

Individual/Household Demographics & Psychographics: Applications in Descriptive & Predictive Research

*By Jim Wheaton
Principal, Wheaton Group*

Original version of an article that appeared “The DMA’s 1997 Research Council Journal”

Individual and household overlay data often play major roles in descriptive as well as predictive research. But the careless use of this demographic and lifestyle information can result in more harm than good when it comes to making accurate marketing decisions. However, by adhering to certain methods of incorporating overlay data into research and by properly interpreting the results, many of these common errors can be avoided.

Descriptive Research Applications

Handling Missing Data

One common error in descriptive research applications results from the fact that individual and household overlay data invariably cannot be applied to a significant percentage of a given file. The portion for which a specific data element cannot be applied generally ranges from 20% to 95%. Therefore, whenever marketing decisions are based on a given demographic or lifestyle variable, often it is implicitly assumed that those individuals for whom data coverage does not exist have the identical profile.

Consider a file in which the average age of the codeable records is 44. Any marketing decisions that result from this information will be appropriate only if the uncodeable portion of the file also has an average age that approximates 44.

Unfortunately, uncodeable individuals almost always are demographically different from the codeable, because representation on major overlay databases is skewed towards older, more stable individuals. I call this the “Ozzie and Harriet factor”:

The extent to which an individual has a mortgage, children, credit cards, and the like, is the extent to which this individual is likely to be represented on a given overlay database. Conversely, those individuals 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.

Let's get back to our example, in which the average age of the codeable records is 44. This is exactly what happened to the client of a major data compiler who – finding this average to be counter-intuitive – sought a second opinion. Fortunately, techniques exist to adjust demographic and lifestyle profiles for the systematic bias that is inherent in missing data. Application of one such adjustment algorithm shifted the average age of the file from 44 to 30. This lower estimate agreed exactly with the client's “gut instinct,” as well as with extensive survey research.

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.

Generally, the magnitude of the swing will be directly proportional to the degree to which the percentage of records with missing data deviates from the norm. Ask your overlay data provider to share with you the typical hit rates for each element. If, for a given element, your file experiences a significantly higher rate than average, then you can be sure that the swing between your adjusted and unadjusted readings will be extreme.

Sometimes, one can derive more from the patterns of missing data than from the actual values themselves. Consider the following table that was derived from profiles that were 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 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 lower income than usual.

Problems with Profiling Local Markets

Marketers often run profiles of multiple local markets. A national retailer, for example, might want to compare and contrast its Los Angeles, California and Boulder, Colorado store customers. As noted above, a demographic and psychographic profile of these two customer groups will – by definition – include only those households for whom overlay data is available. And, the makeup of those customers for whom overlay data cannot be applied is very likely to be quite different in Los Angeles than in Boulder:

- For Los Angeles, these individuals are likely to be ethnically diverse. And, they might well be disproportionately represented by unskilled Mexican citizens who have fled their country's oppressive unemployment.

- For Boulder, these individuals are likely to be ethnically homogeneous. Also, because of its very large University of Colorado campus, Boulder has a huge reservoir of students and recent graduates. These future “Ozzie and Harriets” are, of course, highly educated.

Unfortunately, the algorithms that adjust for the systematic bias inherent in missing data for national files will have limited value in our Los Angeles-Boulder comparison. This is because they invariably work off the implicit assumption that the files are national in character. The more that a given file deviates from this ideal, the less accurate the adjustment. For local markets, these algorithms will have to be manually adjusted on a case-by-case basis.

Hazards of “Marketing to the Mean”

Another frequent mistake is what I'll call “marketing to the mean.” It is critical to look beyond the average to the distributions of a given variable. A real life example is a direct marketer that focuses on high-end merchandise. This company's average customer has an adjusted age of 36. In actuality, however, there exist two distinctly different audiences:

- New-to-the-workforce 18 to 22 year olds with modest salaries but even more modest financial obligations. This group's discretionary income is not compromised by monthly house payments and the like, especially for those who are still living with their parents.
- Affluent consumers in their late-40s to mid-50s whose rising incomes have finally outstripped their obligations. For these individuals, there remains enough money at the end of the month for the enjoyment of luxury items.

In fact, individuals who are the average age of 36 are very poor prospects because many are parents with mortgages – and looming college tuition payments – who have little discretionary power for high-priced merchandise.

