned. If the two databases are reasonably well-aligned, these discrepancies should have only minor impact. Still, we feel that we need to point them out.

The first issue relates to projected totals. It may happen that the projected total number of adults in television households is 30,000,000 in the TAM database and 30,002,000 in the TGI database. To reconcile the differences, we multiply each case weight in the TGI database by a factor of 30,000,000 / 30,002,000 so that the revised TGI database will now project to 30,000,000 as well. Among other things, this means that magazine audience totals will change by that factor as well.

The second issue relates to mismatches in common variables. For example, suppose the incidence of telephone ownership is 59% in the TAM database and 60% in the TGI database. At best, constrained statistical matching will result in 1% of the weighted records containing no phone in the TAM portion and with phone in the TGI portion of the fused record. In this case, we opt to report the TAM value in the fused database. This means that any relationship between TGI phone ownership and other TGI variables will not be preserved in this fused database.

In the statistical matching literature, the review paper by Rodgers (1984) stands out as the most comprehensive discussion of unconstrained and constrained statistical matching. Rodgers concluded that "unconstrained matching procedures have the advantage of relative simplicity and lower costs in terms of computer processing time and memory requirements, at least compared with the constrained optimization matching procedure … The disadvantages of unconstrained procedures, however, are many … Constrained matching procedures avoid the problems just described for unconstrained matching procedures and produced considerably more accurate estimates of covariances, regression coefficients, and of the proportion of explained variance in regression analyses. The costs of carrying out a constrained matching may be considerable, however." (p.99)

Rodgers published his review in 1984 and the computer-related costs may no longer be significant today. As much as Rodgers seemed to endorse the superiority of constrained statistical matching, we would withhold our judgment since his empirical data came from domains other than media research. This conservatism is justified by the famous 'No Free Lunch Theorem': "… there are no context-independent or usage-independent reasons to favor one learning or classification method over another. If one algorithm seems to outperform another in a particular situation, it is a consequence of its fit to the particular pattern recognition problem, not the general superiority of the algorithm. When confronting a new pattern recognition problem, appreciation of this theorem reminds us to focus on the aspects that matter most --- prior information, data distribution, amount of training data, and cost or reward functions. This theorem also justifies a healthy skepticism regarding studies that purport to demonstrate the overall superiority of a particular learning or recognition algorithm." (Duda, Hart and Stork (2001), p.454-455).

It is the stated purpose of this paper to compare the relative merits of unconstrained and constrained statistical matching in the fusion of two specific media research databases, without any ingoing presumptions.

4. VARIABLES & DISTANCE METRICS

The two data fusion algorithms that we wish to study are based upon the principles of statistical matching. For a person in one database, we wish to find one or more persons in the other database who match according to a list of variables. Clearly, these matching variables must be collected in both databases, with the same definitions.

Across the TAM and TGI databases, the following variables exist in common: geography (Mexico City, Monterrey, Guadalajara and the balance of 24 other cities), pay television (cable/satellite/microwave antenna services, including the distinction between Cablevisión and Multivisión in Mexico City), gender, age (12-17, 18-24, 25-34, 35-44, 45-54, 55-64), socio-economic level (ABC+, C, D+, DE), head of household status, housewife ("*ama de casa*") status, presence of children 0-11, single vs. multi-TV set and presence of telephone. These are the variables that were used in the statistical matching. The databases have some other common variables (e.g. the number of light-bulbs in the home, the number of showerheads in the home, ownership of a vacuum cleaner, which are components of the socio-economic level variable), but they were not used.

In statistical matching, it is common to divide the matching variables into critical variables on which people must match and non-critical variables on which it is desirable, but not essential, to match. The choice of which variables is essentially subjective, but supported by statistical analyses and business considerations. For our data fusion here, the critical variables were chosen as follows.

Geography The TGI Mexico study was designed to provide local information separately by Mexico City, Monterrey, Guadalajara and the balance of 24 other cities. In the three large cities, the TGI Mexico study provides local media planning information for television stations, newspapers, radio stations, local brands, etc. It would be unacceptable for matched respondents to come from different geographical regions, since they would be using media and products that are not readily available to them.

Gender The TGI Mexico study contains a number of gender-based product categories, such as women's beauty aids and female hygiene products. It would be an absurdity for men to be using women's products.

Pay TV It is known that only about 50% of viewing in pay-TV homes go to broadcast channels, whereas 100% of viewing in non-pay-TV homes go to broadcast channels by definition. Moreover, the TAM panel carries booster samples paid for by Cablevisión and Multivisión inside Mexico City, who require their TV ratings information be reported separately.

As for the non-critical variables, it is often impossible to find a match on all of them, which means that we have to settle for a less-than-perfect match. In that case, it is necessary to prioritize the non-critical variables in some fashion, so that we match on the more important variables.

The above requirements are encapsulated in a distance metric used to compare any person in one database against a candidate for a match in the other database. This distance metric is defined as follows:

1. The candidate must come from the same geographic region (namely, Mexico City, Guadalajara, Monterrey or the balance of 24 cities)
2. The candidate must have the same pay-TV status (namely, Cablevisión, Multivisión or non-pay TV in Mexico City; pay-TV versus non-pay-TV outside Mexico City)
3. The candidate must be of the same gender
4. If the candidate satisfies (1) through (3) above, set the initial distance to zero; if the candidate does not satisfy (1) through (3), then he/she is not eligible to become a match
5. If the candidate is two or more socio-economic levels away, then the distance is increased by 16
6. If the candidate is only one socio-economic level away, then the distance is increased by 10
7. If the candidate is more than two age groups away, then the distance is increased by 16
8. If the candidate is only one age group away, then the distance is increased by 10
9. If the candidate has a different head of household status, then the distance is increased by 12
10. If the candidate has a different housewife status, then the distance is increased by 12
11. If the candidate has a different presence of children status, then the distance is increased by 8
12. If the candidate has a different multi-TV set status, then the distance is increased by 2
13. If the candidate has a different presence of telephone status, then the distance is increased by 1

A perfect match would be someone with a distance of zero (that is, matching on all the variables). This particular distance metric is just a subjectively weighted sum. Obviously, this is mildly disconcerting since different people will come up with different weights. The experiment reported in Appendix A should assure the reader that the performance of the fusion is rather robust with respect to the specific choice of a distance metric.

