NO FREE LUNCH IN DATA FUSION / INTEGRATION

Roland Soong
Michelle de Montigny

This paper addresses the elusive quest for that one single best method for data integration. We assert that this is a fool's quest since at the heart of learning theory is the famous *No Free Lunch Theorem* which makes this an impossible mission. We show the results from four different projects, each one being a genuine real-life commercial problem, in which we applied a number of standard data integration methods. None of these methods is the best in all applications. For any specific problem, the best approach is to find the method that is crafted according to the exact circumstances.

INTRODUCTION

Given the proliferation of information and the near-impossibility of obtaining good single-source data, methods of data integration have assumed increased importance. A number of different methods of data integration have been proposed and even commercially realized. Clearly, there is a desire to form opinions about the accuracy of these methods in real-life applications.

It would be nice and clean if there is one single method that can be shown to be superior (or equal) to other methods. Unfortunately, at the heart of classification/learning theory is the famous *No Free Lunch Theorem* (see Duda, Hard and Stork 2001),[1)] which denies this intellectually lazy option (to wit, the free lunch).

The exact derivation of the *No Free Lunch Theorem* involves deep mathematics, so we will omit that. We will provide some common sense descriptions of its implications.

The *No Free Lunch Theorem* states that, when considered over the totality of all possible problems, no one algorithm can be better on the average than any other algorithm. Thus, no matter how clever we are in devising a theoretically sound and sophisticated "good" learning algorithm, there will still be problems in which a theoretically unsound and dumb "bad" algorithm will outperform the "good" one. So it is a fool's quest to find that one globally "best" algorithm, and the more pragmatic and achievable goal is to find the 'good' algorithm for a specific problem.

All statements of the form "algorithm 1 is better than algorithm 2" are ultimately statements about a specific problem and cannot be generalized to other problems. Thus, a theoretically sound and sophisticated algorithm can sometimes perform poorly when the algorithm and the problem are ill-matched. Duda, Hart and Stork (2001) advised:

> *"Practitioners must be aware of this possibility, which arises in real-world applications. Expertise limited to a small range of methods, even powerful ones such as neural networks, will not suffice for all classification problems. Experience with a broad range of techniques is a best insurance for solving arbitrary new classifications problems."*

This paper will show some empirical evidence for the *No Free Lunch Theorem*. The authors studied four different problems of data integration. These are real-life media research problems that are encountered in commercial situations. The projects are:

- Project A: Data fusion of a television people meter panel (TAM) with a multimedia product usage study (TGI) to produce target group television ratings.
- Project B: The fusion of three local market databases – a television people meter panel, a radio diary survey and a multimedia product usage study (TGI) – to produce target group mixed media schedules.
- Project C: The scoring of a small local market database for the propensity of product purchase as obtained in a large national database.
- Project D: The superposition of an attitudinal segmentation scheme from a small custom study onto a syndicated multi-media product usage database to produce target group magazine ratings.

For each problem the authors applied five or six of the most common data integration methods, including random duplication, simulation method, unconstrained and constrained statistical matching, predictive isotonic fusion, logistic regression, linear regression and discriminant analysis. This paper covers a total of 11,094 individual data integration projects.

For each project, there are many performance statistics that can be reported in principle. Following the line of reasoning delineated in Soong and de Montigny (2003c), the authors will report on those key measures that bear directly on the validity and accuracy of the data integration for the intended applications and not on anything else.

PROJECT A: TAM-TGI FUSION

Background

This is the classical situation of data fusion. In most major markets around the world, television is the dominant medium and garners the largest share of advertising expenditure. Most often, the television audience currency is measured through a people meter panel (usually known as Television Audience Measurement (TAM)), which has detailed viewing information but usually little else by way of product consumption.

In those major markets with people meter panel there are usually other studies, such as the Target Group Index (TGI), which collect detailed information on product usage and consumption of media. Such studies are used for multimedia planning. It is natural to want to integrate the TAM and TGI databases in order to run television plans based upon target groups defined in terms of product consumption. This application is known as Target Group Ratings (TGRs).

The best known example of a TAM-TGI fusion is given in Baker, Harris and O'Brien (1989), which described in detail the process of fusing the BARB people meter panel with the TGI study in the United Kingdom, along with a discussion of the validation of the results.

Description of Methods

The goal here is to compare the empirical performances of several methods of data integration of TAM and TGI databases. There are as many ways of data integration as the imagination will allow, and we will be using six methods that have been published.