To present a balanced perspective on this topic, it is more common for the adjusted mean to represent the primary audience. In such instances, the problem with “marketing to the mean” is that important secondary audiences are not identified. However, marketing to the unadjusted mean is a double whammy because both the primary as well as the secondary target audiences are missed.

Correct Interpretation of Multiple Overlay Variables:

Another common error is the assumption that the demographic and lifestyle overlay variables that stand out or “pop” on a file all describe the same group of individuals. Assume, for example, that the following characteristics are over-represented on a file of diamond ring buyers: “young,” “male,” “affluent” and “married.” It could be hazardous to conclude that the target audience is young, affluent, married males. There just as likely could exist multiple audiences, such as:

- Young (single) males (of various income levels) who purchased an engagement ring.
- Affluent couples (of various ages) who bought a ring to commemorate an important wedding anniversary.

This distinction has profound marketing implications. Fortunately, multivariate statistical techniques such as CHAID (Chi-Square Automatic Interaction Detection) are available, which have the power to identify situations in which multiple target audiences exist.

CHAID is a member of a broader category of statistical techniques called “tree analysis.” Another well-known member is CART (Classification and Regression Trees). Tree analysis is a wonderfully-insightful profiling tool because of its ability to segment a given universe into multiple, homogeneous segments (e.g., the diamond ring example above). By definition, each homogeneous segment is a potential target market. This is different from regression analysis, which creates heterogeneous segments that can be more difficult for marketers to understand. With regression, the inhabitants of a given segment have only one guaranteed similarity: their predicted future purchase patterns.

Predictive Research Applications

Problems with Static Data

A frequent problem when overlay demographics are incorporated into predictive models is that static data – those sources that are purchased outright and are not periodically updated – change meaning over time. This occurs because of the large percentage of individuals who move every year. This, in turn, results in an ever-increasing overlay rate for older, more stable people compared with their younger, more mobile counterparts.

A live example of this phenomenon is a regression model in which two “political affiliation” overlay variables, “conservative” and “liberal,” both “popped” positively. The reason is that the variables were several years old and rapidly were becoming surrogates for the target audience – stable individuals in their 40’s and 50’s.

Hazards of Short-Term Data Fluctuations

Unfortunately, even non-static data can change meaning over time. An excellent example is “length of residence,” a common overlay variable. Because of peculiarities in the update cycle of at least one major data compiler, for three months every year essentially no one on its file shows a “length of residence” of less than one year. In the absence of an adjustment to reflect this phenomenon, this would be problematic for a model developed for “new mover” merchandise such as window treatments!

Predictive Power of Missing Data

Many statisticians are unaware of the often remarkable explanatory power that is inherent in missing data. Sometimes, for a given individual, the inability to apply specific demographic or psychographic information is more predictive than the information itself. This has to do with the “missing data bias” discussed earlier.

As an example, let's revisit the “length of residence” variable, which is created in significant part by comparing names at specific addresses in phone directories from one year to another. Besides the usual problem of younger, mobile individuals having lower hit rates, we have additional bias because of those demographic groups that have a higher probability of opting for an unlisted telephone number. With the unlisted-number group, it is very likely that the information required to calculate “length of residence” cannot be obtained. These people generally fall into one of the following categories:

- Single women, urban residents, and the very affluent (who opt for unlisted numbers for security reasons).
- The very poor (who cannot afford phones).

Therefore, the absence of “length of residence” information increases the probability that a given individual belongs to one or more of the groups listed above. This might very well be more predictive than the knowledge that a given individual has – for example – resided at his or her address for three years.

In order to capture the predictive power of missing demographic and psychographic information, it is critical that missing data for a given predictor variable be assigned its own value when building a model. This is contrary to the practice of many statisticians, who set missing data to the mean of all the observations for which information exists. Others default to the equivalent Census-level variable, which is an improvement but still not optimal.

A wonderful example of the potential predictive power of missing data is what I refer to as “The Unmodel,” which was constructed to segment several large outside rental lists. The top decile was driven largely by the absence of information on multiple overlay elements. This is because the target audience was comprised of “un-Ozzie and Harriets”; that is, single, downscale, renters of apartment units.

