

## Selecting the Right Predictive Model

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

*Original Version of an article that appeared in the July 7, 2003 issue of DM News*

Prospect name predictive models come in many flavors. “Stimulus/response” models, for example, are built directly off promotional files and corresponding responder information. “Look-alike or “clone” models, on the other hand, are not based on promotional information. Instead, they evaluate how similar in appearance each prospect is to a direct marketer’s customers, in an indirect attempt to capture the dynamics of response behavior.

All other things being equal, stimulus/response models are superior to look-alike models. In other words, if robust promotional and responder information is available, then it should be used! However, look-alike models have their place, especially in start-up situations such as when a new channel is being incorporated into the sales process.

### Differences in Granularity

For both stimulus/response and look-alike models, there are differences in the level of demographic overlay data used to create the independent or “predictor” variables. The following are several such levels, in order of increasing granularity:

Unit	Quantity	Comment
ZIP Code	30,000	Residential-only, without Commercial ZIP’s
Block Group	225,000	1990’s number
Carrier Route	570,000	1990’s number
ZIP+4	29 million	1990’s number
Individual/Household	100 <sup>+</sup> million	Overlay data not available on all records

Generally, the more granular the data, the more predictive it is. Therefore, all other things being equal, the most granular available data should be used to build a model. However, nothing is ever equal within the realm of prospect modeling!

### The Complication of Missing Data

An important issue that exists with Individual/Household data is significant gaps in coverage. Except for Age, Length of Residence, and Estimated Income, it is rare for coverage to exceed about 75%. For self-reported Income, coverage generally is about 35%. With Estimated Income, the high

coverage is only possible because this element is – in its own right – the result of a predictive model. Information such as car ownership and neighborhood Census characteristics is interrogated to create these estimates.

### **Measuring Incremental Power Versus Incremental Cost**

Cost is also a factor. Generally, the more granular the data, the higher its price. Therefore, a financial evaluation should be performed for each level of granularity, in order to determine if its incremental predictive power more than offsets its incremental cost. The financial evaluation compares multiple versions of the model. Assume, for example, that the available levels of overlay data are Block Group and Individual/Household:

- First, a model would be built off only Block Group-level demographic variables; that is, inexpensive data with relatively modest incremental power.
- Then, a second model would be constructed off the above plus Individual/Household-level demographics; that is, relatively expensive overlay data would be added, with potentially significant incremental power.

This approach determines if any Individual/Household data elements are cost effective. And, if it makes sense to incorporate such elements in the final model, the total number will be minimized. This, in turn, will result in the most economical possible model.

### **The Hidden Cost of Volatility**

When performing an incremental financial evaluation, one must be mindful of the potential volatility of highly granular sources such as Individual/Household and ZIP+4-level data. Volatility often results in premature model degradation. This is a difficult-to-quantify hidden cost that can only be counteracted by more frequent model builds. Individual/Household data is a particular challenge.

Volatility occurs because of changes in the underlying data sources. Sometimes, a data compiler will replace one or more original sources, either in whole or in part. Other times, a source will be pulled off the market because of legislation, privacy concerns, or other reasons. This has been a more frequent occurrence over the past few years. Periodically, new sources will come on the market.

From a long-term perspective, it is to everyone's advantage when coverage increases. In the short-term, however, increased coverage jeopardizes the effectiveness of established models. This is because models work best when the distribution of the values associated with the predictor variables remain constant over time.

### **Net/Net Rental Arrangements**

In the presence of net/net rental arrangements, direct marketers pay for only the names that they mail. Under such circumstances, the gross names received from a list manager can be run through a statistics-based predictive model. Then, those names with relatively low scores can be returned

without being mailed. The only financial obligation to the direct marketer is a run charge of about \$6 per thousand.

In the absence of net/net arrangements, the list owner will insist on compensation for a significant portion of the names that are processed, whether or not they are mailed. Traditionally, the industry standard has been that at least 85% of the rented names must be paid for, although more favorable percentages are common.

Without the advantage of net/net rental arrangements, it is very difficult for a model to overcome the cost of paying for discarded names. Consider a list arrangement where 50,000 names are obtained at \$100/M on an 85% net basis. Assume that these names are then screened with a model that results in only 10,000 being mailed. In that case, their effective cost per thousand is \$425; that is, 50 times 85% times \$100, divided by 10. And, that does not include overlay data and processing costs!

### **The Role of ZIP Code Models**

Unfortunately, for all but the largest mailers, it is difficult to obtain net/net arrangements. Hence, the widespread use of ZIP Code-level prospect models. This is because their output is a list of ZIP Codes to be used for selection or omission. It is very easy for list managers to process such a list, and with a very nominal “run charge.”

The downside of ZIP Code models is that they generally display very modest predictive power. The sheer math explains why. Census data is available for about 30,000 ZIP Codes, spread across about 100 million mailable U.S. households. Therefore, the average ZIP Code is comprised of about 3,300 households. Imagine how difficult it is to predict behavior based on the makeup the 3,300 nearest residences!

Another problem with ZIP Code models is that they often perform erratically across lists. It is common for ZIP models to be effective on an overall basis, but fall apart within some portion of individual lists or list types. The reason is what is known as “self-selection bias.”

Consider, for example, a cataloger that sells business-appropriate attire to upscale professional women. A ZIP model indicated – not surprisingly – that working class ZIP Codes are not attractive targets. Nevertheless, for a sizeable minority of lists, the response rate within such ZIP Codes was significantly higher than what was predicted by the model. An analysis revealed that the women on these lists who live in working class ZIP Codes are almost always upscale. The following example provides insight as to why:

One of the lists in question was Neiman Marcus catalog buyers, which is comprised of very few working class women. Therefore, any given Neiman Marcus customer is likely to be upscale, even if the average household within her ZIP Code is not. Such mismatches are common in urban and semi-urban areas, especially those that have been labeled “transition neighborhoods.”

*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, which helps companies understand, integrate and leverage different types of data. Jim can be reached at 919-969-8859, or [jim.wheaton@wheatongroup.com](mailto:jim.wheaton@wheatongroup.com).*