



Dynamic Segmentation Fusion

Dr. James Collins, Senior Vice President, Mediamark Research Inc.
Dr. Gina Pingitore, Chief Research Officer, J. D. Power and Associates

Introduction

Advertisers increasingly need to expand their brands' franchises beyond traditional marketing channels to include multi-/cross-media platforms. At the same time, efficient utilization of such multi-platform media requires increased consumer insights. Obtaining detailed and expansive information from a single research survey respondent, however, is generally impractical for a variety of research quality related and economic reasons.

As an alternative, media researchers and others are vigorously pursuing data integration techniques whereby independent databases are integrated through rigorous, formal statistical procedures. The results are integrated data sources that can be used to develop cross-media consumption and related consumer insights.

One prominent data integration technique is fusion. While the details of particular fusions are numerous, most are of one of two forms, static/monolithic or analysis-time. With static/monolithic fusion respondents from different datasets are matched once using a predetermined hierarchy of common variables. In contrast, analysis-time fusion develops the matching algorithm (near) optimally for each data element of a particular analysis.

Recently Mediamark Research Inc. (MRI) and J. D. Power and Associates (J. D. Power) undertook several fusion projects involving the integration of complex sets of consumption, attitudinal and media data all within the context of automotive media/market research. Results from these fusion efforts have shown that neither the static nor the analysis-time fusion techniques are completely satisfactory. As a result an alternative approach - Dynamic Segmentation Fusion - was developed.

Generally, Dynamic Segmentation Fusion begins by using one dataset to systematically develop a classification/regression tree-based model based upon common variables within the datasets (e.g. demographics, attitudes/opinions, media, product preferences, etc.). Next, each dataset is processed through the model such that respondents from each are classified into corresponding sets of relatively discrete and homogeneous nodes/segments. Respondents from corresponding nodes for each of the datasets are then matched for the fusion.

Some advantages of this technique over its predecessor methods are:

- 1) No a priori assumptions with respect to the hierarchy of common matching variables need be made – the classification model generates the matching scheme
- 2) Persons/records are conveniently organized node-to-node for the purposes of matching
- 3) If there is not a single exhaustive segmentation of persons/records between the two datasets, then the classification can proceed iteratively with one set of persons/records being fused using the dominant segmentation scheme, with the classification tree being rebuilt and/or pruned to classify any remaining persons/records.
- 4) It produces a stable database usable in a variety of different analysis systems and yielding consistent results across all analyses

This paper will 1) explain the Dynamic Segmentation Fusion strategy in more detail, 2) demonstrate the extent to which relationships in the data source are preserved in the fused dataset and 3) discuss the application of this integrated information source in automotive marketing contexts.

Data Sources

The J.D. Power Offline Media Reports (OMR) surveys new car/truck buyers from R.L. Polk's vehicle registration database of new vehicles registered. For this initiative, vehicles registered between May 2005 and April 2006 were randomly selected within a vehicle model stratification. Historical response rates for each model were used to determine the total number of mail outs by model to ensure each model met a minimum quota of fifty returns. The resulting sample included approximately 44,000 respondents from among 325 car and truck models from American, European and Asian manufactures. This final sample was then weighted to the total sales for each automotive make/model for the given time period.

In addition to covering general personal and household demographics, OMR also assesses a variety of media consumption behaviors including readership of approximately 135 magazines as well as radio listening, cable network viewing and internet site visitation. As the primary focus of this study is new vehicle buyers, buying dynamics and preferences, a variety of automotive and transportation behaviors and attitudes are also assessed.



As a broadly focused print research study and most relevant to this project, the MRI survey also measures media consumption patterns with respect to magazines, radio, television, the internet, etc. Also, and similar to OMR, MRI assesses automotive buying preferences and the demographic characteristics of the respondents.

MRI, however, has two important differences with OMS in respect to this data integration project.