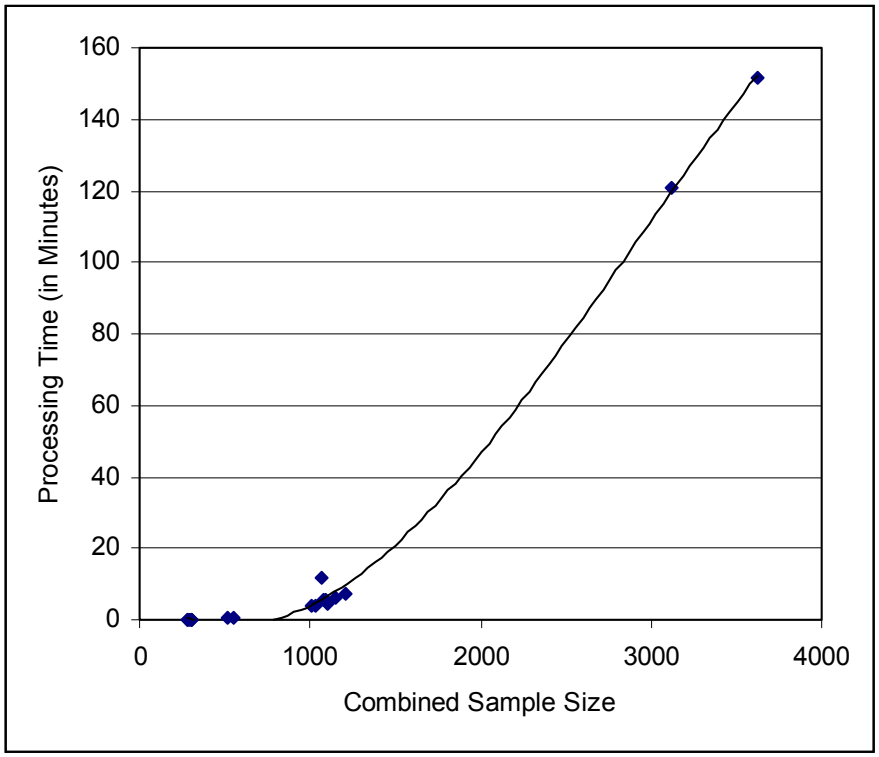
5. COMPUTATIONAL COMPLEXITY

In this section, we will discuss the theory (see Garey and Johnson (1979)) and empirical results for the time computational complexity of the two statistical matching methods. Our benchmark machine is a Dell 550MHz Pentium III PC with 128M memory. At this time, this is in fact a low-end machine and the current top-end machines are three to four times faster.

For unconstrained statistical matching, we compared each of the *n* persons in the TAM database against each of the *m* persons in the TGI database. Therefore, the time computational complexity of our algorithm is of order O(*mn*), where *m* and *n* are the respective sample sizes of the two databases. Actually, our algorithm is faster than that because the problem was structured as 18 segments with the comparisons being necessary only for those cases that fall within the same segment. The total time taken for unconstrained statistical matching is less than 10 seconds. Timing is therefore not an issue here.

For constrained statistical matching, the stepping stone algorithm for the transportation problem has exponential computation complexity in the worst case (see Papadimitriou and Steiglitz (1998)). In practice, the actual performance will be specific to a problem. The total time taken here was 4 hours 56 minutes 31 seconds. In Graph 5.1, we have plotted the processing time against the combined sample size for the 18 segments. The fitted curve is a cubic polynomial. Thus, the empirical performance of the algorithm exhibits time computational complexity of $O(n^3)$; that is to say, doubling the sample size leads to an eightfold increase in processing time.

Graph 5.1. Processing Time (in Minutes) By Total Sample Size for 18 Segments



If the data fusion is interactive in nature, the timings for constrained statistical matching are problematic. But if the data fusion is produced on a delayed basis, as in our actual operations, this is not an impediment.

The performance of constrained statistical matching can be significantly improved with a couple of simple moves. First of all, developing a good initial solution instead of the random one right now will reduce the number of steps needed to reach the optimal solution. This can be achieved as easily as pre-sorting the databases. Secondly, sub-dividing the largest segments will save time, because we know that halving the segment size results in an eightfold reduction in processing time. Thirdly, the 18 segments can be processed simultaneously so that the total elapsed time equals the time needed to process the largest segment. By way of comparison, we report that our Brazilian fusion takes less than 10 minutes on the same machine after following some of these recommendations.

6. DONATION FREQUENCY

For constrained statistical matching, this topic has no meaning. Every respondent appears exactly once, with the same weight as in the original study.

Under unconstrained statistical matching, we seek for each TAM recipient one (and only one) best matching TGI donor. We also imposed a penalty in the scoring scheme against multiple donations. In this fusion, we have 7,385 TAM respondents acting as recipients, and 10,954 TGI respondents acting as donors. Under a simple random model, we would expect that the average donation frequency = 7385 / 10854 = 0.67 times, with 67% of the TGI respondents donating once and the other 33% of TGI respondents not donating at all.

In our actual unconstrained statistical matching, we found that 54% of the TGI respondents were 'unmarried' (that is, they did not donate at all), 34% were 'monogamous' (that is, they donated exactly once) and 12% of them were 'polygamous' (that is, they donated more than once). It is disconcerting to find that so many TGI respondents were ignored altogether even as others were being used multiple times. This happened because the matching had to occur separately within the 18 segments.

In Table 6.1, we show the TAM and TGI sample sizes by segment, with the mean donation frequency being the ratio of the two. In some segments (such as the non-pay TV segments in Mexico City), we have huge surpluses of donors so that 'polygamy' is not an issue whereas the 'unmarrieds' seemed wasteful. In other segments (such as those non-pay TV people outside of Mexico City), we have many more recipients than donors so that 'polygamy' becomes unavoidable.

Table 6.1 Distribution of TAM and TGI sample sizes by segment

Segment	TAM sample size	TGI sample size	Mean donation frequency
Mexico City Cablevisión males	355	744	0.48
Mexico City Cablevisión females	327	709	0.46
Mexico City Multivisión males	232	318	0.73
Mexico City Multivisión females	242	277	0.87
Mexico City non-pay TV males	866	2261	0.38
Mexico City non-pay TV females	929	2693	0.34
Guadalajara pay-TV males	109	192	0.57
Guadalajara pay-TV females	135	144	0.94
Guadalajara non-pay-TV males	560	516	1.08
Guadalajara non-pay TV females	557	540	1.03
Monterrey pay-TV males	78	228	0.34
Monterrey pay-TV females	86	201	0.43
Monterrey non pay-TV males	588	422	1.43
Monterrey non pay-TV females	676	535	1.26
Balance pay-TV males	133	177	0.75
Balance pay-TV females	139	157	0.89
Balance non pay-TV males	673	392	1.72
Balance non pay-TV females	700	448	1.56
TOTAL	7385	10954	0.67

Multiple donations are considered undesirable because of the reduction in sample size and the implied additional weighting, ultimately causing some loss in statistical reliability. We can reduce the instances of multiple donations by collapsing some of the segments. But given our reasoning for choosing the segments, it would seem improper to accept donors from other segments (for example, male-to-female), because we will only be delivering the wrong information with greater reliability.

7. MATCHING SUCCESS RATES

Statistical matching is about finding respondents who match on a list of variables. It may turn out that there is no perfect match in some instances, so that respondents are brought together with some differences in the matching variables.

In Table 7.1, we show the percent of cases in which variables were successfully matched for the two statistical matching methods. The first four variables are guaranteed to 100% successfully matched, because they were the critical variables.

TABLE 7.1 Matching Success Rates for Unconstrained and Constrained Statistical Matching

Matching Variables	% Successfully Matched under Unconstrained Statistical Matching	% Successfully Matched under Constrained Statistical Matching
Presence of television	100 %	100 %
Geographical region	100 %	100 %
Pay TV (including CV and MV)	100 %	100 %
Gender	100 %	100 %
Socio-economic level	97 %*	93 %**
Age	97 %*	86 %***
Head of family	98 %	90 %
Housewife	97 %	89 %
Presence of children 0-11	98 %	88 %
Multi-set television household	90 %	83 %
Presence of telephone	83 %	82 %

* Remaining cases are in adjacent groups
** Another 5% in an adjacent socio-economic group
*** Another 11.0% in an adjacent age group

From Table 7.1, we see that the success rates are higher for unconstrained statistical matching. This is necessarily true since constrained statistical matching is defined as finding the best statistical match subject to constraints whereas unconstrained statistical matching is defined as finding the best statistical match without constraints. At best, constrained statistical matching can only do as well as, but never better than, the unconstrained one.

