FUSION-ON-THE-FLY FOR MULTIMEDIA APPLICATIONS

Roland Soong and Michelle de Montigny, Kantar Media Research

1. ABSTRACT

In a previous paper (Soong and de Montigny (2003a)), we presented a data fusion algorithm called predictive isotonic fusion. This is a rapid fusion-on-the-fly method that is optimized for a target group, while preserving sample sizes and media currencies. This paper examines how this algorithm can be enhanced in the context of analyzing multi-media television-print schedules though the incorporation of print-related information. We showed that this was indeed the case.

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 all the previously separate information. Data fusion products are usually produced on a syndicated basis, whereby the fusion database is produced once and for all and issued to all subscribers.

Syndicated data fusion is usually formulated as a global optimization problem, and is therefore a one-size-fits-all approach. Alternately, there are those who prefer to have data fusions that are customized for specific problems, under the reasonable belief that these might be superior locally optimal solutions. However, such customized solutions cannot be built beforehand since there can be a very large number of different formulations. Therefore, we must wait for the problem to be presented and then that fusion problem must be solved rapidly. Such fast, customized fusions are often referred to as 'fusion on the fly.'

In this paper, we present a 'fusion on the fly' algorithm and we will compare its empirical performance against a syndicated data fusion product in the context of multi-media TV-print planning.

3. DESCRIPTION OF CONSTRAINED STATISTICAL MATCHING

The most prevalent form of syndicated data fusion is the (TAM+TGI)-like fusion. On one side, we have a television audience measurement (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.

There are many ways to conduct (TAM+TGI) fusion. If the objectives are to preserve the TAM and TGI sample sizes and to preserve the media currency values, then there is a well-defined and elegant formulation known as constrained statistical matching (Soong and de Montigny (2001)) which is based upon solving the transportation problem in the field of operations research. Syndicated (TAM+TGI) fusion products based upon constrained statistical matching have been produced in Argentina, Brazil, Colombia, Mexico, Puerto Rico and the United States.

The syndicated fusion products are standardized products, so that all subscribers receive the identical fusion databases. They are constructed through the collaboration of the fusion specialists with the original media research suppliers so that the integrity of the original databases is maintained. Constrained statistical matching will typically take hours to execute, and therefore cannot be used in an interactive environment.

4. DESCRIPTION OF PREDICTIVE ISOTONIC FUSION

There are many variations of 'fusion on the fly' (for example, Czaia (1992), Raimondi and Santini (1997), Baynton (2003)). Our algorithm goes under the name of 'predictive isotonic fusion' and was previously presented in Soong and de Montigny (2003a) in the context of target group television ratings. We will describe this algorithm in some detail here.

As we see it, here are the requirements:

1. The fusion should be optimized for a specific target group which is defined 'on the fly.' The target group definition is potentially complex, such as 'young mothers who have purchased non-prescription vitamins for their children' or 'professionals/managers who have traveled overseas for business at least three times in the last 12 months' and cannot be pre-listed and processed in advance.
2. The fusion should be executed in sufficiently quick time in an interactive environment. That means not more than a few seconds in elapsed time on a typical personal computer.
3. The fusion should preserve the media currencies and target group incidences in the original databases.
4. The fusion should preserve the sample sizes of the original databases.

Requirement #1 is a given fact which we cannot change, so it remains for us to devise a fusion algorithm that runs fast and preserves the sample sizes, target group incidences and media currencies. Just as importantly, we need to establish that this algorithm yields accurate results for our intended application.

We divide our description of the algorithm into two phases. In the first phase, we deploy a predictive model to obtain a predicted score for target group membership for all the cases in both databases. In the second phase, we deploy a quick matching algorithm that preserves the order (i.e. 'isotonic') of those predicted scores. Our method is similar in spirit, but not identical in details to works such as Kahane (1978; reprinted 2001), Rubin (1986), Moriarity and Scheuren (2001) and Moriarity and Scheuren (2002).

Step 1: Predictive Modeling

The practice of predictive modeling consists of the following steps (Weiss and Indurkhya (1998)). There is a database which contains the outcome variable and some predictor variables. We construct a statistical model that relates the outcome variable with the predictor variables. Then we proceed to apply this statistical model onto another database which contains only the predictor variables to obtain the predicted scores for the desired outcome. Predictive modeling is used extensively in database marketing, data mining, direct marketing, credit card solicitation, credit scoring, insurance prospecting, loan approval, magazine subscriber drives, etc.