The first significant distinction with OMR is that MRI is a general population survey of United States adults (Age 18 and older) and uses an area probability sampling methodology with over-sampling based on geography. Final projection weighting/sample balancing is done to standard national population estimates. While this sampling methodology is particularly well suited to magazine audience measurement (and the measurement of other media, consumption behaviors and attitudes and opinions generally) a consequence is that in the MRI sample for relatively small incidence automotive makes/models there are few MRI respondents who own such vehicles. In contrast, J. D. Power explicitly avoids such a condition by constructing their sample so as to explicitly insure that each automobile make/model is adequately represented in the database.

The second relevant distinction is that MRI contains a more extensive assessment of consumer behavior and attitude, opinion and preference measures. It is these measures that are of central importance to this data integration project. In particular, additional behavioral and psychographic characteristics from MRI are fused to the appropriately matched new vehicle buyers comprising the OMR. These measures include a variety of automotive and travel related behaviors as well as attitudes and preferences regarding Technology, Food, Buying Styles, Consumer Confidence, Advertising and Lifestyle. When these characteristics are fused to new vehicle buyers at the individual make/model level, a richer database is available to media planners, marketers and product development teams, facilitating more targeted marketing as well as enhanced product development.

Dynamic Segmentation Fusion

MRI developed Dynamic Segmentation Fusion in large measure to deal with the exigencies and conditions of this and similar fusion projects.

Because of the relatively low incidence for some automobile makes/models in the MRI dataset MRI's doublebase was employed (i.e., respondents from two successive years of data collection were used). Moreover, consideration sets were developed within the MRI database for each of the 325 makes/models measured in the J. D. Power OMR survey.

These consideration sets were developed using data supplied by the OMR respondents regarding what other makes/models were considered when making the buying decision for the particular make/model purchased. In reality these consideration sets were relatively tightly defined – for example the Honda Accord set included Toyota Camry buyers among other comparable vehicles. The pool of Honda Accord “owners” in the MRI dataset was comprised of all Honda Accord owners and owners of other cars in the consideration set, with low occurrence make/model mentions being excluded as noise.

The advantage of this technique is that it did not force a relatively larger number of OMR respondents to use the data of a relatively small number of MRI respondents for lower incidence makes/models. Analyses have generally demonstrated that even though not all MRI individual make/model “owners” do in fact own the particular make/model that these consideration sets do nicely represent individual make/model owner characteristics (demographics and automotive attitudes/opinions).

Initial analysis revealed that a monolithic fusion (i.e., matching on a single hierarchy of common measures irrespective of make/model) would not suffice. This became evident upon initial examination of the OMR data which showed that demographic and attitudinal dynamics of automotive ownership in the United States varied significantly by make/model. Specifically, we found that quite different sets of demographic and attitudinal characteristics were related most strongly to different make/model buying selections.

Examples of distinctive sets of predictive characteristics can be easily seen by considering four different automobile makes/models:

- 1) Toyota Prius - a very fuel efficient hybrid
- 2) Hummer H2 SUT– a relatively large and expensive sports-utility vehicle of military origin.
- 3) Cadillac DeVille – a full-size luxury sedan
- 4) Chevrolet Trailblazer – mid-priced SUV vehicle with family-oriented features

Charts #1 through #4 include the output from make/model specific classification tree analyses. In the respective analyses the dependent variable was ownership of each of the four makes/models and the independent/predictor variables were a series of demographic and automotive-related attitudes and preference measures common to both datasets and used in the fusion itself.

- c. Vehicle that “Stands Out from the Crowd” – Important
- d. Acceleration and Good Handling – Important
- 3) Cadillac DeVille
 - a. Age – Older
 - b. Household Income – High
- 4) Chevrolet Trailblazer
 - a. Married
 - b. Have Children at Home
 - c. Household Income – Middle/High

In short, very different sets of discriminating measures (the common matching variables) are needed for effective matching at the individual make/model level. A single hierarchy of matching variables is not viable.

Hence, stronger fusion strategies would seem to be run-time or point-of-analysis ones such as Gilles Santini’s “Just-in-Time-Fusion” or Soong’s “Fusion-on-the-Fly” as they assume no a priori, monolithic hierarchy of common matching variables for respondent pairing. Rather, by design they possess the flexibility to develop optimal matching algorithms for the measures constituting a particular analysis. In the context at hand, the matching hierarchy would be dependent on the particular automotive makes/models and other measures constituting a particular analysis.