Method A1: Random Duplication

This method assumes that the estimates from the two databases are statistically independent, and hence can be multiplied together to obtain the incidence of their overlap. For example, suppose the incidence of product usage is 50% according to the TGI database and the rating for a television program is 10% in the TAM database. By the random duplication method, the overlap of the product users who watch the television program is 50% of 10% (or equivalently, 10% of 50%) = 5%.

This method depends on the assumption of statistical independence, which is usually suspect. However, many third-party software processors have built this option for their users (to be used at their own peril).

Method A2: Simulation Method

This method is a refinement of random duplication. This method divides the population into discrete cells that are defined in terms of combinations of gender and age groups (e.g. Men 18-24). Within each cell, it is then assumed that the estimates from the two databases are statistically independent, and hence can be multiplied together to obtain the incidence of their overlap.

This method depends on the assumption of conditional independence (or statistical independence conditional on the gender/age-defined strata). For some examples of the simulation method, see Cannon (1988) and Cannon and Seamons (1995). This method is also known as weighted profile matching (Papazian 1980).

Method A3: Unconstrained Statistical Matching

This method was described by Baker, Harris and O'Brien (1989) for the fusion of the BARB and TGI databases in the United Kingdom. In the research literature on classification theory, this is known as the nearest neighbor classifier (Duda, Hart and Stork 2001).[2)]

In this setup, the TGI database is designated as the donor database and the TAM database is designated as the recipient database. For each recipient, the donor who is most similar in terms of a list of common variables (e.g. gender, age, education, income, occupation, television viewing, etc) is located and the donor's product usage information is transferred to the recipient. This results in what looks like a single source database. Further details can be found in Baker, Harris and O'Brien (1989) and Soong and de Montigny (2001).

Method A4: Constrained Statistical Matching

This method was described by Soong and de Montigny (2001) for the fusion of TAM and TGI databases. The algorithm is based upon solving the transportation problem in operations research, and attempts to match each respondent in one database with one or more respondents in the other database based upon similarity in terms of a list of common variables (e.g. gender, age, education, income, occupation, television viewing, etc). This results in what looks like a single source database, with the properties that the full sample sizes are retained and the marginal distributions are preserved.

Method A5: Predictive Isotonic Fusion

This method was first described by Soong and de Montigny (2003a) and is a "fusion-on-the-fly" method that retains the characteristics of constrained statistical matching, while being a fast algorithm that is customized for optimal accuracy for specific product categories.

This method consists of an initial predictive model (namely, logistic regression) that relates the product usage to a list of common predictor variables (e.g. gender, age, education, income, occupation, television viewing, etc) from the TGI database. This model is derived from the TGI database and then applied to score all the respondents in the TAM and TGI databases. Constrained statistical matching is then done on the predicted scores. This method has the properties of full sample sizes and preservation of marginal distributions.

Method A6: Logistic Regression

An example of this method is given by Baron (2001). A logistic regression model is derived from the TGI database that related product usage by a list of common predictor variables (e.g. gender, age, education, income, occupation, television viewing, etc). This logistic regression model is then applied to the TAM database.

The model results in predicted probabilities of product usage for each TAM respondent. These probabilities are multiplied into the respondent weights directly for analysis, whereas Baron (2001) actually randomly assigned people as users and non-users based upon the predicted probabilities.

Empirical Testing

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

In the United States, a syndicated TAM-TGI fusion product then fuses together the television currency in the Nielsen People Meter panel with the MARS OTC/DTC Pharmaceutical Study. As it happens, the MARS database contains a small number of television-related variables, which would permit a split-sample validation test to be conducted.

In the 2004 MARS study, there were 21,054 intab respondents. We randomly divided the sample into two equal halves of 10,527 cases each. Each respondent has a list of 16 demographic variables (which are those that appear in the Nielsen People Meter panel), 20 product usage variables and 74 television-related variables. The goal is to obtain target group ratings (TGRs).

According to Soong and de Montigny (2003c), the issue is not just estimate bias. In theory, the more complicated methods leverage more information, and that may reduce bias but possibly at the cost of increased sampling variance. Therefore, the analysis will consist of 10 repeated split-samples of the 2004 MARS database, from which a root-mean-squared-error (RMSE) statistic was obtained that incorporates both bias and sampling variance.

Results

For the empirical validation, there were 20 product variables and 74 television-related variables, for a total of 20 x 74 = 1,480 target group ratings (TGRs). For each target group, the true estimate in the sample was compared against the estimate produced by the data integration method. By going across the 10 repeated random split samples, the root-mean-squared-error (RMSE) is obtained for each TGR.

Table 1 summarizes the results. The column titled “RMSE” shows the average RMSE across the 1,480 TGRs. A smaller RMSE means that the method has a smaller combination of bias and sampling variance.