In addition, when the two samples differ in composition, constrained statistical matching will fail by at least the size of the discrepancy. For example, if one sample has 50% males and the other sample has 49% males, then constrained statistical matching will have an error of at least 1% on gender.

To the extent that we believe that these matching variables are significant to the fused variables, it was obviously desirable to have the highest possible successful matching rates. To the extent that the matching was less than perfect, the question has to be about the potential impact on the accuracy of the fused data. In Appendix B, we present the details of a study about the incremental explanatory power of the matching variables. We found that there are only some small improvements in overall accuracy after the first 5 or 6 variables. We use those findings to infer that the differences in matching success rates between the two methods do not affect the overall accuracy of fusion.

8. COMPARISON OF INCIDENCES

Under both forms of statistical matching, the TAM data are preserved completely. This implies that all television ratings, audience profiles, duplication and other ancillary statistics remain the same as before the fusion. The same situation does not exist for the TGI data. Unconstrained statistical matching is a form of random sampling, which means that the results are subject to random error; furthermore, if the matching variables are chosen or employed improperly, there may be systematic biases as well.

In theory, constrained statistical matching was designed to preserve all TGI data. In practice, there are some discrepancies due to differences in sample composition between the TAM and TGI studies. For example, consider what might happen in one of the 18 segments: it may turn out that the projected number of persons is 200,000 in the TAM database and 200,500 in the TGI database. There are many reasons: lack of explicit weighting to the segment total, different universe estimates being used, etc. The transportation problem requires that the supply must equal to demand, so we had to re-scale the weights in the TGI database (that is, multiply every weight by 200,000 / 200,500). As a result, the TGI data will not be perfectly preserved, but we do expect the results to be quite close.

For the 270 product variables in the TGI study, the mean incidence was 28.87%. The mean fused incidence was 29.41% under the unconstrained statistical matching and 28.83% under the constrained statistical matching. The mean absolute deviation between the original and fused incidences was 0.85% under the unconstrained statistical matching and 0.09% under the constrained statistical matching.

The discrepancies in product incidences from the constrained statistical matching are negligible in size. The incidences from the unconstrained statistical matching show a systematic upward bias. This does not seem to be a major issue for incidences of product usage.

For the 98 magazines in the TGI study, the mean average audience was 1.066%. The mean fused average audience was 1.180% under unconstrained statistical matching, and 1.064% under constrained statistical matching. The mean absolute deviation between the original and fused average audiences was 0.14% under unconstrained statistical matching, and just 0.01% under constrained statistical matching. To make these numbers more concrete, we show the results for the top 10 magazines in tables 8.1 and 8.2.

TABLE 8.1 Original vs. Fused (Constrained Statistical Matching) Average Audiences for Top 10 Magazines

Magazine Title	Fused AA %	Original AA%	Difference	Relative Difference%
Reader's Digest (Selecciones)	8.38	8.44	-0.06	-0.7%
Eres	7.74	7.72	0.02	0.3%
Muy Interesante	7.51	7.58	-0.07	-0.9%
TV y Novelas	7.50	7.55	0.04	-0.5%
Tele Guía	5.08	5.09	-0.01	-0.2%
Vanidades	5.07	4.97	0.09	1.9%
TV Notas	3.87	3.84	0.03	0.9%
Conozca Más	2.77	2.80	-0.02	-0.9%
Proceso	2.43	2.45	-0.02	-0.6%
Cosmopolitan	2.46	2.41	0.05	2.2%

TABLE 8.2 Original vs. Fused (Unconstrained Statistical Matching) Average Audiences for Top 10 Magazines

Magazine Title	Fused AA %	Original AA%	Difference	Relative Difference%
Reader's Digest (Selecciones)	9.07	8.44	0.63	7.5%
Eres	8.86	7.72	1.14	14.7%
Muy Interesante	7.82	7.58	0.25	3.3%
TV y Novelas	7.29	7.55	-0.26	-3.4%
Tele Guía	5.37	5.09	0.28	5.6%
Vanidades	6.03	4.97	1.06	21.3%
TV Notas	4.50	3.84	0.66	17.1%
Conozca Más	2.72	2.80	-0.08	-2.7%
Proceso	3.09	2.45	0.65	26.4%
Cosmopolitan	2.66	2.41	0.24	10.1%

For constrained statistical matching, the deviations are negligible in size since they amount to something like two parts out a thousand (1.066% versus 1.064%).

For unconstrained statistical matching, the discrepancies cannot be easily ignored. The average audience went from 1.066% to 1.180%, which was an 11% increase in audience size. This will have the effect of over-valuating magazines in the multi-media optimization setting. Although we will not provide the details in this paper, we also found that there were significant discrepancies in audience profiles and duplications.

After we saw these results, we thought about possible remedies. It would be easy to go into the fused database and alter the magazine reading data of some respondents such that the average audience levels will come close to the original levels. But we considered this to be a cosmetic fix, since the problems with the audience profiles and duplications remain and may even be exacerbated. The constrained optimization problem of making the fewest number of data changes to match the average audiences, audience profiles and duplications simultaneously would in fact be much more difficult than the transportation problem itself.

To develop a solution requires an understanding of the source of the problem. One possible cause lies in the way that case weights are being handled. In unconstrained statistical matching, the case weights do not figure at all. Instead since the pre-stratification variables (namely, geography, socio-economic level and pay-TV status) as well as the post-stratification variables (namely, geography, age, sex, socio-economic level and pay-TV status) are being used as matching variables with high degrees of matching success, the case weights are assumed to be automatically accounted for.

However, there is another design feature that was not accounted for. In the TGI study, when there is more than one eligible person in a household, one of them is randomly selected (see Kish (1949)). During the weighting stage, a design weight is used to account for the differential probabilities of selection. Thus, if the respondent is the only eligible person in the household, the design weight is 1; if the respondent is one of two eligible persons, the design weight is 2; and so on.

Consider the example of two TGI respondents who have identical matching variables (age, sex, etc), but one of them has a design weight of 1 and the other has a design weight of 2. If the unconstrained statistical matching does not consider case weights, then each person would be chosen half the time. If weights were considered, the first person would be chosen one-third of the time, and the other person two-thirds of the time. A bias will exist if these two people have different product usage and magazine reading behavior.

Although we have identified this as a potential problem, we have not tried a version of unconstrained statistical matching that accounts for case weights, such as using the number of eligible persons in the household as a matching variable. So it is possible that there are other causes as well.

9. INDIVIDUAL-LEVEL ERROR ANALYSES

Data fusion can be considered as a classification problem, in which people are classified into different groups on the basis of known variables. The accuracy of data fusion can therefore be assessed by the standard techniques used in classification and prediction theory (e.g. Han and Kamber (2001), Chapter 7).

In the split-sample method, a database is randomly partitioned into two mutually exclusive sets. One set serves as the donor sample and the other set serves as the recipient sample. After the data fusion is done, we compare the original versus the fused values for each recipient. For dichotomous data (such as the recent reading of a magazine), the outcomes are summarized in the 2x2 contingency table known as the confusion matrix.