However, a variety of practical reasons prevented the application of these run-time techniques. The two most important barriers being, 1) the need for consistency of specific measure levels from analysis to analysis and 2) limited availability of end-user run-time fusion software.

Thus, it was the appeal of dynamic/analysis-time fusion techniques in combination with static fusion’s ability to generate a stable and persistent database that provoked our thoughts toward what we are terming *Dynamic Segmentation Fusion*

The Basics of Dynamic Segmentation Fusion

Dynamic Segmentation Fusion is comprised of four phases:

Phase #1. Using either the donor or recipient dataset develop a classification tree model for each automotive make/model (dependent variable) separately, with the common linking measures as the independent variables. In doing the classification, minimize the deviation within terminal nodes or leaves and allow for small terminal node sizes. The result will be a model embodying a well defined set of classification rules. (By way of reference, Santini’s Just-in Time-Fusion uses a similar tree-based procedure and



Soong's Fusion-on-the-Fly technique employs multiple regression, functionally quite similar.)

In this case, the classification model was developed within the recipient dataset (OMR) because it possessed the more robust automotive make/model sample size and since make/model related characteristics were what needed to be determined most precisely.

We are fortunate in this fusion as MRI's and the OMR datasets have extensive measures in common and readily available for matching. In the case at hand, important demographics such as Sex, Age, Ethnicity and Census Sub-Region, as well as a variety of household-related characteristics, in addition to a number of automotive-centric attitude/opinion measures and magazine readership variables were available and employed. As noted previously, these measures serve as the independent variables in the classification tree models with ownership of each individual make/model as the dependent one in each of the separate trees. Thus, as there were 325 automotive makes/models measured in the OMR, 325 separate classification trees were developed and used in the matching process.

Phase #2. Using the tree model, donors and recipients for each make/model are classified into separate but isomorphic sets of leaves/nodes resulting in comparable but separate sets of donors and recipients.

Phase #3. Using these separate isomorphic sets, matching of donors with recipients takes place between comparably classified pools of donors and recipients.

As a practical matter, by controlling 1) against excessively high rates of donation which may occur if one or more donor nodes is substantially smaller than the corresponding recipient one, and 2) for important values such as Age or Sex, not all recipients are necessarily initially matched with donors.

Phase #4. Recursively "prune the tree" to collapse leaves/nodes which are least discriminating and re-classify both donors and remaining recipients, repeating Phases #3 and #4 until all recipients are matched with donors.

This recursive approach involves some level of compromise between exactly matching on more measures and limiting the numbers of times an MRI respondent can donate to an OMR recipient. To minimize the extent of this compromise, pruning is undertaken in relatively modest increments with the number of nodes being reduced by approximately 5% on successive iterations.

Results

For fusion, as for any other data integration technique, a central aspect of the validation of the particular exercise is the extent to which the known relationships resident in the donor database (MRI) are maintained within the fused dataset. Therefore, a key validation of our fusion effort was preservation in the fused dataset of the attitudinal characteristics at the individual make/model level from the MRI database.

Charts #5 and #6 (Toyota Prius and Chevrolet Silverado) below depict comparisons between the set of automotive related attitudes and opinions as native to the MRI database with their manifestation in the resulting fused one. As indicated previously, approximately one half of these measures are common to both the MRI and OMR datasets and used as linking variables (independent variables in the classification tree). Thus, at a minimum, if these relationships are not well preserved the general integrity of the fusion is questionable.

As is evident in the charts below the relationships are generally well preserved.