In the present context, the TGI database contains the target group information and a list of variables that can be used as predictors (namely, demographics, television viewing, magazine readership, etc). We construct statistical models that relate target group membership with these predictor variables. Then we proceed to apply the statistical models to the TAM and TGI databases, such that every person in both databases receives a predicted score for target group membership.

Step 2: Isotonic Matching

The TAM and TGI databases are now sorted by the predicted scores. The two databases are then merged together by a process that preserves the order of these predicted scores. A verbal description of this process may be hard to understand. Instead, we have created an illustrative example in Appendix A. With the accompanying tables there, the verbal explanation should be easy to understand.

For a standard (TAM+TGI) setup, the predictive isotonic fusion will take just a few seconds to execute. Therefore, it satisfies the timing requirement.

5. DESCRIPTION OF DATABASES

For the empirical portion of this paper, the syndicated fusion product is the 2002 NTI-MARS product. On one side, we have the Nielsen Television Index, consisting of 11,657 adults who were intab in the Nielsen People Meter panel for one or more days during the first 13 weeks of 2002. On the other side, we have 22,097 adults who participate in the MARS OTC/DTC Pharmaceutical Study during the first quarter of 2002.

If our goal is to compare the accuracy of the fusions, then the NTI-MARS fusion itself will be uninformative. The syndicated fusion and the predictive isotonic fusion will sometimes match different people together, but there is no way of deciding which one is more 'accurate.'

Rather, the standard approach in assessing the accuracy of fusions is through a split-sample foldover analysis of a single source database. The MARS study contains the following relevant data elements:

- Target group information: We chose forty ailment conditions (from acid reflux to yeast infection) from the MARS study, and the target group variables are defined as presence of those conditions during the past 12 months.
- Demographic variables: There are 21 demographic variables (gender, age, geography, etc) and media variables (average daily television hours, presence of cable/satellite, etc) that are present in both databases.
- Television variables: There are 17 television program types, past-seven-day viewing to 34 cable television networks and average viewing hours in 12 dayparts. These variables are not considered to be equal to people meter data in accuracy or resolution, but they have reasonable similarity in profiles that they can be used as approximate surrogates.
- Print Variables: There are 96 magazines which are measured by the frequency of reading method.

The MARS respondent-level database was randomly divided into two halves. One half-sample served the role of the NTI sample (henceforth referred to as the NTI-half sample), and the other half-sample served the role of the MARS sample (henceforth referred to as the MARS-half-sample).

6. DESCRIPTION OF FUSION METHODS

Four different fusion methods were implemented and compared. Our benchmark is the constrained statistical matching used in the syndicated data fusion product. There were three predictive isotonic fusions that used different groups of predictor variables.

Method A: Constrained Statistical Matching

The syndicated data fusion product is our benchmark. The half-samples were divided into 36 mutually exclusive and exhaustive strata defined by age, gender and overall television viewing hours (heavy/medium/light) and we solved the transportation problem within each stratum by matching on the remaining 18 common variables subject to the preservation of case weights (and therefore sample sizes).

Method B: Demographic Predictive Isotonic Fusion

The simplest version uses the same set of common variables in the syndicated fusion. We will call this the 'demographic' predictive isotonic fusion. For each target group, a multiple linear regression model was run with that target group as the outcome variable. There were 56 predictor variables, which were derived by coding the 21 common demographic variables into indicator variables. The resulting model equation was applied to both the MARS-half-sample and the NTI-half-sample, so that every person received a predicted score. The two half-samples were then matched together by the isotonic matching method (as illustrated in Appendix A).

For the demographic predicted isotonic fusion, there were forty fused databases since there is a different predictive model for each target group.

Method C: Demographic/Television Predictive Isotonic Fusion

In the second part of our predecessor paper (Soong and de Montigny (2003a)), we proposed an extension by introducing more television variables. Generally speaking, we can improve the performance of data fusion by introducing better matching/predictor variables. Since our interest there was in target group television ratings, it would seem that having more television-related matching/predictor variables should help.

For syndicated fusion, the high-dimensional matching problem is already stressful. Adding a large number of TV-related matching variables will bring on the 'curse of dimensionality' (Bellman (1961)) to lower matching success rates.

By contrast, predictive isotonic fusion is not subjected to this limitation in the present case. The framework here initially involved fitting a multiple linear regression of 56 predictor variables for a total sample size of more than 11,000 cases. Adding another few dozen more predictor variables will not stress the system.