TABLE 9.1 Definition of Confusion Matrix

	Original Variable: Yes	Original Variable: No
Fused Variable: Yes	True positive	False positive
Fused Variable: No	False negative	True negative

From this confusion matrix, a number of statistics can be defined, of which these are commonly used:

$$\text{Accuracy} = 100 \times \frac{(\text{Number of true positives}) + (\text{Number of true negatives})}{(\text{Total number of persons}) \times (\text{Total number of variables})}$$

$$\text{Sensitivity} = 100 \times \frac{(\text{Number of true positives})}{(\text{Number of true positives}) + (\text{Number of false negatives})}$$

$$\text{Specificity} = 100 \times \frac{(\text{Number of true negatives})}{(\text{Number of true negatives}) + (\text{Number of false positives})}$$

$$\text{Precision} = 100 \times \frac{(\text{Number of true positives})}{(\text{Number of true positives}) + (\text{Number of false positives})}$$

Accuracy is the percent of correctly classified cases, but its weakness is that it lumps the true positives and true negatives together when we are more interested in the true positives. Sensitivity addresses the question: "Of the people who were really magazine readers (or product users), what percent of them were classified as such?" Specificity addresses the question: "Of the people who were really not magazine readers (or product users), what percent of them were wrongly classified as readers (or users)?" Precision addresses the question: "Of the people who were classified as magazine readers (or product users), what percent of them were really that?" Sensitivity and specificity are regularly used in the evaluation of clinical trials, while precision is used for quality control in industrial production.

To understand the expected behavior of the performance statistics, let us set up a baseline. Suppose the incidence of a variable is p (=fraction of the cases with the attribute). Under a random fusion algorithm, a fraction p of the cases is randomly selected and assigned this attribute and the remaining people are assumed not to have this attribute. The expected fractions can therefore be found in this table (where $q = 1 - p$).

	Original Variable: Yes	Original Variable: No
Fused Variable: Yes	p^2	pq
Fused Variable: No	pq	q^2

The expected values of the performance statistics are as follows:

$$\text{Accuracy} = 100 \times \frac{(\text{Number of true positives}) + (\text{Number of true negatives})}{(\text{Total number of persons}) \times (\text{Total number of variables})} = 100\,(p^2 + q^2)$$

$$\text{Sensitivity} = 100 \times \frac{(\text{Number of true positives})}{(\text{Number of true positives}) + (\text{Number of false negatives})} = \frac{100\,p^2}{p^2 + pq} = 100\,p$$

$$\text{Specificity} = 100 \times \frac{(\text{Number of true negatives})}{(\text{Number of true negatives}) + (\text{Number of false positives})} = \frac{100\,q^2}{q^2 + pq} = 100\,q$$

$$\text{Precision} = 100 \times \frac{(\text{Number of true positives})}{(\text{Number of true positives}) + (\text{Number of false positives})} = \frac{100\,p^2}{p^2 + pq} = 100\,p$$

The insight that can be drawn from these baseline figures is that these four performance statistics depend on the incidence. Other things being equal, a variable with a higher incidence should have higher sensitivity and precision, and lower specificity. Other things being equal, a variable with more extreme incidence (that is, close to either 0% or 100%) will have a higher accuracy.

In Tables 9.2 and 9.3, we show these performance statistics for the split-sample method as applied to the TGI data under both constrained and unconstrained statistical matching. We have added a column titled "baseline" that gives the expected values of these performance statistics under random fusion.

TABLE 9.2 Mean Individual-level Performance Statistics for Product Usage

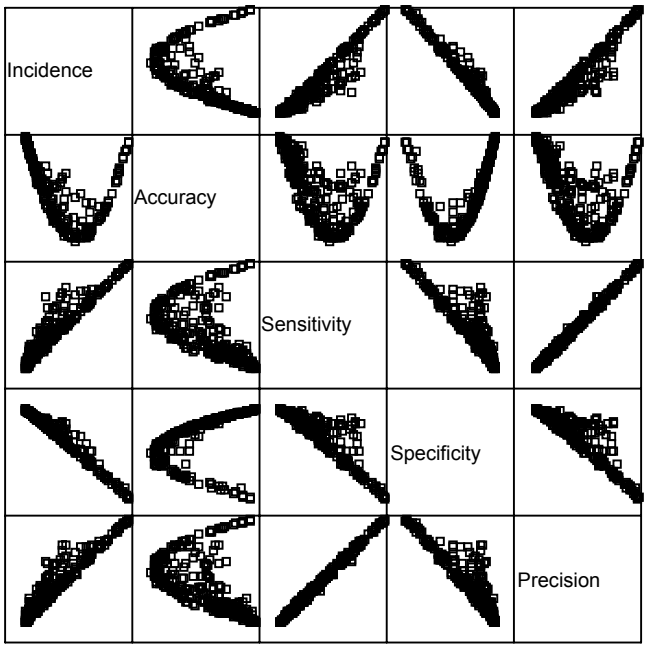
Performance Statistics	Baseline	Unconstrained Statistical Matching	Constrained Statistical Matching
Accuracy	59.2%	75.2%	75.1%
Sensitivity	28.4%	58.2%	56.8%
Specificity	71.6%	82.0%	82.6%
Precision	28.4%	56.2%	57.0%

TABLE 9.3 Mean Individual-level Performance Statistics for Magazine Audiences

Performance Statistics	Baseline	Unconstrained Statistical Matching	Constrained Statistical Matching
Accuracy	99.0%	97.9%	98.0%
Sensitivity	1.1%	6.5%	5.9%
Specificity	99.0%	98.9%	99.0%
Precision	1.1%	5.4%	5.9%

With respect to these performance statistics, there is not much difference between constrained and unconstrained statistical matching. More visible are the differences between product usage and magazine audience, which we know is the consequence of the difference in incidences. This dependency is illustrated in Graph 9.1, in which we have plotted scatterplots between the incidences and the performance statistics for the 270 product variables.

GRAPH 9.1 Scatterplot Matrix of Incidence and Performance Statistics for TGI Product Usage
Split-sample Analysis of Constrained Statistical Matching



The first lesson that we draw here is that it is inappropriate to compare individual-level performance statistics across different studies or different data items, because differences occur as a result of different incidence levels. However, for the same set of variables within the same study, we can compare the individual-level performance statistics from different fusion algorithms.

Our second observation is that some of the individual-level performance statistics seemed to be quite low in absolute terms. For example, the sensitivity and precision measures are only between 5% to 6% for magazine audiences. But we would argue that the individual-level performance statistics are in fact not the relevant criteria to accept or reject a data fusion.

The individual-level performance statistics are based upon comparing individual data items for matched individuals. If one person happened to come across a magazine during a business trip somewhere, then the matching person needs to do the same or else an error is declared. If someone switches the television from channel 2 to channel 4 this minute, then the matching person needs to reach for the remote control and do the same at the same moment, or else an error is declared. When put in those terms, any individual-level comparison such as data fusion seemed hopeless because human behavior between total strangers cannot ever be expected to track so closely.

But this is not what we expect our samples to do for us in applications, whether it is data fusion or observational studies based upon matched samples. We do not analyze the behavior of any specific individual in a sample and base our decisions upon what he/she does. Rather, we analyze the aggregate-level behavior of the entire sample, or at least of sufficiently large subgroups. When we compare two groups, we ask questions such as whether or not the groups have the same overall incidences of the behaviors of interest.

Here, an analogy with the science of physics is appropriate. In quantum mechanics, the position and momentum of an individual quantum particle cannot be known exactly as a result of Heisenberg's uncertainly principle and can only be specified in terms of the probability function known as Schroedinger's wave function. Nevertheless, in statistical mechanics, we can derive the ensemble behavior of a system of individual quantum particles, which is the world as we know and function in.

