

Applying Archetypal Analysis in Marketing Research

© Paul Riedesel, Action Marketing Research 2008

Archetypal analysis is a mathematical procedure for decomposing a multivariate dataset as a function of a set of underlying archetypes or ideal types. While used for many years in the physical sciences, these methods were first introduced to marketing research practitioners by Louviere and Carson at the 1998 Advanced Research Techniques Forum. Little public use has been made in our field since then.

'Tis the pity. Archetypal analysis is an interesting alternative to the more-familiar methods of cluster analysis used to represent consumer heterogeneity (segments). And it avoids a major fallacy in virtually all segmentation solutions—one we suspect every good researcher is aware of, but few talk about.

This note seeks to elevate the practice of archetypal analysis within marketing research by:

- Explaining its rationale and contrasts to conventional segmentation
- Reviewing the algorithm
- Illustrating its application via a major study of American Baby Boomers
- Discussing practical issues in analyzing and reporting archetypal data

Why Bother with Archetypes?

Our argument is not against cluster-based segmentation, which can be very useful. Rather, our argument is for adding a new tool that can be a respectable alternative to understanding and acting on consumer heterogeneity.

The basic rationale for archetypal consumer analysis and cluster-based segmentation is the same. Consumers (a term meant to include business buyers) differ in their attitudes, needs, behavior, and other characteristics. Because it is usually impractical for a firm to create a unique marketing mix for each consumer, firms rely on various means of grouping consumers. Some segments may be given higher priority. Different products, distribution channels, marketing communications, etc. may be aligned with each segment.

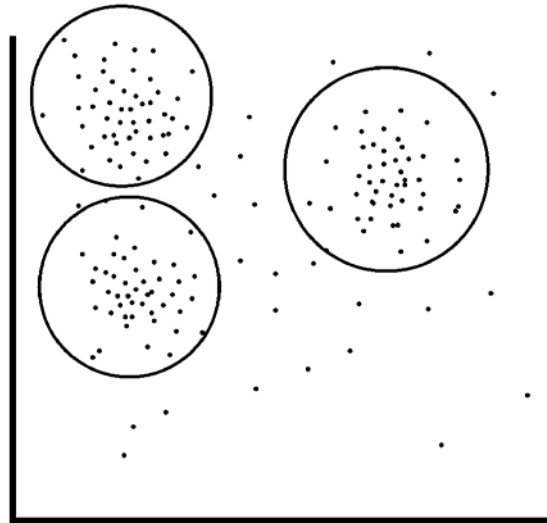
A fundamental problem faced by researchers as they process segmentation data is that the classification of consumers into a few discrete segments is almost always arbitrary to some degree. (Exceptions that come to mind include classification on the basis of criteria such as age, gender, or company size; but these are instances of *a priori* segmentation rather than statistically-derived segmentation).

The statistician performing cluster analysis on market research data knows down deep that the discrete segments he/she will eventually report are useful fictions. We know that the data we typically use are like a vast, multidimensional cloud of points. Yet the client expects each respondent to be placed in one and only one segment. We must therefore carve up that cloud into a small number of regions and call each one a segment.

Consider an example with just two continuous variables. The cases are plotted against the X and Y axes. A three-segment solution is obvious (unusually so). In this illustration, outliers are unclassified, a common but not universal practice.

The classification rules could have been such that every case is assigned to the segment/cluster to whose center it is closest. Or the circles could be larger. Either way, the segments would be that much more heterogeneous (a.k.a. mushy).

This type of classification is not necessarily "wrong." It can be very useful. In fact it can be most gratifying for the researcher when the segments become an integral part of how a company does business. And yet, and yet. The well-tempered researcher is aware that small changes in the classification rules could result in a different story with different marketing implications.



We have had the experience of listening to ordinary consumers react to the presentation of a conventional segmentation scheme. What do they say? While some will admit "That's me, all right", it is at least as common to hear them say "No, I'm some of this and some that; none of those describe me exactly." We may be tempted to write them off as outliers, except there are too many of them. This is not a fatal flaw of standard segmentation procedures in our view, but is another reason to caveat our conclusions (not that clients pay them much heed).

Archetypal analysis, in contrast, keeps us on solid ground. It reflects reality as consumers see it, whether we do or not.