So this fusion is called demographic/television predictive isotonic fusion. For each target group, a multiple linear regression model was run with that target group as the outcome variable. There were 119 predictor variables, consisting of 56 demographic predictors and 63 television predictors. The resulting model equation was applied to both the MARS-half-sample and the NTI-half-sample, so that every person received a predicted score. The two half-samples were then matched together by the isotonic matching method (as illustrated in Appendix A). Again, there were forty such databases since the procedure was implemented separately by target group.

Method D: Demographic/Television/Magazine Predictive Isotonic Fusion

The demographic/television predictive isotonic fusion seemed natural for the application of target group television ratings. But our present interest is in the application of mixed media television-print planning, where it seems natural to think that magazine-related information is important too. In fact, it is distinctly unseemly that the word 'magazine' has not been mentioned anywhere in the descriptions of the first three methods when magazines 'hold up half the sky' in this application.

In this fusion, we used the MARS-half-sample to build two different predictive models. In the first instance, we built a multiple linear regression model based upon the 56 demographic predictors and 63 television predictors. The resulting model equation was applied only to the NTI-half-sample, so that every person there received a predicted score. We cannot use any magazine predictors because the true NTI sample does not possess that information.

In the second instance, we built a multiple linear regression based upon the 56 demographic predictors, 63 television predictors and 96 magazine predictors (which are the frequency-of-reading-based probability of reading to the 96 measured magazine titles). The resulting model equation was applied only to the MARS-half-sample, so that every person there received a predicted score. The two half-samples were then matched together by the isotonic matching method. Again, there were forty such databases since the procedure was implemented separately by target group.

It may seem strange that two different predictive models are being applied within the same fusion. This can be re-conceptualized as follows. Within each database, we were simply looking for the maximum amount of relevant information that permits us to rank the cases from the highest to the lowest in terms of predicted propensity of target group membership. Once the sorts are done, the numerical values of the predicted scores do not figure in the isotonic matching which is based solely on the sort order. Therefore, the fact that the scores were derived from different models is immaterial.

In any case, the justification for this method is going to depend on the validation exercises in the next section.

7. SPLIT-SAMPLE ANALYSES

As described in our other paper for this symposium (Soong and de Montigny (2003b)), there are many criteria to assess the accuracy of data fusion. Our stated philosophy is for the data fusion supplier to select the criteria and state the reasons why those criteria are relevant to the applications.

In this case, the application is multimedia television-print planning. Given a single source database, the typical multimedia analysis software system will require the following elements:

- The target group
- Target group magazine ratings
- Target group television ratings
- Target group pairwise duplications

Knowledge of these terms permit the evaluation of the reach/frequency characteristics of mixed media schedules for the target group. Optimization can be done by considering these characteristics over different schedules. If these variables are fused accurately, all is well; if they are not fused accurately, then the ensuing analyses and decisions are subject to error.

Evaluation of Target Group

All four fusions are constrained fusions in that the original samples appear with their original data and survey weights, so the target group incidences are preserved.

Evaluation of Target Group Magazine Ratings

Since all four fusions are constrained fusions, with target group and magazine readership data coming from the same database, the target group magazine ratings are automatically preserved. This consistency is extended to the preservation of pairwise inter-magazine duplications within target groups.

Evaluation of Target Group Television Ratings

Since the target group information comes from one database and the television information comes from the other database, the accuracy of target group television ratings depends on the quality of the fusion process. This is addressed by a split-sample foldover test on the MARS sample.

The split-half foldover tests led to the creation of a number of respondent-level fused databases. There is only one database for the constrained statistical matching but, for each the predictive isotonic fusion, there were 40 such databases since the procedure is customized separately by target group.

Within each respondent-level fusion database, a ‘record’ contains the following information:

- The record weight
- The target group variables, common variables and television surrogate variables from the NTI-half-sample
- The target group variables, common variables and television surrogate variables from the MARS-half-sample

The assessment of the accuracy of the target group television ratings of the various fusions will be based upon comparing the original and fused data within the NTI-half-sample. There are 40 target groups and 63 television measures for 2,520 combinations. For each combination, we calculated the original target group rating for the NTI-half-sample by using the true target group and television viewing information and then we calculated the fused target group rating for the NTI-half sample by using the fused target group information with the true television viewing information.