Chart #5 – Toyota Prius Automotive Attitude/Opinion Profile Comparison

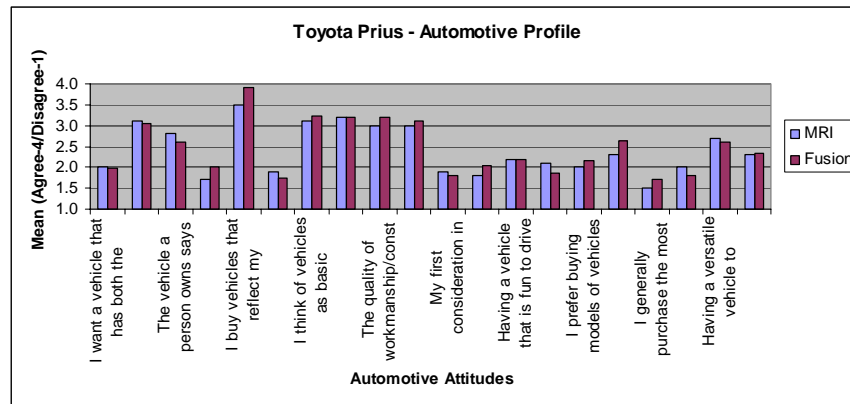
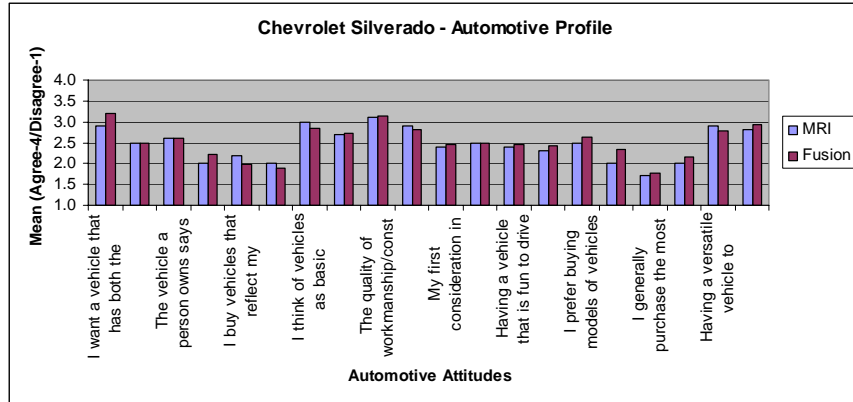


Chart #6 – Chevrolet Silverado Automotive Attitude/Opinion Profile Comparison



While it is reassuring that the MRI automotive attitude/opinion profile is well preserved within the fused dataset this is almost a de minimis condition; it is expected given the explicit control exercised throughout the fusion on many of these measures. Therefore, we also examined other attitude/opinion items for which there was minimal or no explicit control - for example, extensive batteries relating to buying styles and financial attitudes. Charts #7 - #10 depict the profiles, again for Toyota Prius and Chevrolet Silverado, for these two attitude/opinion batteries. Again, and reassuringly, the profiles are quite similar.