Table 1
SUMMARY OF PERFORMANCE RESULTS FOR TARGET GROUP TV RATINGS

Method	*RMSE*	*Winner-Take-All*	*Mean Rank*
A1: Random duplication	3.49	10.7 %	4.60
A2. Simulation method	2.83	9.1 %	4.39
A3. Unconstrained Statistical Matching	1.57	16.4 %	2.93
A4. Constrained Statistical Matching	1.92	9.7 %	3.84
A5. Predictive Isotonic Fusion	1.47	20.5 %	2.49
A6. Logistic Regression	2.01	33.6 %	2.85

Of the six methods, random duplication is the worst. The simulation method is slightly better, but still not very good. The other four methods are data integration methods that at least have the possibility of leveraging all the available information and are therefore likely to be better than simplistic approaches that discard much of the available information. On the basis of RMSE, predictive isotonic fusion is the best method, followed by unconstrained statistical matching, constrained statistical matching and logistic regression.

When assembling a set of methods to attack the identical problem, there is the tendency to want to view this as a competitive tournament. In competitive terms, the results can be presented in a couple of ways. For each TGR the method that produces the smallest RMSE can be identified. The column titled "Winner-Take-All" shows the percentage of times that each method has the lowest RMSE. For each TGR, the methods can be ranked according to RMSE (rank 1 for smallest RMSE and rank 6 for largest RMSE). The column titled "Mean Rank" shows the mean rank for each method. A smaller mean rank indicates better relative performance.

Bear in mind that such tournament results can be misleading. Essentially, tournament results depend totally on the list of invitees and they mean nothing outside.

From the "Winner-Take-All" column it can be observed that all methods win sometimes. So a theoretically sound and sophisticated method such as logistic regression does not win over a theoretically unsound and dumb method such as random duplication all of the time. This was precisely the point of the *No Free Lunch Theorem*.

On the "Winner-Take-All criterion," logistic regression did surprisingly well better than its position on RMSE. A detailed examination of the results showed that this method consistently performs well on a subset of the TV variables, but not as good on the others. Historically, the “fusion-on-the-fly” proponents have insisted on customized solutions for each product variable instead of one-time-only one-size-fit-all solutions. Meanwhile, the television database has been treated with as one homogenous entity. Here, logistic regression exhibits heterogeneous performance behavior on the TGRs.

The syndicated fusion product of the Nielsen People Meter panel and the MARS survey is based upon the method of constrained statistical matching, although the actual implementation contains some more bells and whistles than the simplified version used for this article. The results here do not suggest that the constrained statistical matching is necessarily superior to the other methods. However, there are some other issues that are outside the scope of the project described here. Those issues deserve an exposition because they determined the ultimate choice.

Consider the method of logistic regression. For Project A, we have narrowly defined the application as target group ratings. In practice, the MARS survey is a full-fledged magazine readership survey and the syndicated fusion product is used for mixed media TV-print planning. In that framework, logistic regression simply cannot be used since there is no mechanism to accommodate magazine readership information.

Consider the method of unconstrained statistical matching. For Project A, we have approximately equal sample sizes for the split samples. In reality, the Nielsen people meter has a daily intab of 9,000 adults while the MARS study has an intab of about 21,000 cases. Unconstrained statistical matching on the syndicated database would have resulted in large losses in the MARS sample size (namely, fewer than 9,000 cases will be used as donors and the rest discarded) as well as distortions in magazine audience estimates (Soong and de Montigny (2001)). It is in fact not an acceptable option for MARS and its clients.

Consider finally the method of predictive isotonic fusion. It is true that this method preserves full MARS sample size and magazine audience estimates. When working within a multiple-person planning team environment, the syndicated constrained statistical matching is easier to manage than the customized “fusion-on-the-fly”-style predictive isotonic fusion where every user may be coming up with their own formulations.

These are the extraneous considerations that come into decisions about data integration methods.

PROJECT B: MULTI-SOURCE FUSIONS

Background

As far as we can tell, this application has not been attempted before. Here is the situation: in a local market, there are multiple media currency systems: a TAM people meter panel for television audience data; a diary sample for radio audience data; and a TGI study of multimedia usage and product usage. The objective here is to bring together the various databases into a single multimedia planning system.

Description of Methods

The goal is to compare the empirical performances of several methods of data integration.

Method B1: Random Duplication

This method assumes that the estimates from the three databases are statistically independent, and hence can be multiplied together to obtain the incidence of their overlap. For example, suppose the incidence of product usage is 50% according to the TGI database, the rating for a television program is 10% and the rating for a radio station is 1%. Under the random duplication method, the overlap of the product users who watch the television program and listen to the radio is 50% x 10% x 1%.

