

## **Common Errors in the Use of Overlay Data**

*By Jim Wheaton  
Principal, Wheaton Group*

*Original version of an article that appeared in the September 12, 1994 issue of "DM News"*

Individual and household overlay data often play major roles in descriptive as well as predictive research. But the careless use of this 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 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.

## 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 well-known fashion magazine whose average adjusted subscriber age is 36, but who in actuality has two target audiences:

- New-to-the-workforce 18 to 22 year old women who view the magazine as a “wish book.
- Affluent women in their late-40s to mid-50s who reference the magazine when making purchase decisions.

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

## 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) have the power to identify situations in which multiple target audiences exist.

## Predictive Research Applications

### Problems with Static Data

A frequent problem when overlay demographics are incorporated into predictive models is that static data – sources purchased outright and 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 real-life 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, for example, has 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 missing data’s potential predictive power 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” – single, downscale renters of apartment units.

Consider, for example, 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 Credit Card*

Card Missing	No	Yes
1.33%	0.98%	0.69%

*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

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).

**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. It is also important to consider

profile distributions rather than means when drawing marketing conclusions. And 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 finally, recognize and take advantage of the fact that missing data often can provide remarkable explanatory power.

*Jim Wheaton is a Principal at Wheaton Group, and can be reached at 919-969-8859 or [jim.wheaton@wheatongroup.com](mailto: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](http://www.datauniversity.org).*