Chart #7 – Toyota Prius Buying Styles Profile Comparison

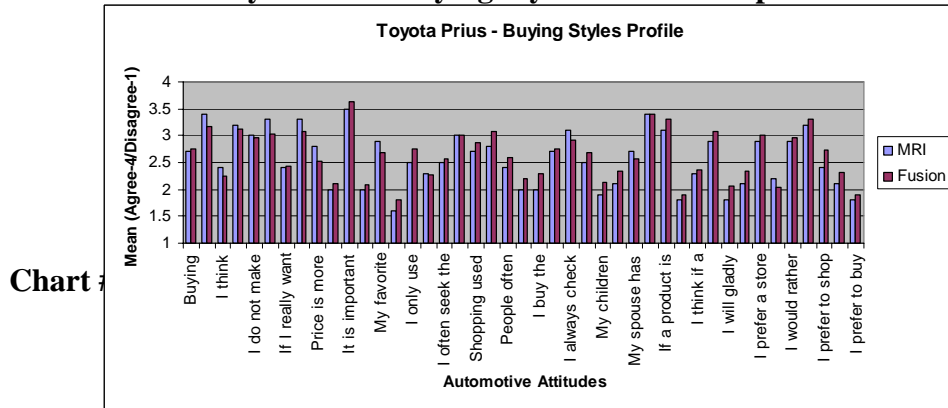


Chart #8 – Chevrolet Silverado Buying Styles Profile Comparison

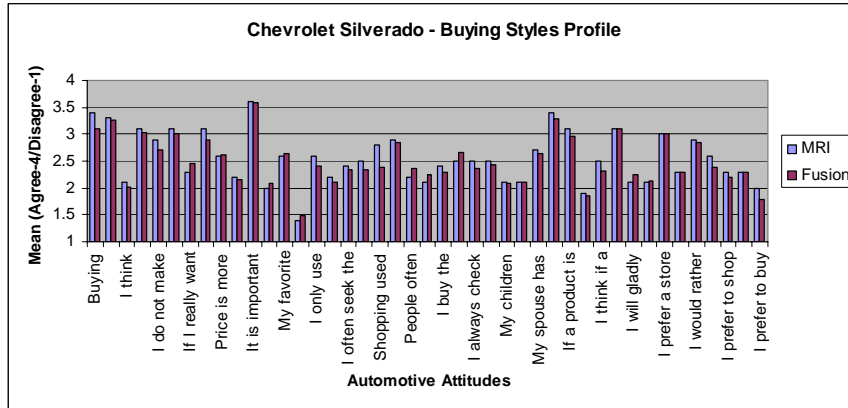


Chart #9 – Toyota Prius Financial Profile Comparison

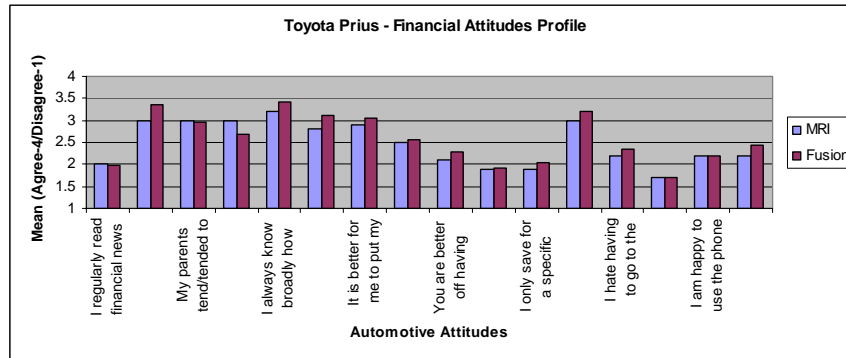
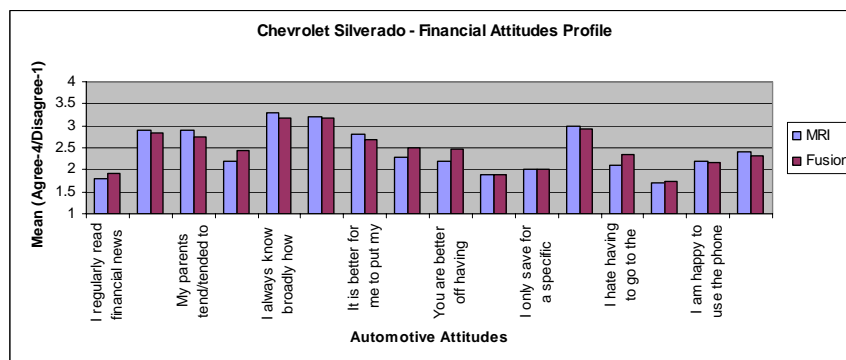