Method B2: Simulation Method

This method is a refinement of random duplication. This method divides the population into discrete cells that are defined in terms of combinations of gender and age groups (e.g. Men 18-24). Within each cell, it is then assumed that the estimates from the three databases are statistically independent, and hence can be multiplied together to obtain the incidence of their overlap.

Method B3: Unconstrained Statistical Matching

In this setup, the TGI database and the radio databases are designated as the donor databases and the TAM database is designated as the recipient database. For each TAM recipient, we locate the TGI donor who is most similar in terms of a list of common variables (e.g. gender, age, education, income, occupation, television viewing, etc) and we transfer the donor's product usage information to the recipient. For this same TAM recipient, we locate the radio diary keeper donor who is most similar on the list of common variables and we transfer the donor's radio listening information. This results in what looks like a single source database.

Method B4: Constrained Statistical Matching

This requires two constrained statistical matching steps. In the first step, the TGI database is integrated with the TAM database by matching on a list of common variables (e.g. gender, age, education, income, occupation, television viewing, etc). In the second step, the combined TGI-TAM database is integrated with the radio database, again by matching. This results in what looks like a single source database, with the property that all three full sample sizes are retained and the marginal distributions are preserved.

Method B5: Predictive Isotonic Fusion

We derive a predictive model (namely, logistic regression) that relates the product usage to a list of common predictor variables (e.g. gender, age, education, income, occupation, television viewing, etc.) from the TGI database. We apply this model to score all the respondents in the TAM, TGI and radio databases. Each product usage variable will have its own predictive model.

Constrained statistical matching is then done on the predicted scores, first to bring the TAM and TGI databases together, and again to bring the TAM-TGI database with the radio database. This results in what looks like a single source database, with the property that all three full sample sizes are retained and the marginal distributions are preserved. There will be a separate database for each target group.

Method B6: Enhanced Predictive Isotonic Fusion

This is an enhancement of the generic predictive isotonic fusion that leverages all the available information from the various systems. This enhancement was described in Soong and de Montigny (2003b).

In the present context, we derive three predictive models (namely, logistic regression models) from the TGI database. For the TAM database, the predictive model is based upon demographic and television variables. For the radio database, the predictive model is based upon demographic and radio variables. For the TGI database, the predictive model is based upon demographic, television and radio variables.

Constrained statistical matching is then done on the predicted scores, as in Method B5. This results in what looks like a single source database, with the properties that all three full sample sizes are retained and the marginal distributions are preserved.

None of the other methods (B1 through B4) mentioned can leverage the additional information in this asymmetrical fashion because those other methods all rely on variables that must be common to two or more databases.

Empirical Testing

We are currently working on the even more ambitious project of integrating the people meter panel, the radio diary sample, the newspaper study and the TGI study in a Latin American country. We would have liked to report data based upon that project, but could not assemble all the pieces in time for this paper.

We therefore used the 2004 Mars study here. In the 2004 MARS study, there were 21,054 intab respondents. We divided them into three equal parts of 7,018 respondents each. Each respondent has a list of 16 demographic variables (gender, age, income, education, occupation, etc), 20 product usage variables, 7 television-related variables and 10 radio-related variables.

The relevant validation criteria are listed and discussed in detail by Soong and de Montigny (2003c). The most significant here are the target group TV ratings, target group radio ratings and the target group cross-media pairwise duplications. We will not deal with the target group TV ratings here, as this subject was covered in Project A above, but will address the other two items.

Results

For the empirical validation, there were 20 product variables and 10 radio-related variables, leading to 20 x 10 = 200 target group ratings. For each target group, we compared the true estimate in the sample against the estimate produced by the data integration method. We did not have the time to do multiple repetitions, and so our results are based upon one set of split samples. The results are shown in table 2.

Table 2
SUMMARY OF PERFORMANCE RESULTS FOR TARGET GROUP RADIO RATINGS

Method	***Mean Absolute Deviation***	***Winner-Take-All***	***Mean Rank***
B1. Random Duplication	2.83	7 %	3.65
B2. Simulation Method	2.93	5 %	3.65
B3. Unconstrained statistical matching	2.95	25 %	3.46
B4. Constrained statistical matching	3.08	15 %	4.18
B5. Predictive isotonic fusion #1	2.79	16 %	3.32
B6. Predictive isotonic fusion #2	2.42	33 %	2.66

With the caveat that this was based upon only one choice of random split samples, we note that the method that stood out in the pack is the enhanced predictive isotonic fusion. It is gratifying to find out that using more directly relevant information does indeed lead to better performance.