We therefore argue that the more relevant criteria of evaluation should be the aggregate-level analyses, such as those presented in the next section of this paper. As for the individual-level performance statistics, they are not entirely useless. For one thing, they are useful in comparing different statistical matching algorithms on the same databases (as in Appendices A and B).

10. AGGREGATE-LEVEL ERROR ANALYSES

The evaluation of data fusion has less to do with any TAM-only or TGI-only issues. The heart of the matter has to be: Did the TGI information get fused to the right TAM people? And, conversely, did the TAM information get fused to the right TGI people?

If we have a single-source database in which we have demographics, television viewing, magazine readership and product usage data from everyone, this would be easy. We would sub-divide database into a donor portion and a recipient portion, execute the data fusion and compare the fused and actual data for the recipients. The reality is that we do not have such a single-source database; after all, if we had one, there would be no need for data fusion.

Instead, we have a less-than-ideal, next-best situation. Within the TGI survey, there are a number of television-related questions. The TGI respondents were asked about television viewing by daypart and for specific program types. It is true that the paper by Ephron and Peacock (1999) showed that the levels of TGI-like tv data are systematically different from the information captured by TAM-like systems. Nevertheless, the TGI and TAM do exhibit the same patterns (e.g. women watch more soap opera, men watch more sports, etc), and we will use the TGI sample as if it were a single source database containing demographics, television, magazine and product data.

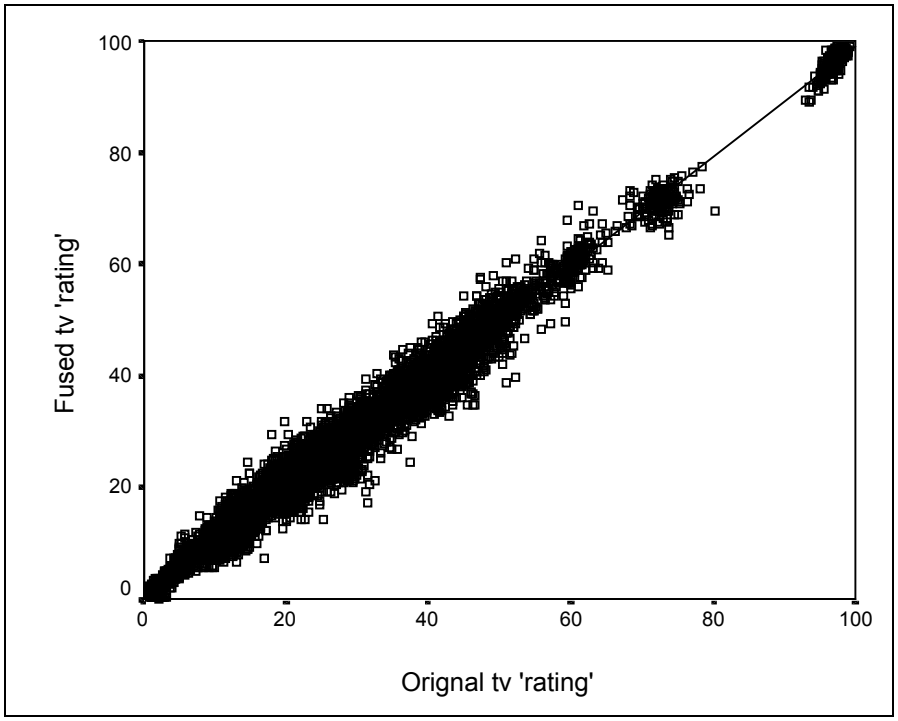
We randomly divided the TGI sample into a donor sample and a recipient sample. We fused a total of 53 television variables (27 dayparts and 26 program types) from the recipient sample onto the donor sample, using both unconstrained and constrained statistical matching.

The split-sample results can be visualized as a table of numbers. In this table, there are 14,310 rows which correspond to the 'ratings' for the combinations of (270 product variables) x (53 television variables). There are two columns of numbers, one for the original values and the other one for the fused values.

For each row (= a particular combination of product and television, such as credit-card holders for the Monday-Friday 8pm-10pm daypart), the first column is the percent of persons in the product group who fit the television usage behavior according to the original data, and the second column is the percent of persons in the product group who fit the television usage behavior according to the fused data.

If data fusion were perfect, the two columns would be identical to each other. In practice, they are somewhat different. The task at hand is to measure these differences. The simplest graphical display is a scatterplot.

Figure 10.1 Original versus Fused TV 'ratings' by Product Users, using Unconstrained Statistical Matching



If data fusion were perfect, all the points would fall right on the 45-degree diagonal line. As it stands, this scatterplot shows considerable dispersion. The R-squared statistic is 0.9597, meaning the 95.97% of the variance in one column is accounted for by the other column. This is a high level of variance accounted for, and does not match the amount of dispersion implied by the visual display.

The reason that the visual display does not seem to match the high value of the R-squared statistic is that the two-dimensional scatterplot does not provide any sense of the density of the points. There are a lot of numbers here (14,310, to be exact), and the single pixel that falls near the 45-degree line may in fact represent many different numbers. So we need some more precise measurements beyond the pictorial representation.

A quantitative measure of the closeness of fit is the ratings index, defined as the 100 times the fused TV 'rating' divided by the original TV 'rating' within a product user group. Here is an example: among persons who have made home improvements in the past 12 months, 43.7% watch national news regularly according to the original data and 39.9% do so according to the fused data. The rating index is 100 x 39.9 / 43.7 = 91.

We compute the rating index for each of the 14,310 entities. Since these are not statistically independent entities, none of the conventional tools of hypothesis testing can be applied. Rather, we can only provide descriptive statistics to summarize their distribution. But when we attempted to summarize these rating indices, we observed that they were highly volatile and non-robust by being influenced by a few outlying observations.

The extreme values of the rating index occur in situations where the TV 'ratings' are small. For example, the maximum value of the rating indices from constrained matching occurred for this situation: 10.6% of persons have life insurance; among these people, the original incidence of watching religious programs is 0.6% but the fused incidence is 2.3%, resulting in an index of 405. Of course, in this particular example, it may be that people who purchase their own life insurance may not be religious but the fusion process failed to reflect this subtle point.

As another example, the minimum value of the rating indices from constrained statistical matching occurred for this situation: 10.5% of persons have personal income of 500 pesos (USD 50) or less; among these people, the original incidence of watching cooking programs on television is 1.97% but the fused incidence is 0.15%, resulting in an index of 8. Neither this example nor the preceding one carry any real-life commercial interest.

Due to the presence of these outliers, moment-based measures such as arithmetic mean, standard deviation, variance, skewness and kurtosis will behave poorly with respect to their ability to summarize the data (see Andrews et al (1974). That is to say, their numerical values will be influenced by the outliers and do not represent the majority of the observations. Instead, we use as our summary measures the median as the measure of central tendency and the interquartertile range (the distance between the 25th and 75th percentiles) as the measure of dispersion, because their breakdown points are higher (Hoaglin, Mosteller and Tukey (1983)).

In Table 10.1, we show these two summary statistics for unconstrained and constrained statistical matching.