Chart #10 – Chevrolet Silverado Financial Profile Comparison





Application of the MRI/JD Power and Associates OMR Fused Dataset

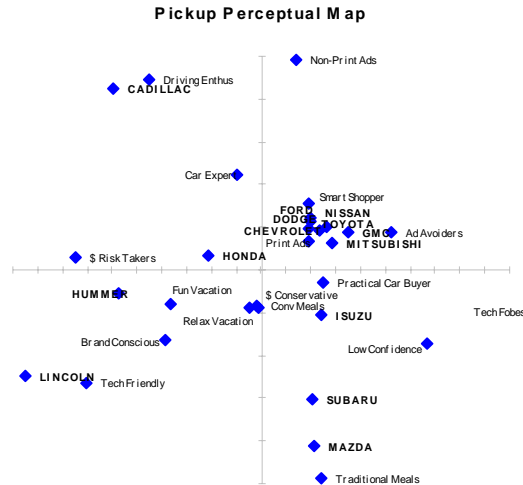
Automotive manufacturers aim to identify meaningful differences between buyers of specific models within each vehicle segment. This is because model level discrimination helps both in the product design phase as well as in the sales and marketing efforts. But finding meaningful differences is challenging due, in part, to the fact that acquiring the necessary depth of insight from a single respondent in a single survey is itself challenging. As researchers know, executing extensively detailed research poses substantial and complex challenges: longer surveys have lower survey response rates, increased missing data rates, and overall higher study costs. Additionally, there are numerous findings indicating that even if researchers entice respondents/consumers into completing a lengthy survey, intra-respondent variation decreases notably suggesting respondent “fatigue”/“burn out” that may result in limited insights.

Because of this, we decided to use the fused dataset to determine whether we could find meaningful differences between models within each segment. In doing so, we not only demonstrate one example of the application of this fusion effort, but we also provide some face validity evidence as well.

We started by factor analyzing all the psychographic questions within fused data – automotive and non-automotive attitudes/opinions. This resulted in 17 different categories. We then used these factors in separate discriminate function analyses within each automotive segment (e.g., compact car, midsize car, midsize SUV, etc.). These discriminate function analyses showed significant differences between brands within a vehicle segment.

To illustrate these differences, we constructed perceptual maps. As an example, Chart #11 below shows the mapping of various models within the pickup truck segment. As can be seen, by using the fused dataset we were able to identify significant differences among the items within this segment. Brands offering traditional pickups group together close to Smart Shoppers. Lincoln and Hummer H2 attract buyers with similar attitudes about being technologically friendly and brand conscious; however Hummer buyers are more willing to take risks with their investments.

Chart #11



We also wanted to examine whether the fused dataset could be helpful in identifying differences between buyers and new targets (i.e. consumers shopping in that segment who purchased a different model). These types of analyses can be used to help marketers understand how to better position both their advertising messages and placement to capture new buyers.

Table #1 below depicts an example showing the percent of respondents selecting the top box response for selected MRI Personal Values LifeMatrix items. Specifically, the comparison is between those purchasing the Chevrolet Malibu and those purchasing a midsize car who did not consider a Malibu (Conquest Target).

Table #1 Malibu versus New Targets

	Malibu Buyers	New Targets	Delta
Being creative, imaginative	50.0%	82.0%	32.0%
Fitting into nature	58.1%	75.8%	17.7%
Wanting to explore and learn about new things	71.6%	89.2%	17.6%
Seeking adventure and risk	24.3%	41.5%	17.2%

The non-Malibu automotive buyers were substantially more creative and adventure seeking than the Malibu buyers. One practical implication of these data is that companies, such as Chevrolet can develop greater insights into how best to develop an advertising message which will more effectively resonate with these new target automotive consumers.

Conclusions

Advertisers, their agencies and the media serving them all crave broader, deeper and more refined insights into the relationships among consumers' purchase behaviors, media consumption habits and the variety of attitudes, opinions and dispositions motivating them. Regrettably, often times the best way for researchers to measure one consumer dimension is to exclude consideration of most others. Yet it is exactly in this constellation of assessments that the most genuine and actionable consumer insights often lie.

Hence, the persistent interest on the part of research providers and users in all manner of data integration strategies, methods, tools and techniques. Certainly, and on theoretical grounds inevitably, all data integration techniques entail some level of information loss – regression to the mean – some loss in the precision and resolution of the relationships being *constructed* from what is *true* (but unknown).

Ultimately though, as important as validation is to all data integration exercises (“How close to *true* are the *constructions*?”), an ultimate test of data integration, as with research generally, is a utilitarian one: “Does it allow us to make better decisions?”.

Thus, the ambition of this work is necessarily two-fold. First, to put up for review and consideration by the research community a particular fusion strategy, one grounded in prior technique, and from which we and our clients have derived value. Perhaps more ambitiously though, we hope others will find use for these ideas in their own work and that it serves to encourage and enrich consideration of the place and contribution of data integration in the utility media and market research continually strive to offer.