Table 3 shows the results for the pairwise TV-radio duplications.

Table 3
SUMMARY OF PERFORMANCE RESULTS FOR TARGET GROUP PAIRWISE TV-RADIO DUPLICATIONS

Method	*Mean Absolute Deviation*	*Winner-Take-All*	*Mean Rank*
B1. Random Duplication	0.33	15 %	3.88
B2. Simulation Method	0.34	3 %	4.20
B3. Unconstrained statistical matching	0.30	29 %	2.86
B4. Constrained statistical matching	0.34	11 %	4.37
B5. Predictive isotonic fusion #1	0.30	14 %	3.06
B6. Predictive isotonic fusion #2	0.28	39 %	2.52

Again with the caveat that this was based upon only one choice of random split samples, we note that the enhanced predictive isotonic fusion stood out. Using more directly relevant information has its benefits.

Although unconstrained statistical matching appears to be within striking range of enhanced predictive isotonic fusion, this is misleading because we have conveniently set our test situation based upon three equal sized split samples. In our ongoing Latin American project, the people meter panel is several times smaller than the other samples. Under those circumstances, unconstrained statistical matching would suffer tremendous losses in sample sizes, and hence reliability.

PROJECT C: PREDICTIVE MODELING

Background

Predictive modeling is used extensively in database marketing, data mining, database marketing, credit card solicitation, credit scoring, insurance prospecting, loan approval, homeland security, etc. (see Weiss and Indurkhya 1998).

In the context of media research, we have encountered the following scenario. On one hand, we have a large national database that deals with a specific subject with a large sample size. An example would be the MARS study that deals with OTC/DTC pharmaceutical product usage. On the other hand, we have a smaller local market that provides a local media currency (such as television, or radio, or newspaper). We are asked to assess the feasibility of integrating the product usage information from the national database with the local market database.

This scenario is different from the traditional ones delineated so far in Projects A and B above. In those other projects, it is assumed that the databases are based upon the same population universe. For Project C, the databases are based upon different universes, which could have very different demographic distributions as well as incidences of product usage. Right away, the notion of constrained statistical matching to preserve incidences is irrelevant.

Description of Methods

The goal here is to compare the empirical performances of five common methods of data integration.

Method C1: Total Independence

This method assumes that the incidence of product usage is homogeneous everywhere. Thus, if the incidence of product usage is 10% in the national database, then everyone in the local market will have a 0.10 probability of being a user.

Method C2: Simulation Method

This method is a refinement of Method C1. The population is divided into discrete cells that are defined in terms of combinations of gender and age groups (e.g. Men 18-24). Within each cell in the local market, we assign the probability of product usage as found in the same cell in the national study. Thus, the overall incidence in the local market will be different from the national average to the extent that product usage varies by gender/age groups and the local market has a gender/age distribution that is different from that of the national population.

This method depends on the assumption of conditional independence (or statistical independence conditional on the gender/age-defined strata). Danaher and Rust (1992) gave an example of linking segmentation studies by assuming conditional independence.

Method C3: Unconstrained Statistical Matching

The national database is designated as the donor database and the local market is designated as the recipient. For each recipient, we locate a donor who is most similar in terms of a list of common variables (e.g. gender, age, education, income, occupation, television viewing, etc) and we transfer the donor’s product usage information to the recipient.

Method C4: Logistic Regression

This is the most classical application of predictive modeling. From a training database (namely, the national database), a logistic regression model is constructed that relates product usage with a list of predictor variables that are also available in the local market database (e.g. gender, age, education, income, occupation, etc). This logistic regression model is then applied to the local market to assign predicted probabilities of product usage for each respondent.

Method C5: Linear Regression

In Soong and Montigny (2003a) it was pointed out that there exists a large number of statistical methods that can be used in predictive modeling, such as multiple linear/nonlinear regression, discriminant analysis, logistic regression, probit regression, tobit regression, proportionate hazard regression, neural networks, support vector machines, kernel methods, nearest neighbor matching, AID, CHAID, CART, MARS and so on. However, Soong and Montigny (2003a) asserted that predictive modelers have found that the specific choice of method seemed to make little or no difference in a fixed problem.

To illustrate this assertion, we use the method of linear regression here. The approach is the same as Method C4. But while logistic regression yields estimates that are interpreted as probabilities of product purchase, linear regression yields numerical scores that do not have physical interpretation (for example, they can be less than zero or greater than 1). It was the contention of Soong and de Montigny (2003a) that the only thing that matters is the relative position of the scores as opposed to their actual numerical values. We have therefore included a linear regression to illustrate this point.

