Sawtooth Software

RESEARCH PAPER SERIES

Segmentation – How to Do It Badly and Well

Keith Chrzan
Sawtooth Software, Inc.

3210 N. Canyon Rd., Provo, Utah
+1 801 477 4700
www.sawtoothsoftware.com

Segmentation – How to Do It Badly and Well

Keith Chrzan, Sawtooth Software

Segmentation helps marketers understand how groups of customers differ with respect to the products, messaging or positioning that appeal to them. Understanding these differences gives marketers more leverage in designing or selling products to their customers.

Besides choice modeling, segmentation is the most common thing we do for clients of Sawtooth Software's analytical consulting division. I aim in these four pages to distill learnings from 35 years of experience with hundreds of segmentation studies: what factors make segmentation difficult and what we can do to counteract them and succeed anyway.

This paper concerns unsupervised segmentation (e.g., cluster analysis and model-based/latent class clustering). Other types of segmentation we cover in the previous article, Four Common Types of Segmentation Analysis.

What Makes Segmentation Difficult

Three kinds of combinatorial complexity and one bit of nebulousness make segmentation studies about the most difficult kind of studies marketing researchers perform:

Combinatorial complexity #1 – the number of possible ways of partitioning data

Imagine you have a small (n=150) segmentation study and through a stroke of luck (maybe it came to you in a dream, or a genie in a bottle told you or whatever) you happen to know for a fact that your data contains four segments. You're all set, right? Wrong. It turns out there are more than 2 x 10^{90} ways of dividing 150 respondents into four segments. That's more than the number of atoms in the universe (about 10^{82}) - 500 million times bigger, in fact. There is just no way you can look at all those solutions to decide which is best between now and Tuesday when your report is due. Uh-oh.

Combinatorial complexity #2 – choosing basis variables

Your survey contained 100 questions. Which of those will you use as basis variables (i.e. as the inputs to the segmentation) and which will be the profiling variables (the variables used to characterize but not define segments)? Each question can either become a basis variable or not, so we face another decision, one with 2^{100} (or 1.26 x 10^{30}) possible sets of basis variables (that's about a trillion times as many sets as there are gains of sand on the earth).

Combinatorial complexity #3 – segmentation algorithms

Similarly, there are different ways of pre-treating your segmentation data, different segmentation algorithms and different ways of measuring similarities and differences among respondents, making for thousands of possible ways to run your segmentation analysis.

Nebulousness

Finally, in physical sciences like biology, segmentation uses measurable physical characteristics (e.g., genes) to discern how different kinds of conifers, salmon or bacteria group together. Those plants or animals or bacteria really do have natural groupings, and segmentation needs only to find them. In marketing, however, our human subjects' differences reside in their beliefs or attitudes, perhaps as manifested by their behaviors. Seeking to create groups of customers whose primary differences are psychological, marketing researchers have to deal with the limitations and biases of survey research.

Advice for Successful Segmentation

Despite the four factors working against us, we can still successfully execute a segmentation study. The steps below can help.

Don't turn segmentation into a black box

More than in most other kinds of analysis, segmentation analysts will make many methodological decisions in the course of a segmentation. Having client buy-in is critical for the success of segmentation. I like to make sure clients agree on which variables are basis variables (i.e. which go into the algorithm that makes segments) and which are profiling variables (all the other variables in the study, on which segments can be compared once produced). I like them to know which analysis decisions I'm making and the rationale for those decisions. Rather than deliver a single solution I think is best, I prefer to cull the many solutions I've investigated into 3-5 solutions I think pass statistical muster, and let my client, with her greater knowledge of the market, have the final say.

Determine whether your data are segmentable

You may hear critics of segmentation suggest that market segments are imaginary, that the groups into which we put survey respondents are no more real than if we grouped the randomly-distributed seeds inside of a watermelon. Though sometimes true, it usually is not: I like to reassure myself and my clients by testing whether our data are clumpy or whether our cases really are randomly distributed and incapable of meaningful segmentation. If we segment on just two variables, it would be easy enough to run an XY scatterplot in Excel and let our eyeballs tell us if our data is clumpy. With 20 variables, however, we cannot conduct an eyeball test because our eyes cannot visualize data arrayed in 20 dimensions. We can, however, use the Hopkins statistic, which produces a value ranging from 0.0 to 0.50. In the version of the Hopkins statistic available in R, the Hopkins tends to be lower for data containing more clumpiness, while it moves closer to 0.50 the more evenly our data is distributed. While no particular cutoff divides clumpy from random data, marketing research data sets tend range from about 0.20 to 0.45, and only above 0.40 have I had much trouble segmenting.

Choose a small number of basis variables

Segmentation studies should have at least 100 respondents per basis variable, which limits the number of basis variables in many segmentation studies. And sometimes even if you have 100 or more respondents per basis variable, a problem called the "curse of dimensionality" can still wreak havoc on your segmentation study. In a nutshell, having too many basis variables works against the objectives of a segmentation study: segments tend to be less well-differentiated the more basis variables we use.