Table 10.1 Summary Statistics of Rating Indices by Product User Groups

Statistic	Unconstrained Statistical Matching	Constrained Statistical Matching	Original Split Samples
Median rating index	100.9	100.9	100.8
Interquartile range	17.1	13.5	13.0

The median ratings indices are close to the ideal of 100 for both unconstrained and constrained statistical matching. However, it is not sufficient to be correct on the average since this could be the result of large canceling errors of opposite signs. So we need to check the distribution of the rating indices. The interquartile range (the distance between the 25th and 75th percentiles) of the rating indices is slightly higher for unconstrained statistical matching than for constrained matching, which is probably related to the sample size drop during that process.

There are two major reasons for the original and fused TV 'ratings' to be different. The first one is due to the inadequacy of the data fusion, whether it is the weakness of the statistical technique or the lack of explanatory power in the matching variables.

The second reason is that there is some sampling error in the sample splitting process. For example, 50% of the TGI people may be watching television during weekday primetime; after the random splitting, the incidences may become 51% in the donor sample and 49% in the recipient sample just by chance, and this would be reflected in the original versus fused ratings.

In the last column of Table 10.1, we show the median rating index and the interquartile range of the rating indices for the original split samples. The subject of data fusion does not come up at all since they refer to direct tabulations within the two split samples, so this is a pure measure of the sampling error during the random splitting. We observe that the interquartile range there is only slightly smaller than those from the fused data.

In Table 10.2, we present the results for the summary statistics by magazine readers for unconstrained and constrained statistical matching. The median rating indices are both close to the ideal of 100. Overall, the interquartile ranges for the rating indices of magazine audiences are much larger than those for product users . The interquartile range for constrained statistical matching is smaller than unconstrained statistical matching, which is probably related to the sample size drop in the latter method.

Table 10.2 Summary Statistics of Rating Indices by Magazine Readers

Statistic	Unconstrained Statistical Matching	Constrained Statistical Matching	Original Split Samples
Median rating index	99.4	100.0	100.3
Interquartile range	53.9	41.0	40.7

The split-sample analysis should not be regarded as a panacea in the analysis of the errors associated with data fusion. We will review some of the limitations here.

Firstly, the split-sample analysis is applied to the set of surrogate television viewing variables in the TGI database. The analysis in this paper covers 53 of these variables (27 variables and 26 television program types). Whilst these are the most common summary statistics, they are still but a small fraction of the variety of television environments that exist out there. Above all, they are only surrogates.

Secondly, the TGI database was randomly split into two portions. Sampling error occurs in the sense that different random splits would have resulted in different numbers in the split samples. We could have repeated this exercise multiple times using different random splits and use the average of those runs. But this was not feasible at this time due to the computational time that would be required by constrained statistical matching.

Thirdly, the Mexican data fusion is one in which 10,954 TGI respondents is fused with 7,385 TAM respondents. For the split-sample analysis, the TGI respondents were split into a donor sample and a recipient sample. Therefore, we have different levels of sample sizes in the two situations. To be precise, the split-sample analysis is based upon a situation with smaller sample sizes. We know that the ability to achieve successful matching is a function of the sample size, because it is harder to find successful matches on the same list of variables when the sample size is smaller. As a result, the split-sample analysis will tend to overstate the data fusion error.

Fourthly, for the purposes of the data fusion in Mexico, we identified 18 critical segments as combinations of geography, pay TV status and gender. These segments were critical in the sense that, for a person in one segment in one database, the potential matching cases must come from the same segment in the other database.

The two databases do not have the same sample distributions within these critical segments. The reason is primarily one of deliberate design, wherein certain segments were intentionally over-sampled to achieve certain report intab goals. Those design objectives were different for the two databases due to different commercial needs. As a result, the sample size ratios between the two databases varied between 0.34 to 1.72 across the critical segments. The TGI split samples were simply proportionate random divisions of the whole sample and they do not reproduce the disparate sample size ratios that existed in the actual data fusion. The split-sample analysis may therefore yield unrealistic results.

11. CONCLUSIONS

In this paper, we reviewed our detailed analyses to compare two methods of data fusion for the Mexican TAM and TGI databases. In the end, we decided to go with constrained statistical matching. The primary reasons are the preservation of the TGI incidences and the retention of the full TGI sample size. These were extremely important reasons because of the promise that the TGI estimates were accurate and precise ones based upon a sample of that particular size. It would have been very difficult to justify significantly different magazine audiences or product usage incidences, or to rationalize a reduced sample size.

Our conclusions are not prescriptive. We do not claim that constrained statistical matching is the only right way to conduct data fusion. In another time and another situation, we can see that we or someone else may arrive at different decisions. This is not a half-hearted qualification, because we have developed concrete proposals in which unconstrained statistical matching was unequivocally more logical and appropriate. One purpose of our paper here is to describe the suite of analytical tools that are available to evaluate data fusion, in order to reach an informed decision under the particular circumstances. We have also been careful to point out imperfections and limitations that we see in these tools.

In this paper, we presented a comparative analysis in which we looked at two different methods of data fusion applied to the same databases, with all other parameters being held constant (such as the critical segments, the distance metric and the split samples). This provided us with a great deal of understanding about which analytical tools are useful in helping us to make choices among methods.

The reader may be disappointed in not finding a set of rules that will definitively inform as to whether a particular data fusion was acceptable or not. Our perspective is that we must compare the data fusion against the alternative of not doing it. The needs for both target group ratings and multimedia planning are ubiquitous, and data fusion provides a solution in the absence of a workable single source database. If we reject data fusion, we are in fact asking the users to revert back to using demographic surrogates, which is less than optimal, much more inaccurate and error-prone than data fusion. At a minimum, therefore, data fusion can be positioned as a marked improvement over current practices.

The reader may also be disappointed at seeing how data fusion performs on some of the key evaluation measures. Furthermore, it would seem that there are intrinsic barriers which cannot be surpassed by using more sophisticated statistical techniques. We do not believe that such pessimism is justified. For this paper, we did not have space to include the set of simulations that we have performed. Using TGI split-sample analyses, we simulated what might happen if we were able to include certain key items as matching variables. For example, having just a few magazine questions raised the individual-level performance statistics on magazine audiences by huge amounts. Unfortunately, those key items do not exist on the TAM database as yet, so nothing can be done in the short run.

The media studies were designed and conducted for their own original purposes, with data fusion being an afterthought as a way to add value. In the long run, data fusion can be improved immensely by having the right matching variables. Therefore, getting the cooperation of the research suppliers to make this happen is of the utmost importance.

REFERENCES

David F. Andrews, Peter J. Bickel, Frank R. Hampel, Peter J. Huber, William H. Rogers and John W. Tukey (1972) *Robust Estimates of Location*. Princeton, NJ: Princeton University Press.

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

Richard S. Barr and J. Scott Turner (1978) A new linear programming approach to microdata file merging. *1978 Compendium of Tax Research*. Washington, DC: Office of the Treasury.

Abraham Charnes and William W. Cooper (1954) The stepping stone method of explaining linear programming calculations in transportation problems. *Management Science*, 1, 49-69.

George B. Dantzig (1963) *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.

Belur V. Dasarathy (ed.) (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alamitos, California: IEEE Computer Society.

Richard O. Duda, Peter E. Hart and David G. Stork (2001) *Pattern Classification*. New York: John Wiley & Sons.

Erwin Ephron and James Peacock (1999) MRI and media-mix. *ARF Television Research Workshop*, October 28, 1999. New York: Advertising Research Foundation.

Michael R. Garey and David S. Johnson (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*. San Francisco, CA: W.H. Freeman and Company.