Empirical Testing

The test here is rather severe and exacting. On one hand, we have a national database from the 2003 MARS OTC/DTC pharmaceutical study, consisting of 21,106 respondents. On the other hand, we have the sub-sample of the 2004 MARS OTC/DTC pharmaceutical study that was determined to be in the New York Designated Market Area, consisting of 943 respondents. The two databases are therefore truly different in temporal and spatial aspects, and this will be a test of the goodness of prediction.

For this situation, we consider a set of 12 predictor variables (e.g. gender, sex, income, education, occupation, etc) and 10 product usage variables. We build the predictive model on the national sample and we apply it to the local market sample.

The goodness-of-fit measure of the predictive model could be summarized in terms of standard measures such as the correlation coefficient, R2, likelihood ratio and so on, but they do not provide directly relevant information about the business aspects. Predictive modelers have a more appealing approach, as they would sort the validation into deciles (or quintiles) and examine the actual product usage incidences in the decile. Under the null hypothesis of no effect, the top decile would account for 10% of the users. More effective models will have higher product incidences in the upper deciles.

Results

Table 4 below shows the results for the top decile (10%) analysis. The column titled "Mean Index" shows the actual product incidence among those in the top 10% of the predicted scores in the local market sample indexed by the total incidence. We actually did not need to do the total independence method (Method C1), since that index is 100 by definition. Under the "Winner-Take-All" column, we count the percentage of times that a method has the highest index for each product. Under the "Mean Rank" column", we calculate the mean rank achieved by each method across the products.

Table 4
DECILE ANALYSIS OF INDICES

Method	*Mean Index*	*Winner-Take-All*	*Mean Rank*
C1. Total Independence	100	0 %	4.80
C2. Simulation Method	160	43 %	2.00
C3. Unconstrained statistical matching	135	10 %	3.20
C4. Logistic Regression	154	13 %	2.20
C5. Linear Regression	148	33 %	2.20

Table 5 shows the results for the top quintiles (20%). The column titled "Mean Index" shows the actual product incidence among those in the top 20% of the predicted scores in the local market sample indexed by the total incidence. The other two columns are derived in the same way as in the previous projects.

Table 5
QUINTILE ANALYSIS OF INDICES

Method	*Mean Index*	*Winner-Take-All*	*Mean Rank*
C1. Total Independence	100	0 %	5.00
C2. Simulation Method	152	80 %	1.40
C3. Unconstrained statistical matching	117	0 %	3.60
C4. Logistic Regression	141	15 %	2.50
C5. Linear Regression	142	5 %	2.50

The total independence method was included here as a null baseline, and it did not perform well at all. Absent any information, one might favor the more sophisticated logistic regression method. Instead, we actually found that the very straightforward simulation method worked even better.

There is little or nothing to choose between the two regression methods, which confirms the assertion made by Soong and de Montigny (2003a). Linear regression is a lot faster to execute than logistic regression and the performance characteristics are similar.

Unconstrained statistical matching did not perform well relative to the other methods. Actually, when we think about what is happening, this is very much expected. The version of constrained statistical matching here calls for one-to-one matching. For the 943 local market respondents, one and only one best match was chosen from the national sample of 21,106 respondents. The final fused database therefore contained information from only of 943 donors and the remaining 21,106 – 943 = 20,163 cases never came in at all. This was a severe loss in sample size, hence reliability. By contrast, the simulation method and the two regression methods managed to leverage the full information from the entire national sample.

Predictive modelers are actually less interested in choosing among competitive techniques, since their practical experience is that it makes little or no difference given the same working conditions. Instead, they focus on improving the "working conditions", such as finding a better predictor variable. Picking the right method can move the top decile index up by 10 points, but finding the right predictor variables may well move it by 100 points.

PROJECT D: SEGMENTATION

Background

This has become an increasingly common situation. On one hand, we have these large national syndicated databases (such as the MARS OTC/DTC Pharmaceutical Study). Necessarily, such databases tend to have broad coverage of the general application area (namely, healthcare). They have some information on specific areas but are perhaps not detailed enough for advertisers.

Thus, on the other hand, advertisers have gone out to run their own in-depth custom surveys to drill down on a specific category (such as pain relief medicine), obtaining a lot of qualitative and quantitative information. These custom studies often lead to a market segmentation scheme that will drive their marketing strategies and plans. However, it is not possible to translate this market segmentation scheme into a media plan since the planning database (such as the MARS study) does not have the segmentation information.

Our challenge is therefore to find a way of integrating the segmentation scheme in the custom study into the large syndicated study. Obviously, accuracy is a requirement. The specific objective is to produce target group ratings, where the target groups are defined in terms of the segmentation scheme.

