

Unleashing the Power of A Demographic & Psychographic Database Enhancement

*By Jim Wheaton
Principal, Wheaton Group*

Original version of an article that appeared in the October 1998 issue of “Target Marketing”

A demographic and psychographic enhancement of your customer or prospect database can improve dramatically your target marketing program. This is because a major part of targeting is determining what to promote. You must tailor the message to the life-stage and needs of each individual. And, an important tool for understanding life-stage and needs is third-party demographic and psychographic overlay data.

Most marketers believe that a database enhancement is a simple process. After all, how hard can it be to overlay demographic and psychographic variables and then interpret their averages and distributions?

Unfortunately, the answer is that it's quite a bit harder than you might think! And, that's the reason for this article: to provide you with a step-by guide for successfully incorporating demographic and psychographic overlay data into your target marketing program.

Reasons for Enhancing a Customer or Prospect Database

There are four major reasons for enhancing a database with overlay demographics and psychographics:

- To create profiles. Averages and distributions can be computed for a broad range of variables such as “age,” “income,” “marital status,” and “presence of children.” Feed this information to a good creative staff and the result will be interesting insights into how to tailor the promotional message to the characteristics of each individual.
- To generate segments, which divide a database into groups of identical, or “homogeneous,” individuals (or households). Many companies create a manageable number of homogeneous life-stage segments and then target them with customized promotions. Consider a few of the permutations that can be created from our aforementioned “age,” “income,” “marital status,” and “presence of children” variables:
 - Young, affluent singles.
 - Young, middle-class parents.
 - Middle-aged, affluent parents.
 - Older, middle-class “empty nesters.”

- As input to statistics-based predictive models, which methodically interrogate multiple variables to predict future behavior. Examples of such behavior are response rate and purchase volume.
- To increase the “rentability” of a customer database, especially for non-hotline names. In this application, enhancement data is used to identify pockets of customers with increased responsiveness to the renting company’s offer.

The Importance of Name and Address Hygiene

Name and address hygiene are critical to an optimal enhancement. No matter how good the overlay data, it’s worthless if the customer or prospect database displays such poor hygiene that the matching algorithm cannot function properly.

Hygiene problems take two forms. The first occurs when a database record does not contain an up-to-date address. The second takes place when a record, although reflecting the current residence, contains garbled or incorrect address elements such as Street Numeric or ZIP Code. Both types of hygiene problems depress the percentage of records that can be appended with enhancement data and contribute to deliverability problems.

Multiple Methods of Compilation

In order to properly interpret and use enhancement data, it’s important to understand the forms in which it is offered for sale. The following are the possibilities, along with some caveats:

Aggregated data:

- Census-based geographic units can be purchased at the Census Tract and Block Group levels. There were 49,961 Census Tracts and 225,876 Block Groups in the 1990 Census. Most of this data is derived from the Census statistics itself, although some enhancement companies offer elements that have been compiled from proprietary sources.

As the unit of data aggregation gets larger, the variation of the variable values tends to narrow. Therefore, there will be a smaller percentage of very high, and very low, values. This will affect the criteria for any select. Consider a promotion targeted to neighborhoods with high “median household income.” There will be a higher percentage of Blocks Groups than there will be Census Tracts with “income” greater than or equal to – say – \$80,000.

- Postal-based geographic units are offered at the ZIP Code, Carrier Route and ZIP+4 levels. As of the mid-1990’s, there were about 43,000 ZIP Codes, 570,000 Carrier Routes and 29 million ZIP+4 Codes. Most of this data is also based on the Census statistics. In such instances, the enhancement companies use proprietary algorithms to translate Census-based geographic units into postal-based units.

Non-Aggregated data:

- Household-level data comprises the majority of non-aggregated data. “Income,” “marital status,” “presence of children,” and psychographics are examples of variables that generally are compiled at this level.

Such elements are particularly easy to misinterpret. Assume that I am interested in automobiles but my wife is not. Unfortunately, my wife will be tagged with an “automotive interest” during a database enhancement. Understanding this phenomenon can mean the difference between a smart and a not-so-smart marketing move. “Interest in automobiles” once “popped” on the customer database of a well-known woman’s magazine. It would not have been a good move for the editorial staff to add content on the rebuilding of a fuel injection system!

- Individual-level data, which often is limited to “gender” and “age.”