David E. Goldberg (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Jiawei Han and Micheline Kamber (2001) *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Frederick S. Hillier and Gerald G. Lieberman (2000) *Introduction to Operations Research*. New York: McGraw Hill.

Frank L. Hitchcock (1941) The distribution of a product from several sources to numerous localities. *Journal of Mathematical Physics*, 20, 224-230.

David C. Hoaglin, Frederick Mosteller and John W. Tukey (1983) *Understanding Robust and Exploratory Data Analysis*. New York, NY: John Wiley & Sons.

Leslie Kish (1949) A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.

Melanie Mitchell. (1996) *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Christos H. Papadimitriou and Kenneth Steiglitz (1998) *Combinatorial Optimization: Algorithms and Complexity*. Mineola, NY: Dover Publications Inc.

James L. Reilly (2000) *The Development and Evaluation of Statistical Matching Applications*. Master of Science in Statistics thesis, University of Auckland, New Zealand.

Willard L. Rodgers (1984) An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2(1), 91-102.

George A. Vignaux and Zbigniew Michalewicz (1989) Genetic algorithms for the transportation problem. In *Methodologies for Intelligent Systems (*Ras, Z. (ed.), 4, 252-259. Holland: North-Holland.

Note: This is a short list of references that we cited in this paper. We maintain an extensive bibliography in a Microsoft Word document located at the World Wide Web URL (http://www.zonalatina.com/datafusion.doc). By last count, there are over 800 citations in that document.

APPENDIX A. OPTIMAL IMPORTANCE WEIGHTING BY GENETIC ALGORITHM

A statistical matching algorithm involves matching persons by a list of variables. In some instances, it may occur that there is no perfectly matching person available, necessitating the selection of persons who match on only some of the variables. Some variables are deemed to be more important that others; for example, we may consider housewife status to be more important than the presence of telephone if we are dealing with household product purchases because of the direct relevance.

Statistical matching algorithms will therefore assign certain importance weights such that those variables with higher importance weights receive higher priority in the event that a perfectly matching donor could not be found. The objective here is to find a set of importance weights that maximizes the accuracy of statistical matching.

This optimization problem is a difficult one. A brute-force direct search is out of the question because the possibilities are too many. Unlike techniques such as discriminant analysis, there is no closed-form optimal solution here because statistical matching is based upon the local search for the nearest neighbor(s).

We chose to attack this problem with a genetic algorithm method. Roughly speaking, the genetic algorithm is an intelligent search method that is based upon the principles of genetics and natural selection (see Goldberg (1989) and Mitchell (1996) for further information on the subject). Of all things, genetic algorithms have been applied to solve the transportation problem (Vignaux and Michalewicz (1989)).

Description of the Genetic Algorithm

We begin with some definitions of some basic terminology. We define a 'bit' as an entity that is either on (usually assigned a value of 1 or TRUE) or off (usually assigned a value of 0 or FALSE).

For each matching variable, we will assign a bit string of length of four denoted by "XXXX". Each X can assume a value of either 1 or 0 (in genetics, this would be an indicator for the presence or absence of a genetic trait at the locus). The first X is assigned a value of 8, the second X assigned a value of 4, the third X assigned a value of 2 and the fourth X is assigned a value of 1. This bit string is a representation of the importance weight when we summed up the values of all the X's that are 1's.

Example 1: The first matching variable has a bit string of '1111'. Then its scoring weight is 8+4+2+1 = 15.

Example 2: The second matching variable has a bit string of '1001'. Then its scoring weight is 8+1=9.

Example 3: The third matching variable has a bit string of '0000'. Then its scoring weight is 0.

Altogether, there are 2 x 2 x 2 x 2 = 16 different bit strings of length four. Furthermore, each bit string corresponds uniquely to a scoring weight of 0, 1, 2, … 15.

Suppose that we had used only three matching variables in our statistical matching algorithm and they have scoring weights given by the bit strings in Examples 1, 2 and 3 respectively. We can represent this assignment of scoring weights as a bit string of length 12 composed of the three bit strings one after another: 1111 1001 0000.

Our actual genetic algorithm works as follows:

Step 1: Initialization

Within each of the 18 segments, we have 7 matching variables. Since we are assigning a bit string of length 4 to each matching variables, the bit string for a candidate solution will be a bit string of length 7 x 4 = 28.

In the beginning, we generate randomly a population of 100 candidate solutions each defined by a bit string of length 28. Each bit has a probability of ½ to be either 1 or 0.

Step 2: Evaluation

For each candidate solution, we run through the split-sample method for the unconstrained statistical matching. The TGI sample was randomly divided into two halves: one is a recipient sample and the other is the donor sample.

For each person in the recipient sample, we calculate the distance of this person from every person in the donor sample. This distance is based upon the sum total of the importance weights of the variables that disagree, with partial credits in adjacent age groups and socio-economic level. The best match is the person with the shortest distance, which is zero if all variables match. If there is more than one donor with the shortest distance, we choose the one who has been a donor the fewest times so far. If there is still more than one person, we choose one at random.

For each person in the recipient sample, we can now compare his/her actual answers with the donated data on the 270 fused product variables. The goodness of fit measure for this candidate solution is the individual-level accuracy statistic defined in Section 9 of this paper. We record what the best solution is so far.

Step 3(a) Evolution: Crossover

We have 100 candidate solutions and we have the accuracy statistic for each solution. We assign an inverse rank to these candidate solutions (1 to the worse, 2 to the second worst, … , 100 to the best).

We construct the next generation by mating pairs of current candidate solutions. We would select a pair of current candidate solutions with probabilities proportional to their inverse rank. This means that the best current candidate solution is 100 times more likely to be selected than the worst current candidate solution. This pair is called a set of parents. Then we perform the crossover operation as follows.

Imagine that the two 'parents' have their bit strings of length 28 represented symbolically as:

Parent 1: **XXXX XXXX XXXX XXXX XXXX XXXX XXXX**
Parent 2: YYYY YYYY YYYY YYYY YYYY YYYY YYYY

We generate a random integer number between 1 and 28, and call that K. The first descendant of these two parents is constructed as the first K bits of the first parent plus the last (28 - K) bits of the second parent. The second descendant of these two parents is constructed as the first K bits of the second parent plus the last (28 - K) bits of the first parent. For example, if K=8, then we will have

Descendant 1**:** **XXXX XXXX** YYYY YYYY YYYY YYYY YYYY
Descendant 2: YYYY YYYY **XXXX XXXX XXXX XXXX XXXX**

Abstractly speaking, this process will breed a new generation which incorporates different components with the advantage going to the fitter elements of the population. If we take the best part of one good solution and add it to the best part of another good solution, we may obtain a solution better than either.

We repeat this crossover fifty times to obtain 100 new candidate solutions.

Step 3(b) Evolution: Mutation

We introduce the random element of mutation by systematically going through all the individual bits in every bit string of the 100 new candidate solutions and flipping them (that is, 1 becomes 0 while 0 becomes 1) with a low probability (1 out of 100).

Mutation is important because its role in creating surprising, possibly superior, candidate solutions that go beyond the stock in previous generations.

We repeat steps (2) and (3) for 100 times. By this time we will have evaluated 100 x 100 = 10,000 candidate solutions in an intelligently directed search. It is observed that the accuracy increases slowly in the beginning and then there is no more improvement after a while. We save the best solution found across all the generations.