The technical difficulties include the fact that the custom study is rarely based upon a nationally representative sample. For reasons of economy, the sampling frame is either screened from a national sample (e.g. discontinue the telephone interview or web session if the respondent does not qualify) or else it is based upon special lists in which the incidences are believed to be high. Right away, the notion of constrained statistical matching to preserve incidences is irrelevant.

Description of Methods

The goal here is to compare the empirical performances of several methods of data integration for this situation. There are as many ways of data integration as the imagination will allow, and we will be using five methods that have been published.

Method D1: Total Independence

This method assumes that the media usage levels in each segment is identical to that of the total population. If 5% of the population reads a magazine, the same 5% will appear in each and every segment.

Method D2: Simulation Method

This method is a refinement of Method D1. We divide the population into discrete cells that are defined in terms of combinations of gender and age groups (e.g. Men 18-24). Within each cell in the local market, we assign the probability of media usage as found in the same cell in the national study to the incidence of the segment in the same cell of the custom study.

Method D3: Unconstrained Statistical Matching

The custom study is designated as the donor database and the syndicated study is designated as the recipient. For each recipient, we locate a donor who is most similar in terms of a list of common variables (e.g. gender, age, education, income, occupation, etc) and we transfer the donor's segmentation information to the recipient.

Method D4: Multi-group Discriminant Analysis

This is the classical method of classifying individuals into one of several groups using a set of predictor variables (see, for example, Duda, Hart and Stork 2001.[3] We use the custom study to derive a discriminant analysis model that assigns probabilities of belonging to several segments (such that the sum of those probabilities add up to 1.0 since the segments are mutually exclusive and exhaustive) based upon the predictor variables that are common to both databases (e.g. gender, age, education, income, occupation, etc.).

We then apply this discriminant analysis model to the national study, so that each respondent is assigned probabilities of belonging to the various segments. We could make a random assignment according to those probabilities, but we do not. Instead, we kept the probabilities and apply them to the respondent weights. This is a cleaner approach that is not subject to the vagaries of the randomized assignment process.

Method D5: Enhanced Unconstrained Statistical Matching

The first four methods may be the obvious things to do under the 'working conditions.' In practice, we should always try to improve the 'working conditions.' In the empirical testing, we had a situation in which we analyzed the segmentation data and deduced that while it was not feasible to ask the entire battery of attitudinal questions, it was possible to ask four simple yes/no questions and still be able to get a very effective predicted classification of segmentation membership.

Thus, this enhanced version of unconstrained statistical matching involves four more predictor variables than those used in Methods D1 through D4.

Empirical Testing

The test database is the 2004 MARS OTC/DTC Pharmaceutical Study. The sampling frame for this study actually corresponds to the situation that we have described. On one hand, we have a nationally representative sample of 8,560 respondents. On the other hand, we have 12,494 other respondents who were sampled from specialized lists of people who are likely to have various types of ailments.

Both samples are assumed to share 19 common variables, including demographics (gender, age, income, education, occupation) and some basic health indicators (current health rating, insurance coverage, prescription drug coverage).

The segmentation scheme was determined via a k-means clustering algorithm applied to a battery of questions about specific actions taken as a result of seeing health-related advertising. The details of this segmentation scheme is reported in White, Draves, Soong and Moore (2004).

For each method, we integrated the segmentation information from the oversample portion of 12,494 respondents onto the national portion of 8,560 respondents. This permits us to compute target group magazine ratings by segment. There are 4 segments and 100 magazines, so that each method yields a total of 4 x 100 = 400 target group magazine ratings. The estimate from the data integration is then compared to the actual target group magazine rating.

According to Soong and de Montigny (2003c), the issue is not just estimation bias. In theory, the more complicated methods leverage more information, and that may reduce bias but possibly at the cost of increased sampling variance. Therefore, our analysis will consist of 10 jackknife replicates from the 2004 MARS database, from which we obtained a root-mean-squared-error (RMSE) statistic that incorporated both bias and sampling variance.

Results

Table 6 below shows the summary performance measures. The column titled “RMSE” shows the average RMSE across the 400 TGRs. A smaller RMSE meant that the method has a smaller combination of bias and sampling variance. For each TGR, we can identify the method that produces the smallest RMSE. The column titled “Winner-Take-All” shows the percentage of times that each method has the lowest RMSE. For each TGR, we can rank the methods according to RMSE (rank 1 for smallest RMSE and rank 5 for largest RMSE). Under the column titled “Mean Rank”, we show the mean rank for each method. A smaller mean rank indicates better relative performance.