Before taking a short detour into the mathematics, we offer this distinction between cluster analysis-based segmentation and the new tool of archetypal analysis:

- Segmentation makes the implicit assumption that there are several "average" consumers. They are found in the dead center, statistically speaking, of each cluster. The practice is then to create discrete groupings around each one.
- Archetypal analysis assumes instead that there are several "pure" consumers who are on the "edges" of the data. All others are considered to be mixtures of these pure types. Those mixtures and the relative paucity of pure types are very important in the analysis and application.

So how can we use this construct in a systematic way?

Doing the Numbers

The basic mathematics for doing archetypal analysis are most accessible in a paper by Cutler and Breiman published in 1994.

Just like most cluster-based segmentation analysis, archetypal analysis begins with a matrix of data with n respondents and m variables. The algorithm solves several regression equations at once with the object of maximally explaining the variance within the data. The user must specify the number of archetypes to be solved for. If/when a solution is found, we know:

- The archetype weights for each respondent
 - They sum to 1.00 for each respondent
 - If a person has a weight of 1.00 on an archetype, he or she is the archetype
- The relative dominance of each archetype in the sample
 - This is just the mean of all respondents' weights on that archetype
- The degree to which each variable is associated with each archetype

Archetypal analysis uses a form of iterative regression. This summary is based on the formulation by Professor Adele Cutler.

Assume:

- n consumers (survey respondents)
- m variables
- p archetypes to be solved for
- \mathbf{X} is an observed data matrix of dimensions $n*m$
- \mathbf{Z} is a matrix to be solved for of archetypes of dimensions $p*m$
- \mathbf{a} is a matrix to be solved for of mixture weights of dimensions $n*p$
- \mathbf{B} is a matrix to be solved for relating consumers to archetypes of dimensions $p*n$

The object of the procedure is to maximally explain the data in \mathbf{X} or more precisely, to minimize the residual sum of squares (RSS) in $\hat{\mathbf{X}}$ given these two alternating least squares problems:

- $\mathbf{X} = \mathbf{aZ}$
- $\mathbf{Z} = \mathbf{BX}$

Each row of \mathbf{a} represents one respondent.

- The sum of values in that row = 1.00
- Each value is constrained to be ≥ 0.00 and ≤ 1.00

Each row of \mathbf{B} represents one archetype.

- The sum of values in that row = 1.00
- Each value is constrained to be ≥ 0.00 and ≤ 1.00

Algorithm:

1. Initialize \mathbf{B} matrix
2. Solve for \mathbf{a} using constrained least squares
3. Solve for \mathbf{B} using constrained least squares
4. Check for improvement in RSS; return to 2) if needed

Archetypes and Baby Boomers

We examined what we refer to as the cultural heterogeneity among American Baby Boomers in a proprietary study conducted in 2007. It used an online sample of 2,000 Americans born between 1946 and 1964. Data were balanced demographically, with particular attention to education. The study punctures various media-inspired myths about this large generation, but those findings are secondary to our purpose here of illustrating the use of archetypal analysis.

The survey collected a range of data about attitudes, lives, and behavior with respect to domains such as:

- Musical tastes
- Politics
- Religion and spirituality

- Media use
- Respect for public figures
- Financial status
- Sex, drugs and rock 'n roll
- Identification as Baby Boomers
- Leisure activities
- Environmentalism

Preliminary analysis quickly made it clear that women and men needed to be analyzed separately. And as any good marketer would predict, the idea that Boomers are all alike is utter nonsense. We finally settled on four male archetypes and four female archetypes that both fit the observed data well and made intuitive sense. Only one pair of female-male archetypes were similar enough to be combined for purposes of reporting (leaving seven different archetypes).

These seven archetypes (or ideal types) provide the DNA for U.S. Baby Boomers in a cultural sense.

- **Cultural Conservative.** This was the only unisex archetype. He/she is not at home in what is seen as an immoral dominant culture. Clings to traditional values and ways of living.
- **New Ager.** Male. Closest to the media stereotype of the Boomer. Identifies as such, loves the old music, very liberal in politics. Reflective, spiritual, self-indulgent. Read all the right books.
- **Left Behind.** Male. Little identity with Boomers, little involvement in the community, religion, or even voting. Less educated, less affluent, bigger consumer of television.
- **Grindstone.** Male. Educated but not self-reflective, almost anti-religious. Some identification with Boomers but not into the music or self-indulgence. Politically tolerant and environmentally conscious.
- **My Way.** Female. Broke the mold for women's roles, and knows it. Assertive and confident. Identifies as a Boomer and loves the music. Politically tolerant but with a strong sense of personal responsibility. Religious and spiritual, environmentally conscious.
- **Survivor.** Female. Homebody and caretaker. Feels some affinity for the Boomer generation, but is on the margins. Not involved much in the community. Respectability is important.
- **Second Wave Sister.** Female. Feels she is living a life different from her parents, but has little affinity for the Boomer generation and its music. Politically liberal and socially tolerant. On younger end and feels she may have missed the gold ring.