In the context of this project, we apply the genetic algorithm separately to each of the segments. At the end, we had evaluated 18 x 10,000 = 180,000 candidate solutions. Since the segments were independently processed and

the solutions can be combined across segments, this is in fact equivalent to having evaluated $(10{,}000)^{18} = 10^{54}$ different candidate solutions.

In the initial generation, we always seed one of the candidate solutions to consist entirely of 1's in the bit string. This special solution represents one in which the matching variables are given equal importance to each other.

Results

Given that the stated goal was to maximize the individual-level accuracy, here are the most important results:

- 74.9% accuracy was achieved by assigning equal importance weights to the 7 matching variables
- 75.2% accuracy was achieved by the unconstrained statistical matching using the set of importance weights that was ultimately used, as reported in Section 4 of this paper.
- 75.3% accuracy was achieved by the best genetic algorithm solution

Discussion

The best genetic algorithm solution was just slightly better than the simple one that we defined, but it is significantly more complicated because the importance weights are different by segment. Some of these differences are easily interpreted. For example, in the best solution, socio-economic level is assigned smaller importance weights in the Mexico City pay-TV segments because the US$30 pay TV bill pre-supposes a certain amount of disposable income; but then assigning a large importance weight to socio-economic level would have no impact anyway because most of those people would be from the same upper level. In the end, following the principle of Ockham's razor, we opted for the simple solution because it was more transparent, much easier to implement and essentially just as accurate.

The results above are for unconstrained statistical matching only. The total computer time taken to process 180,000 evaluations of the split-sample method on the TGI database was about 20 hours. Since it takes less than 10 seconds to run the unconstrained statistical matching once versus almost 5 hours for constrained statistical matching, it was not feasible to repeat the above exercise for constrained statistical matching. Nevertheless, based upon the results in Section 9 of this paper, there should not be any marked difference between constrained and unconstrained statistical matching in terms of the accuracy statistic.

The assertion that the specific choice of importance weights does not make a difference should not be generalized beyond the present configuration of sample sizes and matching variables. In this particular case, we were able to achieve high matching success rates for the matching variables (see Section 7 of this paper). Using any distance metric based upon reasonable importance weights, for example, someone who matches on 9 variables should be a better match than someone else who matches only on 3 variables.

This lack of sensitivity of the results to the choice of parameters is sometimes known as the 'flat maximum effect'; it has also been called the 'curse of insensitivity' (see Duda, Hart and Stork (2001)). The presence of this phenomenon in statistical matching has been noted by Rodgers (1984), who wrote in his extensive review of the subject that there is no "evidence that would suggest any alternative that would be consistently superior to a simple subjectively weighted sum of the absolute differences between values on the X variables." But we can imagine that if we had 50 matching variables or if our sample sizes were smaller, there may be only a small number of matched variables found for any donor, and then the optimal assignment of importance weights may matter more.

APPENDIX B. STEPWISE STATISTICAL MATCHING

Statistical matching involves the matching of persons on a list of variables. We can obviously use only those variables that are in common between the two databases and we should use only those variables that have a direct bearing on the accuracy of the fusion results.

In this section, we will present the results of an exercise designed to gauge the impact of introducing variables into the matching in a stepwise fashion. These results are of interest because they show the trade-offs as we bring in more variables.

The first aspect that we will treat is the matching success rate. In Table B1, we show the matching success rate by seven matching variables (listed in the columns). The rows of the table represent the steps by which variables were introduced, in accordance with the system of importance weights described in Section 4 of this paper. Thus, step 0 is the baseline of random fusion in which people were randomly matched together in unconstrained statistical matching. The table entries are the matching success rates. For example, in Step 0, only 28% of the persons were correctly matched on socio-economic level, and so on.

In Step 1, we introduce the 18 segments defined by geography, pay-TV status and gender. At this point, people will be perfectly matched on those 3 variables. In Step 2, we matched socio-economic level within the segments whereupon the matching success rate becomes 100% for socio-economic level but other variables such as age do not improve. In Step 3, we match age as well whereupon we become 99% correct on age.

TABLE B1 % Successfully Matched under Unconstrained Statistical Matching in Stepwise Process

Step Number	% Socio-economic Level	% Age Group	% Head of Household	% Housewife	% Presence of Children	% Multi Set	% Presence of Phone
Step 0: Random fusion	28%	21%	56%	60%	49%	56%	73%
Step 1: + 18 segments	31%	21%	63%	73%	50%	59%	74%
Step 2: + SE Level	100%	21%	64%	72%	52%	61%	79%
Step 3: + Age Group	100%	99%	77%	78%	55%	63%	79%
Step 4: + Head of Household	99%	99%	100%	86%	56%	63%	79%
Step 5: + Housewife	99%	98%	100%	100%	59%	64%	80%
Step 6: + Presence of Children	97%	97%	100%	100%	98%	64%	80%
Step 7: + Multi-set	98%	97%	100%	100%	98%	93%	80%
Step 8: + Presence of Phone	98%	97%	100%	100%	98%	93%	93%

The most obvious conclusion is that the matching success rate approaches 100% whenever that particular variable is brought into the process. The matching success rate also improves when correlated variables are introduced. For example, adding gender in Step 1 helps to identify housewives; and so on. It is also clear that the matching success rates for the first variables begin to decline as more and more variables are added. Therefore, we cannot add more and more variables indefinitely, as there are some negative consequences.

The lesson learned from the preceding table is that we should only match on those variables that matter. In Table B2, we show the individual-level performance statistics obtained from applying the split-sample method to the 270 TGI product variables under unconstrained statistical matching.

TABLE B2 % Individual-Level Performance Statistics for the Split-sample Method for TGI Product Variables under Unconstrained Statistical Matching in Stepwise Process

Step Number: Variable Added	Accuracy	Sensitivity	Specificity	Precision
Step 0: Random fusion	59 %	28 %	72 %	28 %
Step 1: + 18 segments	73 %	57 %	80 %	53 %
Step 2: + Socio-economic Level	75 %	57 %	82 %	55 %
Step 3: + Age Group	75 %	58 %	82 %	56 %
Step 4: + Head of Household	75 %	58 %	82 %	56 %
Step 5: + Housewife	75 %	58 %	82 %	56 %
Step 6: + Presence of Children	75 %	58 %	82 %	56 %
Step 7: + Multi-set	75 %	58 %	82 %	56 %
Step 8: + Presence of Phone	75 %	58 %	82 %	56 %

From Table B2, it is quite apparent that the performance statistics have leveled off after Step 3, beyond which adding new variables no longer helps. We note that these statistics are averaged across the 270 product variables. When we looked at specific variables, we found some instances where the introduction of a new matching variable led to significant improvement. Examples are the presence of children for infant care products and the presence of telephone for long-distance telephone usage. But such local improvements may not show in the global average.

So our final decision on using this set of matching variables was based upon two things: (1) the matching variables were found to be correlated with the fused variables, and therefore their inclusion will improve the overall accuracy of the fusion; and (2) the inclusion of a matching variable does not decrease the matching success rates for the other variables.

The results in Appendix B is for unconstrained statistical matching only. The closeness of the performance statistics between constrained and unconstrained statistical matching in Section 9 would suggest that the conclusions should also apply to constrained statistical matching.

We should also caution the reader that the results should not be taken as the inherent limitations of data fusion as such. Rather, these results reflect the total predictive power of this set of common variables that are available at this particular moment in time. Not reported here are experimental results which show that marked improvement can be achieved by inserting more questions into the two studies to use as matching variables, such as a few simple questions about magazine readership.