1. Introduction

We want more information than consumers are prepared to give: we want to know what products they buy, what TV programs they watch, websites they visit, magazines they read, and more. Trying to get all this information from a single piece of research is problematic and entails compromise: we may have to accept lower response rates, or perhaps less detailed data, such as claimed TV viewing rather than the metered data that forms the TV currency. And last but not least, custom single-source research is expensive.

Data fusion sidesteps these issues by merging existing databases into a single respondent level database. This paper explains the data fusion process, discusses validation of the process and gives some examples of how data fusion can provide a deeper understanding of consumers.

2. What is data fusion?

Data fusion is a method of integrating data. What sets it apart from other approaches is that it is a respondent-level matching of data sets: respondents from one survey are paired with respondents from another survey. The matching uses the common characteristics of the two surveys; these are usually demographic (household and persons characteristics), geographical, media and any other relevant information.

The principle underlying data fusion is that the common characteristics can reliably predict the behavior of consumers in terms of (for example) TV viewing by product users. A very simple example is as follows:

---

**Simple example illustrating the data fusion principle**

1. **Canned fruit users** tend to be **older people**
2. **Older people** tend to be **heavier TV viewers**
3. **Canned fruit users** tend to be **heavier TV viewers**

---

**THE FUSION PROCESS**

- **Data Base A** (eg. NPM)
  - Common characteristics
  - Data unique to A (TV)

- **Data Base B** (eg. Homescan)
  - Common characteristics
  - Data unique to B (Consumer Purchasing)

**FUSION** (linking via Common Characteristics)

**Fused Database**
- Common characteristics
- Data from A and B (TV Viewing of Product Purchasers)
Of course life is more complicated than this, and data fusion recognizes this complexity, by simultaneously taking into account many demographic and other classifications to create an accurate picture of the consumer across all the common characteristics of the linked surveys. This means that data fusion is enhanced targeting. It moves planning away from blunt demographics to a precise view of product users and media habits.

3. Linking variables: the key to a reliable data fusion

The common characteristics used to link the two surveys in the data fusion are referred to as "linking variables" or "fusion hooks". The reliability of a data fusion is related to how well the linking variables explain consumer behavior, both product purchase and media use. This can be assessed by statistical techniques, and the evidence from such analyses is that data fusion is reliable. Certainly data fusion will always provide a more precise view of consumers than simple demographic targeting, making planning more efficient.

Inevitably, some aspects of consumer behavior are more predictable than others: the purchase of baby products is directly correlated with presence of 0-2 year-olds in the household for example. Some argue that for products such as diapers or baby food, simply analyzing by the demographic of presence of 0-2's is sufficient and there is no need for data fusion: however data fusion gives a deeper understanding of this segment, for example in understanding heavy users of tinned baby food versus light users, or the difference between diaper brands.

Data Fusion: who, why and when

Data fusion originated in the United States in the 1970's and has been used in media research in Europe since the late 1980's. It is now used globally in media and marketing research. Despite originating in the US, its use has lagged behind Europe until recently, perhaps because access to the respondent-level data that is essential to the process was relatively more restricted. In the last couple of years, there has been a proliferation of data fusion products from the Nielsen Company and other organizations, with a wide client base. It is no longer true to say that "America doesn’t buy fusion".

The main reasons for using data fusion are:

a) It is cost-effective and convenient.
- it requires no additional (and expensive) primary research
- the data fits into existing analysis systems relatively easy
- it can harness the best available data, including currency measurements

b) Data fusion gives respondent level frequency enabling schedule analysis and optimization.

c) It is the most accurate method of targeting in the absence of single source data.

d) It improves planning efficiency, making clients more profitable.

4. Data Fusion Validation

Any modeling technique requires validation, and data fusion is no exception. The only truly direct and reliable validation of data fusion is to have a perfect single source data-set as a benchmark, but therein lies the Catch 22 of data fusion validation: if a perfect single source data-set existed, there would be no need to produce a data fusion!