Consider the univariate relationship to response of several “Unmodel” predictor variables, where the “Missing” categories all correlate very highly with response:

Response Rate by Income

Income Missing	LT \$15 M	\$15M-\$20M	\$20M-\$40M	\$40M-\$50M	\$50M-\$75M	\$75M-\$100M	\$100M-\$125M	GT \$125M
1.21	1.04	0.87	0.79	0.72	0.68	0.66	0.65	0.62

Response Rate by Age

Age Missing	18-25	26-29	30-35	36-39	40-45	46-49	50-59	60-69	70-75	GT \$125M
1.21	0.81	0.62	0.52	0.57	0.68	0.77	0.92	0.97	0.87	0.78

Response Rate by Credit Card

Card Missing	No	Yes
1.33%	0.98%	0.69%

The resulting performance was quite good for a prospecting model, with “lift” – top 10% to average – of 209 (and “lift” – top 10% to bottom 10% – of 475).

“For the sake of science,” a test was run on this data set to measure the percentage of the model’s segmentation power that was contributed by missing data relative to “actual data.” For each of the independent (i.e., “predictor” variables), the values were collapsed to “yes/no” binaries that correspond to the presence or absence of data.

Consider, for example, the “Income Missing” category from the preceding illustration. All eight of the “valid values” categories corresponding to “LT \$15M” to “GT \$125M” were collapsed to a “1,” meaning “Income Present.” The “Income Missing” category became a “0.” Remarkably, the resulting model displayed segmentation power – “lift” – that was two-thirds that of the actual model put into production. (As an aside, this highlights the folly of setting missing values to the mean when constructing predictive models.)

Problems with Increased Data Coverage

Predictive models work best when the distribution of the values for a given variable, including the percentage with missing values, remains constant over time. From a long-term perspective, it is to

everyone's advantage when a data compiler increases its coverage of a given variable. Unfortunately, in the short-term this jeopardizes the effectiveness of established models built during the time of lower coverage.

Not so long ago, a major data compiler increased by 50 percent the number of households for which it had “presence of children” information. It did so by tapping into a significant additional primary source. Consider the effect of this development on an existing predictive model:

We've already established that missing data for “presence of children” should be assigned its own value when building a model. However, with the dramatic increase in the “hit rate” for this variable, the percentage of records with a value of missing will significantly decline.

This is not necessarily a problem if the overall demographic profile of the incremental source is identical to the base, original source. But what if the new source represents parents who are, say, significantly older and more affluent? The unfortunate answer is that the model's segmentation power is likely to be compromised.

Financial Evaluation of Overlay Data

All of this leads to the question of whether or not to include individual/household overlay data in a model build. The short answer is that this type of data should be used only if its incremental contribution to segmentation power is greater than its incremental cost:

For customer models in which transaction information is available, the financial evaluation should be done off three versions of the model:

- Transaction-only predictor variables.
- The above, plus aggregate-level (e.g., Block Group) demographic variables; that is, inexpensive overlay data with relatively modest incremental power.
- The above, plus individual/household-level demographics; that is, relatively expensive overlay data with potentially substantial incremental power.

For prospect models, where transaction information is – by definition – unavailable, the financial evaluation requires only model versions #2 and #3.

And finally, when performing this incremental financial evaluation, remember that the volatility of individual/household-level data has the potential to cause the premature degradation of a model. This is a difficult-to-quantify hidden cost that can be counteracted only through more frequent model builds.

Conclusion

The use of overlay data can have a powerful impact on direct marketing research, if applied properly. To ensure the effective incorporation of overlay data and the correct interpretation of results, there are several rules to keep in mind.

First, for descriptive research, demographic and lifestyle profiles must be adjusted to reflect the “Ozzie and Harriet” bias that is inherent in major overlay databases. Be especially mindful of missing data's ability to produce misleading results when profiling and comparing local markets. It is also important to consider profile distributions rather than means when drawing marketing conclusions. And, marketing to an unadjusted mean can cause additional problems. Finally, never assume that multiple overlay variables that “pop” on a file all describe the same group of individuals.

For predictive research, incorporate static data into models with caution, recognizing that their meanings will change over time as they become surrogates for older, more stable individuals. Also, be mindful of the fact that even non-static data can change meaning as suppliers update their databases. And, recognize that increased data coverage generally will degrade an existing model. Additionally, recognize and take advantage of the fact that missing data often can provide remarkable explanatory power. Finally, incorporate individual/household data into a model only if its benefits outweigh both its obvious as well as its hidden costs.

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.