Explanation of Analytical Statistics

Fusion method: Predictive isotonic fusion using demographics only

Target group: Persons who suffered from depression in the past 12 months
Television measure: Watched Fox News in past 7 days

Original target group rating = 31.6
Fused target group rating = 30.1

Difference = 30.1 – 31.6 = -1.5
Absolute difference = 1.5

%Relative difference = 100 x (30.1-31.6)/31.6 = -4.7%
%Absolute relative difference = 4.7%

The following table shows the summary statistics for the four methods:

Table 1. Summary statistics for original vs. fused target group ratings by four fusion methods

Fusion Method	Mean difference	Mean absolute difference	Mean %relative difference	Mean %absolute relative difference
Constrained statistical Matching	1.78	2.50	5.7%	10.2 %
Predictive Isotonic: Demographics only	0.97	2.71	6.5%	11.0 %
Predictive Isotonic: Demographics/TV	0.48	2..02	2.7%	8.6 %
Predictive Isotonic: Demographics/TV/Print	-0.04	1.90	-1.5%	8.4 %

We can also count the number of times in which one fusion method is closer to the original target group rating than another. Using the constrained statistical matching method as the benchmark against the three predictive isotonic fusions, we find demographics-only is closer 45% of the time, demographics/TV is closer 57% percent of the time and demographics/TV/print is closer 61% of the time.

When the same matching/predictor variables are available, predictive isotonic fusion is comparable to constrained statistical matching. Whereas predictive isotonic fusion is optimized for each target group and therefore has a smaller bias (as reflected by the mean difference statistic), it is a more complex method that incurs larger variances (as reflected by the mean absolute difference statistic).

The predictive isotonic fusion method is able to incorporate more predictor variables than constrained statistical matching. When additional media-related variables are introduced, the performance is improved. For a purely television application here, even the magazine reading variables improved the performance. The performance statistics are obtained from a split-sample foldover analysis, so they are realistic estimates.

Evaluation of Target Group Pairwise Duplications

Since the target group information and magazine readership information come from one database and the television information comes from the other database, the accuracy of target group pairwise television-magazine duplications also depends on the quality of the fusion process. So this is addressed by a split-sample foldover test on the MARS sample.

There are 40 target groups, 63 television variables and 96 magazine titles, which yields a total of 40 x 63 x 96 = 241,920 pairs of television-magazine duplications. In practice, not all of these pairs would be utilized. There are pairs of media vehicles (e.g. *MTV* and *Modern Maturity, ESPN* and *Ladies Home Journal*, etc) which would probably never occur in real schedules. Therefore, for each target group, we considered only those media vehicles that have indices of greater than 105. This reduces the number of target group television-print duplications to 76,093 combinations. The summary statistics are shown in Table 2.

Table 2. Summary statistics for target group television-print duplications by four fusion methods

Fusion Method	Mean difference	Mean absolute difference
Constrained statistical matching	0.40	0.57
Predictive Isotonic: Demographics only	0.25	0.65
Predictive Isotonic: Demographics/TV	0.17	0.56
Predictive Isotonic: Demographics/TV/Print	0.11	0.53

The outcomes are consistent with the results found for the target group television ratings. When the same matching/predictor variables are available, predictive isotonic fusion is comparable to constrained statistical matching. Whereas predictive isotonic fusion is optimized for each target group and therefore has a smaller bias (as reflected by the mean difference statistic), it is a more complex method that incurs larger variances (as reflected by the mean absolute difference statistic).

The predictive isotonic fusion method is able to incorporate more predictor variables than constrained statistical matching. When additional media-related variables are introduced, the performance is improved. There is a measure of justice to see that incorporating the magazine readership yields the best performance of all.

8. DISCUSSION

In this paper, we described the open-source algorithm that is based upon the well-understood practice of predictive modeling followed by a quick sort-and-match. We showed that this fast algorithm does not suffer any loss in accuracy compared to the more elaborate method of constrained statistical matching.

The more valuable observation is that predictive isotonic fusion has the ability to accommodate many more predictor variables. We showed that the incorporation of more media-related predictor variables, especially the magazine variables, can improve performances. This is true for the applications for target group television ratings as well as target group multimedia television-print planning.

However, we do not see that the world will be moving into 'fusion-on-the-fly' applications immediately. The convenience of syndicated data fusion is a single database that is issued to all users, which guarantees a consistency. Customized fusion-on-the-fly applications generate multiple fusion databases, and such proliferation may cause confusion in environments where large numbers of people are working on many projects at the same time.

BIBLIOGRAPHY

Baynton, Paul (2003) Data integration or fusion? *ARF/ESOMAR Week of Audience Measurement (Mixed Media Session)*, Los Angeles (USA), 169-181.

Bellman, R.E. (1961) *Adaptive Control Processes*. Princeton University Press: Princeton, NJ.

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

DeGroot, M.H., Feder, P.I., and Goel, P.K. (1971) Matchmaking. *Annals of Mathematical Statistics*, 42, 578-593.