We have already noted that, as the unit of data aggregation gets larger, the variation of the variable values tends to narrow. The same phenomenon exists for any aggregated data element that is compared with its individual- or household-level equivalent.

Consider a company that wants to target up-scale households. And say that it selects all households with incomes of at least \$100,000. The problem is that the *income* variable will not be able to be applied to the entire database. (More later on the inevitable problem of missing data.)

For the uncoded portion of the database, it would be unwise to default to households with a ZIP Code-level “income” value of at least \$100,000. This is because there are not many such households. A better strategy might be to make the selection at the Carrier Route level, drop the median to – say – \$80,000, and pull in other data elements to further qualify the select.

- Estimated data, for both household- and individual-level elements. “Income” and “wealth” are perhaps the most widely used.

Each of the enhancement companies has its own estimation algorithms. “Income,” for example, might be driven by information such as the type, number and age of the automobiles owned. Block-Group-level Census information, such as “median income” and “median house value,” are also common inputs.

One must use caution when mixing and matching the same estimated data element from several companies. For any one source, the estimates will tend to be directionally accurate. Across sources, however, one often is systematically different from another.

Finally, it’s important to keep in mind that estimated data elements will be less accurate than non-estimated elements.

Deciding on the Enhancement Variables

There are hundreds of demographic and psychographic data elements on the market. It's important to decide which ones to purchase. The answer is simple: test as many as possible! For predictive models and segmentation systems, append as many of the available elements to the analysis file as is practical. Then, enhance the full database with only those which survive to the final algorithm.

The Problem of Missing Data

Unfortunately, overlay data cannot be applied to every record on a database. For many data elements, the enhancement rate is well under fifty percent. Many marketers do not realize that the demographics and psychographics of codeable and uncodeable records are fundamentally different. The explanation lies with the two reasons for uncodeability:

- No data exists on the household. Often, this is the case for households that do not own items such as homes, automobiles and credit cards.
- There has been a change of address that is not reflected on the enhancement company's overlay database. There are several reasons for this, some of them technical issues relating to the NCOA process. Sometimes, it's as simple as the fact that an NCOA form has not been filled out.

The individuals with the greatest chance of not owning items such as homes, automobiles and credit cards tend to be young renters. These people generally are single and not affluent. Such individuals are also the ones who most often change addresses.

This "missing data bias" towards the young, the renters, the single, and the non-affluent must be taken into consideration when interpreting demographic and psychographic profiles. After all, profiles – by definition – do not reflect this "missing data" portion of a customer or prospect database.

Fortunately, statistical techniques exist to adjust for this bias. One example is a well-known magazine that enhanced its database with "age." It found a fourteen-year difference between the adjusted and unadjusted average age of its customers. Although this is an extreme case, swings of between four to six years are common.

Correct Interpretation of Multiple Overlay Variables

Another common error is to assume that the enhancement variables that "pop" on a database describe the same group of individuals. Assume, for example, that the following characteristics are over-represented: "young, married, affluent, and male." It would be hazardous to conclude that the target

audience is young, married, affluent males. There just as likely could exist multiple audiences, such as:

- Young (single) males (of various incomes)
- Affluent couples (of various ages).

This distinction has profound marketing implications. Fortunately, multivariate statistical techniques such as Tree Analysis are available which have the power to identify situations in which multiple target audiences exist.

Conclusion

Demographic and psychographic enhancement data can dramatically improve a target marketing program because this information is critical to determining what to promote. Specifically, such data can be used to profile a database, create homogeneous “life-stage” segments, assist in building powerful predictive models, and increase the “rentability” of customers.

However, the overlay and interpretation of demographic and psychographic data is not as simple as it might seem. The customer or prospect database must display excellent name and address hygiene. One must be mindful of the multiple methods of compilation, namely Census and ZIP-based aggregated data, as well as household and individual-level non-aggregated data. Also, testing must be done to determine the appropriate enhancement variables to use. And finally, it’s important to be mindful of “missing data” bias as well how to correctly interpret multiple variables.

To sum it all up, it’s best to tap into the knowledge of a data enhancement professional when attempting to unleash the power of demographic and psychographic overlay elements.

Jim Wheaton is a Principal at Wheaton Group, and can be reached at 919-969-8859 or jim.wheaton@wheatongroup.com. The firm specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. Jim is also a Co-Founder of Data University www.datauniversity.org.