Table 6
SUMMARY MEASURES OF PERFORMANCE FOR TARGET GROUP MAGAZINE RATINGS

Method	*RMSE*	*Winner-Take-All*	*Mean Rank*
D1. Total Independence	1.58	5 %	3.32
D2. Simulation Method	1.42	2 %	3.07
D3. Unconstrained statistical matching	1.40	3 %	3.65
D4. Multi-group discriminant analysis	1.42	46 %	2.57
D5. Enhanced unconstrained statistical matching	0.68	45 %	2.37

The method of total independence fared the worst. In terms of RSME, there is not much to choose among methods D2, D3 and D4, but the multi-group discriminant analysis is better than the other two on the “winner-take-all” and mean rank measures.

The enhanced unconstrained statistical matching method is the best of all. This is not because the technique is superior inherently, but this implementation was able to use some auxiliary information that was directly relevant to the problem. Although we did not have the time to do so, we are confident that both the simulation method and multi-group discriminant analysis will also show significant improvements if we could incorporate that same information.

CONCLUSIONS

So far, we have shown the results from four different data integration projects for which we had conducted a total of 11,094 different data integration exercises. What did we learn from all this?

We recall that we were motivated to illustrate the *No Free Lunch Theorem*. Indeed, no single data integration method was consistently superior to all other methods. In fact, some methods cannot even be applied. For example, the logistic regression did well in Project A but cannot be deployed for Project B; as another example, constrained statistical matching did not even make sense for Projects C and D.

Within each project, there are multiple outcomes (e.g. target group ratings). We have been reporting on the basis of global performance. But we have seen evidence that a data integration method can perform consistently well for a subset of the outcomes while being mediocre on others (e.g. logistic regression in Project A).

For the projects here, our conclusions are confined to the specific sample configurations and sets of variables. Thus, if there is a new project that is different from the four projects here, either in problem definition, or sample configuration, or sets of variables, we cannot then confidently predict how any of these methods will perform.

For example, if we have Project A based upon much smaller sample sizes, then the methods may perform differently. If we have a different set of product usage variables, we may have better or worse performances from these methods. If we have a different set of predictor variables for Project C, we may have better or worse performances from these methods. If we have a different segmentation scheme for Project D, we may have better or worse performances from these methods. If we choose to have a different application from the same project, we will have a different set of validation criteria and our performance assessments may be different for the methods.

Hopefully, we will have communicated a healthy skepticism to our readers with respect to methods that are alleged to possess overall superiority. If someone asserts without proof that there is a best method for a brand new situation, it would be fair to attribute that person either as an ignoramus or a prevaricator. And if the proof is offered in terms of the performance of that method on an unrelated problem, all the worst for that person.

We do not wish to drive our readers to distraction and despair. While we say that we cannot generalize arbitrarily to any unseen new problem, we do know enough to tell what is likely to work. The best approach is to focus on the aspects that matter most: problem formulation, application purposes, relevant factors, data availability, data distribution, evaluative criteria, user requirements, application software restrictions, and so on. For each problem, we will find that a method that is crafted to match the exact circumstances is likely to be the most successful.

The other important lesson that we draw is that very often there is little or nothing to choose among various methods given the same working conditions. But when we work hard to improve those working conditions such as by finding more powerful predictor variables (as in Method D4), the performance can be improved dramatically across all methods. This approach will yield much higher payoffs than the pursuit of a better method.

There is no free lunch here. We have to work hard to earn it.

FOOTNOTES

1. Duda, Hard and Stork (2001), chapter 9.
2. Duda, Hart and Stork (2001), chapter 4.

3. Duda, Hart and Stork (2001), chapter 5.

REFERENCES

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

Baron, R. (2001). A new practical approach to data fusion. Paper presented at ARF Week of Workshops, Chicago, IL.

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

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

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

Duda, R.O., Hart, P.E. and Stork, D.G. (2001). *Pattern Classification*. New York: John Wiley & Sons.

Hosemer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons.

Papazian, E (1980). Using product usage data in media selection. *Marketing and Media Decisions*, 15(7), 16-20.

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

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

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

Soong, R. and de Montigny, M. (2003c). Foundations of split-sample foldover tests. Eleventh Worldwide Readership Symposium, Cambridge, MA, USA.

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

White, H.J., Draves, L.P., Soong, R. and Moore, C. (2004). Measuring the effect of direct-to-consumer communications in the world's largest healthcare market. *International Journal of Advertising*, 23(1), 53-68.

THE AUTHORS

Roland Soong is Chief Technical Officer, KMR, United States.

Michelle de Montigny is Executive Vice President, KMR, United States.