A bad way to reduce the number of variables is factor analysis: some folks like to throw every possible variable into a factor analysis, rely on it to identify the variables most related to each resulting dimension/factor and then use only those most-related variables as bases for the segmentation. Factor analysis makes fine sense when testing psychometric theories about which theoretical constructs are real and which are not, and about which variables measure them. But tossing a hodge-podge of variables into it and hoping that factor analysis will identify the best variables for segmentation is a great way to fail. For segmenting, we want measures that will differentiate respondents in ways of marketing relevance. Instead, factor analysis gives us those that share the most variance with other variables – which is not the same thing (often the different factors group by question type – rating scales on one factor, percentages on another and so on). This means that the items that make the most interesting differences across segments may slip through the factor analysis altogether. Using factor analysis to identify basis variables can be the methodological equivalent of an own goal in soccer (or of shooting yourself in the foot, if you're not a soccer fan).

Instead, variable selection methods exist that work well to reduce a candidate set of basis variables down to a smaller set that work better. See the paper by Chrzan and White in the 2022 *Sawtooth Software Conference Proceedings* for a more detailed discussion.

One final note about choosing basis variables: all else being equal, choose basis variables with few or no missing values – not least because the fewer variables have missing values, the more accurately we will be able to predict the segment membership of future survey respondents with a typing tool.

Pretreat data to make it amenable to the segmentation algorithm

Standardization - If our data are measured on different metric scales, we should standardize them. If we have a mix of metric and non-metric variables (e.g., ratings and categorical variables) then standardization isn't going to help and we need to choose a method that accommodates variables measured with different scales (see the "Matching segmentation algorithms to variable types" section below).

Outliers can really wreak havoc on some segmentation methods (e.g., k-means cluster analysis) so eliminate them prior to running your segmentation. Many segmentation packages (including the convergent k-means clustering in Sawtooth Software's CCEA package) have routines built in for finding and eliminating outliers.

Tandem cluster analysis – Some analysts use a "tandem cluster analysis" approach when they have rating scale variables (factor analyze, then cluster on factor scores). Unfortunately, factor scores smooth out some of the very lumpiness in the data that makes cluster analysis work well. It would make more sense to factor analyze and then keep the top loading item on each factor rather than to use the factor scores themselves.

Use robust algorithms

Traditional clustering methods can be unstable, readily finding poor but locally optimal solutions. Because of this well-known instability, avoid running a single cluster analysis in SPSS or SAS, say, and trusting that it will be a good solution. Instead, use robust methods:

- Latent class analysis finds local optima, so running latent class from many sets of random starting points allows us to search more potential solutions and to find solutions closer to the global optimum
- Sawtooth Software's convergent k-means program (CCA, part of the CCEA package) runs up to 30 k-means analyses, with intelligent starting points, for any given number of segments, then searches for the solution that has the most in common with the other solutions (i.e. the solution with the greatest convergent validity) among those searched
- Sawtooth Software's cluster ensembles program (also part of the CCEA package) which runs a "meta clustering" of a large number (by default 70) of diverse cluster analysis solutions
- Robust k-means with very many (e.g., hundreds or thousands of) different sets of starting points followed by a search for the best fitting solution among them
- Partitioning around medoids (k-medians) which is robust to the presence of outliers

For segmentations where all the variables are metric (i.e. rating scales, counts, percentages) you can use PAM (an R package), latent class clustering (available in R as the mclust package and in Latent Gold software), convergent k-means or cluster ensembles (from Sawtooth Software's CCEA package) or robust k-means (available in the R Cluster package).

Finally, when you want to segment on a mix of metric data and ordered or unordered categorical data you can use Latent Gold's latent class clustering (unlike the latent class packages available in R, Latent' Gold's package allows basis variables of mixed scale types). If you want to stay within a cluster analysis framework, PAM also allows mixtures of disparate variable types.

Beware the interaction between sample size and algorithm

Cluster analysis (particularly k-means or convergent k-means) scales up (and down) nicely for large or small data sets. "Small" and "large" data sets might range mean a segmentation study has only 150 respondents (not too uncommon in B2B applications) or 10 million records (e.g., segmenting a database of customers), respectively.

Latent class analysis can start to bog down as sample size gets large (more than a few thousand) and because it is a "model based" method, latent class clustering can quickly run out of degrees of freedom when segmenting small samples of respondents (note that categorical variables exacerbate this problem).

Balance statistical evidence and subject matter expertise when deciding on the number of segments

The fit statistics available for cluster analysis and latent class analysis do a poor job of identifying the true number of segments in a data set.

While cluster analyses support a number of fit statistics (e.g., the silhouette number available in R and other software packages, the reproducibility statistic reported in Sawtooth Software's CCEA, the BIC statistic in latent class analysis, the ensemble of 30 fit indices collected in R's NbClust package) my experience has been that none of these reliably identifies the best number of segments (they seem to fail frequently when you have artificial data sets with known segment membership, a topic covered by Chrzan and White in their 2021 paper in the *Sawtooth Software Conference Proceedings*).

As a result, the informed judgment of people familiar with the market under study should contribute to the decision of the number of segments present in a data set. This subject matter expertise is easily valuable enough to over-ride the weak statistical evidence provided by fit statistics in clustering programs.

Summary

Segmentation analysts face a number of serious methodological challenges and some seriously misleading pseudo-solutions, but methods exist to give the careful analyst the best chance of success in segmentation.