That said, there are many ways of assessing the reliability of a data fusion. First, there are checks and balances applied to the fusion production, to ensure that the various elements that go into the process are optimized. Second, the fusion results can be compared with relevant available data. This sort of comparison needs to be treated with caution as there may be measurement differences that over-ride any possible data fusion bias. For example, comparing a result from an NPM/Homescan fusion with single source data from a recall study is likely to highlight inaccuracies in respondent recall of both TV viewing and product purchase.

A third approach is a split sample evaluation and this method is probably the best way of evaluating the reliability of data fusion results.
Split-sample Validation

Split-sample validation, also known as foldover validation, is a commonly used technique in model validation. It works as follows:

a) A data-set is split into two parts A and B.
b) Part B is fused to A.
c) The real results from Part A are compared with the fused results from Part B.

NielsenConnections has conducted extensive validation of the fusion process using a split-sample approach. The chart below shows real and fused weekly TV viewing levels for product users, across 1200 products (note that these are simulated product classifications applied to the NPM for the purpose of the validation). It can be seen that the fused data track the real data closely.

Further statistical analysis allows us to assess whether the differences between the real and fused data are evidence of bias in the data fusion process, or whether they are within the margin of error we would expect given the sample size. On a positive note, the converse of bias is accuracy – 0% bias is 100% accuracy. By testing a large number of results, we can reliably assess the level of bias (and therefore accuracy) of the data fusion, and generate accuracy statistics. The validation work we have done to date shows that the accuracy levels of our fusion are in the range of 90-95%.
5. Data Fusion Application – Selling Cars in Chicago

This example shows how data fusion can maximize the communication of a cross-media plan. It is based on a three way data fusion of TV, Internet and Outdoor in Chicago, using May 2006 LPM data as a base. Onto this base we fused internet data from Nielsen Online’s NetView panel, and billboard/poster exposure data from Nielsen Outdoor.

One immediate benefit of this fusion approach was the quality of the measurement: in each case, state of the art electronic measurement was employed:
- People meters for TV
- Internet meters for online
- Personal GPS tracking for outdoor

Aside from the difficulties posed by recall measurement for these media, the electronic measurement allows a level of detail – individual programs, websites and billboards that simply would not be available otherwise.

The people meter panel also has car ownership data that allows us to move beyond demographics to target product purchasers. The fusion combined all of the data elements into a respondent level data base allowing campaign evaluation against a target – in this case, SUV/Minivan buyers. An example of the power of this is given here: the chart presents the results of an optimization exercise, showing the trade-off between CPM and reach as the media mix shifts. One element was held constant – the weight of the campaign was fixed at 1000 GRPs, starting with all the GRPs in TV and progressively moving 50 GRPs out of TV and into Internet and Outdoor equally. The sweet spot for optimizing total campaign reach in this case was at 600 GRPs in TV, 200 in each of TV and Internet. Moving more GRPs out of TV reduces CPM further but at the cost of reach. And although the findings are similar for adults 18+ and for SUV/Minivan prospects, they are not identical: the optimal reach for adults 18+ occurs at 55% TV rather than 60% for SUV/Minivan prospects.

The fused database also allows us to understand how the media interact. We can see how the different elements of the schedule combine to generate the overall reach delivery of 93%.

Without data fusion, one would resort to random duplication to estimate the total reach delivered by the campaign. In this example, the reach estimate delivered by random duplication is 86%, significantly lower than the 93% shown by fusion. Media interactions are not random – TV and outdoor for example are by their nature complementary (you’re either indoors or outdoors after all) and a campaign that uses both will have a wide reach.

Campaign reach across the three media (93% reach overall)
6. Conclusion

No human endeavor is perfect and all research contains the potential for error and bias. Data fusion works within the realm of the possible, not the unattainable. A good fusion can use the best data available, such as media currency data and electronic measurement that minimizes respondent bias. A fused database using the best measurement data available will provide an accurate, flexible and actionable data-set, to optimize decision making and maximize efficiency for our clients.