Goel, P.K. and Ramalingam, T. (1989) *The matching methodology: some statistical properties*. Lecture Notes in Statistics, Volume 52. Springer-Verlag New York: New York, NY.

Kadane, J.B. (1978) Some statistical problems in merging data files. In *1978 Compendium of Tax Research, Office of Tax Analysis, Department of the Treasury*, 159-171. U.S. Government Printing Office: Washington, DC. Reprinted in *Journal of Official Statistics* (2001), 17, 423-431.

Kadane, J.B. (2001) Some statistical problems in merging data files. *Journal of Official Statistics*, 17, 423-431.

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

Soong, R. (2002) Quick vs. optimal algorithms in data fusion. *Zona Latina*, January 2002. (http://www.zonalatina.com/Zldata215.htm)

Soong, R. and de Montigny, M. (2001) An anatomy of data fusion. *Tenth Worldwide Readership 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 In 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) Foundations of split-sample foldover tests. Proceedings of the *Eleventh Worldwide Readership Symposium*, Boston (USA).

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

Appendix A. Illustrated Example of Isotonic Matching

After the predictive modeling step, all the cases in the TAM-half-sample and the TGI-half-sample have received predictive scores. The goal now is to create a matching. According to DeGroot, Feder and Goel (1971), the maximum likelihood pairing is to sort the predicted scores and then to link the corresponding pairs (that is, the largest values together, the second largest values together, and so on). Since this pairing preserves the order of the predicted scores, Goel and Ramalingam (1989, Section 3.1.1, p.76-78) named it 'isotonic matching.'

When survey weights are present on databases of unequal sample sizes, the method must be adapted. We will illustrate with a small example. In Table A1, the two databases are each sorted in order of these predicted scores. It is noted that one database contains 4 cases and the other database contains 5 cases, and they both sum to the same projected weight of 2,000. The fusion database is shown in Table A2.

Table A1.
Example of TAM-sample and TGI-sample with predicted scores.

TAM-sample ID	Weight	Predicted Score
TAM-1	600	0.75
TAM-2	400	0.50
TAM-3	300	0.25
TAM-4	700	0.10
Total	2000	

TGI-sample ID	Weight	Predicted Score
TGI-1	300	0.80
TGI-2	400	0.60
TGI-3	200	0.30
TGI-4	500	0.20
TGI-5	600	0.05
Total	2000	

Table A2.
Fusion Database

Fused ID	TAM-sample ID	TGI-sample ID	Weight	TAM Predicted Score	TGI Predicted Score
Fused-1	TAM-1	TGI-1	300	0.75	0.80
Fused-2	TAM-1	TGI-2	300	0.75	0.60
Fused-3	TAM-2	TGI-2	100	0.50	0.60
Fused-4	TAM-2	TGI-3	200	0.50	0.30
Fused-5	TAM-2	TGI-4	100	0.50	0.20
Fused-6	TAM-3	TGI-4	300	0.25	0.20
Fused-7	TAM-4	TGI-4	100	0.10	0.20
Fused-8	TAM-4	TGI-5	600	0.10	0.05
Total			2000		

Isotonic matching works by marching down the two half-samples from the top to the bottom, one record at a time. At first, we look at the first records (TAM-1 and TGI-1). We write into the fusion database a record corresponding to TAM-1 and TGI-1 and the smaller of the two weights. Then we subtract these weights from the two original databases. Thus, TAM-1 is still present in the TAM-half-sample, but with a reduced weight of 600-300 = 300 whereas TGI-1 is completely removed from the TGI-half-sample.

We repeat the process on the revised half-samples. So the next fusion record to be written out is (TAM-1 and TGI-2), after which TAM-1 is completely accounted for and TGI-2 is reduced to 400-300 = 100. This process is continued and will eventually terminate with everyone accounted for.

Isotonic matching is equivalent to the northwest-corner rule that is sometimes used to jumpstart the stepping stone algorithm for the transportation problem (Soong (2002)). Therefore, it has the unimodularity property of creating a fusion database in which the total number of records is no more than the sum of the two input databases minus one. The computational complexity is linear in the sample sizes, and the execution is therefore instantaneous on (TAM+TGI)-like databases.

In the fusion database, the sum of record weights is the same total as in the two original databases. Furthermore, since each original TAM and TGI person is present in the fusion database --- sometimes in more than one record --- with the same relative weight, this method satisfies the requirements to preserve sample sizes, media currencies and product usage incidences. This method is therefore a constrained data fusion.