

Enhance Lists with Overlay Data

*By Jim Wheaton (co-author)
Principal, Wheaton Group*

Original version of an article that appeared in the April 7, 2003 issue of "DM News"

Demographic overlay data can improve your company's top and bottom lines. This is because it will supplement what you already know about your customers, inquirers and prospects, thereby sharpening the focus of your targeting efforts.

However, all of this is contingent on properly interpreting the output reports from a demographic overlay. This is something that most direct marketers assume to be a straightforward process. They reason that anyone can understand the averages, means and frequency distributions that comprise the variable-by-variable profiles. Therefore, they do not hesitate to draw sweeping conclusions that can have profound strategic and tactical ramifications.

The Problem of Missing Data

Unfortunately, it can be more difficult than one might think to interpret the profile reports that are generated by the demographic overlay process. There are several analytical traps that frequently snare the untrained.

We will focus here on missing data, which corresponds to that proportion of customers, inquirers and/or prospects for which overlay information does not exist. We will illustrate how it complicates the interpretation of profile reports, and can be responsible for significant distortions.

An Example

A dramatic example occurred when a well-known data compiler performed a demographic overlay on the subscriber files of several magazines owned by a publishing conglomerate. The results were organized into a formal presentation, and delivered to the client's CEO. One of the findings was that active subscribers to one of the magazines had an average age of 44.

Effectively, that was the end of the presentation! Because of this assertion, the compiler immediately lost all credibility in the eyes of the CEO. Here is what had happened:

In order to assist in selling ad pages, the publisher had done extensive survey research on its subscribers. Many of the hundreds of data elements on the compiler's overlay file did not intersect with the publisher's research. However, the Age element did. As a result, the CEO knew the average age of the title in question to the nearest tenth of a year, which was just a hair over 30.

Reasons for the Miscalculation

Individual and household overlay data generally cannot be applied to a significant portion of a given file. The magnitude generally ranges from 15% to 95% for specific data elements. In this case, about 80% percent of the magazine's active subscriber file could not be overlaid with age data, which is extreme for this element.

When calculating the average age, the compiler had focused entirely on that portion of the subscriber file for which Age had been successfully appended. By doing this, the compiler had implicitly assumed that those individuals for whom Age was not known had the identical profile.

Individuals who can be coded with a given data element are almost always demographically different from those who cannot. This is because representation on major overlay databases is skewed towards older, more stable individuals. The explanation lies with the two reasons for not being codeable:

- There has been a change of address that is not reflected on the database. Technical reasons contribute to this effect, some of which are related to the NCOA process. Sometimes, it is as simple as the fact that an NCOA form has not been filled out.
- No data exists for the individual. The extent to which an individual has a home, automobile, credit cards, children and the like, is the extent to which he or she is likely to be represented on a given overlay database. Conversely, those who cannot be matched to an overlay database tend to be young renters who move frequently. These people generally also are not affluent and not married.

Generally, the magnitude of the swing between reported and actual age is directly proportional to the degree to which the percentage of records with missing data deviates from the norm.

Getting back to our publishing example, the 20% of the file that was appended with age did, indeed, have an average age of 44. However, the other 80% averaged just 26.5. This can be seen by solving a simple simultaneous equation, as follows:

- $(20\% \times 44) + (80\% \times \text{Unknown Age}) = 30$.
- $8.8 + (80\% \times \text{Unknown Age}) = 30$.
- $80\% \times \text{Unknown Age} = 21.2$.
- $\text{Unknown Age} = 21.2 \text{ divided by } 80\%$.
- $\text{Unknown Age} = 26.5$.

In other words, the vast majority of the magazine's active subscribers were in their 20's. However, the overlay process had entirely missed this core target market. Imagine the problems that would have resulted had the publisher begun to focus on acquiring 44-year olds!

Admittedly, a fourteen-year swing between an adjusted and unadjusted data element is the exception rather than the rule. However, swings of between four to six years are very common.

Fortunately, techniques exist to adjust demographic and lifestyle profiles for the systematic bias that is inherent in missing data. Although space does not exist in this article to discuss these computationally intensive processes, it is important for direct marketers to know that they exist.

Another Example

Sometimes, more can be derived from the patterns of missing data than from the actual values themselves. Consider the following table that was constructed from profiles run for a well-known company off a large prospect mailing:

Percent of Records with Missing Data

Data Element	Converters	Non-Converters
Age	73%	60%
Income	50%	40%
Marital Status	75%	68%
Presence of Children	51%	41%
Length of Residence	51%	41%

The rate of missing data among the converters (i.e., the target audience) is unusually high. This is true on an absolute basis as well as when compared with the non-converters. It suggests that the target audience is comprised of young, mobile, lower-income individuals. In fact, the high rate of missing data among the non-converters reflects the fact that the entire prospect universe is younger, more mobile, and has a lower income than usual.

Final Thoughts

In the presence of missing data, how can direct marketers draw sound conclusions about the demographic composition of their customers, inquirers and prospects? This will be the subject of a future article. We will outline how to properly evaluate profile reports, and how to determine the quality of the ones supplied by your compiler. Also, we will focus on the questions that you, as an educated consumer, should be asking.

Jim Wheaton is a Principal at Wheaton Group, which specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. He is also co-founder of Data University. Jim can be reached at 919-969-8859, or jim.wheaton@wheatongroup.com.