



## Predictability 2.0: Refining Your Householding Algorithm

by Greg Martin

Column published in [DMReview.com](http://DMReview.com)  
January 11, 2007

Any company about to update their householding (or customerholding) process should become familiar with the opportunities as well as the pitfalls before undertaking this process.

Depending on your business model and industry, your company may need a great customerholding process, but may not need a householding process or vice versa; however, you can certainly work with both. Because revising a complex matching algorithm can be applied generically to either process, I will use the term householding to refer to both in this article.

To begin with you will need three things: first, the tool and platform for householding; second, a statistical tool for doing random selections; and third, you will need staff with the available time to run the matching, conduct sampling, build reports and inspect outcomes. If this process needs to be accomplished on a tight deadline and your company does not have significant staff availability, then you should consider outsourcing.

The basic idea here is to apply the classic champion-challenger approach to the selection algorithm. This method will help you to avoid making a common mistake, which is to simply have someone select and inspect households by hand. If done manually, you will not know whether you are solving systematic problems (which is what you want to do), or a unique data issue, which may or may not create a better overall solution.

Throughout this process, it is important to keep in mind how the results will be used within the company. Consider the sponsoring area, its current pain points and any legal ramifications. Engage the internal constituencies (direct marketing for data usage and front line sales for data input) as early as possible in the refinement process and ensure they know how to share future change requests to the matching process.

Some additional considerations to keep in mind before undertaking this process include making sure the sampling tool can handle volume of data, having a tool to view the output datasets, ensuring your matching tool allows for exclusion files and determining whether you need a separate business matching routine.

### Steps in the Testing Process

The first step in this approach is to run both algorithms; the outcome of which is to attach the household identifier (HH ID) to each input record.

At this point, you should have two data sets that are identical, except for the HH ID attached to each input record. You should then merge these two datasets; the results of which will be one dataset with two different HH IDs at the end of each record. You can now generate your first result metrics. These are:

- ▶ Number of households that were the same,
- ▶ Number which merged, and
- ▶ Number that split

Your focus from this point forward should be on the changed households.

Use a statistical sampling method to select 50 households that merged and 50 households that split. Inspect each of these households and tally good versus bad. The fields you will be using are the ones that go into the matching algorithm, e.g., name, address, SSN, driver's license, etc. Some matching tools will have codes indicating why two records were matched, which can be helpful especially in trying to figure out why records ended up in a large household. (It is sometimes not obvious why the matches have occurred when there is a chain of events matching record A with record B and record B with record C. If you look at just A and C, it may not make immediate sense). If there are multiple records, this will take significant effort.

The final result is worth working for. If you have better results with the new algorithm, then you have a new champion. Remember that this is about trade-offs, not perfection.

At this point the algorithm can be further tweaked to either make it a tighter match which will result in fewer merges, but more splits - that is, a lower cross-sell ratio. Or, make it a looser match - fewer households with a higher cross-sell ratio. You should also inspect your largest households, where you can usually find fertile ground for exclusion data.

Now that you have a better household algorithm in place, you can turn your attention to cross-sell, tenure, head of household, best address or even relationship profitability and other related metrics that can help drive your company's success.

...

Greg Martin, vice president-West Coast, Quaero, is the business technology linkage expert in implementing large-scale data warehouse projects with in-depth knowledge in the financial services industry. He welcomes any questions or comments on this article and can be reached at [marting@quaero.com](mailto:marting@quaero.com).