A seemingly natural next question is "How big was each group?" But archetypal analysis is not about defining groups . . . at least not in the familiar way.

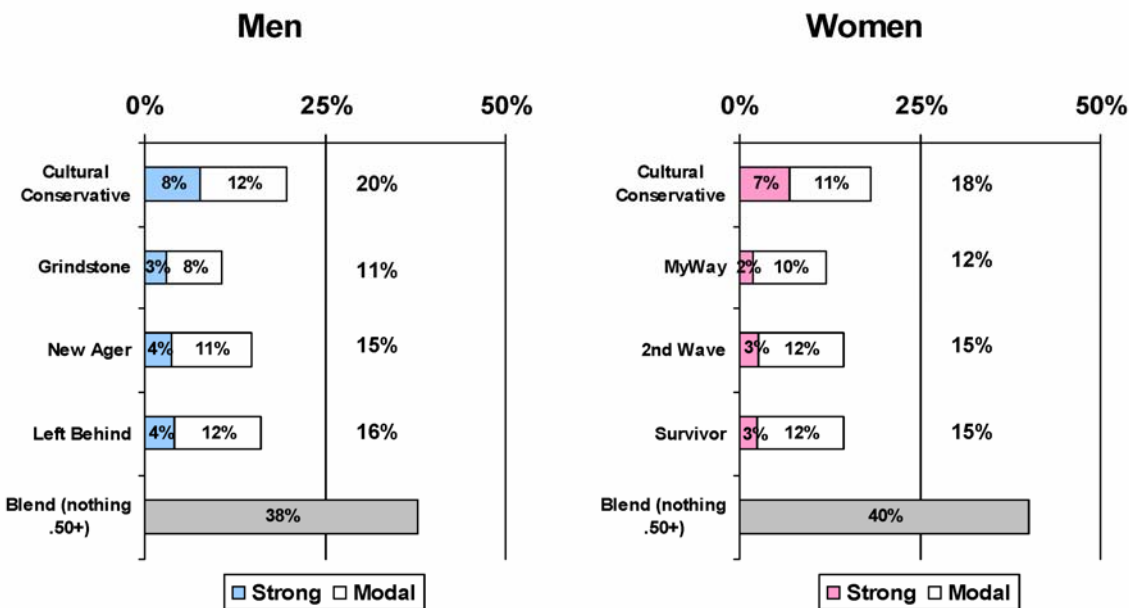
The distribution of the various types of archetypal DNA (i.e. the archetype weights) within the generation is fairly uniform—from 21% to 28% among both men and women. What is startling for those accustomed to thinking in terms of conventional segments is how few relatively "pure" embodiments there are of the seven (eight if you prefer) archetypes.

Recall that each survey respondent is defined as the product of the four gender-specific archetypal weights. One heuristic we employed was to tally the number of Boomers who have any archetype weight of .70 or higher. Call them "strong" instances. Only 16 percent can be treated as near-pure representatives of one or another archetype. Instead, 82% of men and 86% of women in our sample are blends.

A more lax, though potentially useful, standard is to define anyone with a weight of .50 or more as at least a "modal" embodiment of that archetype. With this rule, more than half the sample can be crunched into single buckets. And what are the largest of these groups? Why, surely the former hippie, male New Agers and the feminist, crystal gazing My Way women, right? Wrong. Of all the Boomers who are not

mixes of these archetypes, the largest contingent aligns with the Cultural Conservative archetype. Our friends in advertising need to look again at their suppositions about the Baby Boom generation.

Distribution of Strong, Modal, and Archotypically Blended Baby Boomers



Even though this study was deliberately not focused on a particular category or product, any firm trying to communicate with Boomers should be challenged to raise some questions?

- How well do their messages actually play against the different cultural archetypes?
- How relevant to each archetype are the emotional benefits and values being emphasized?
- What is the distribution of the archetypal DNA across its (presumably tighter) target market?

Making the Most of the Method

Marketing research is nothing if not facile at borrowing and adapting ideas from other disciplines. Once upon a time, "conjoint" was something of interest only to academic psychologists, for instance. We believe that archetypal analysis could and should be applied more often to marketing problems. Like any transplanted method, it is sure to be adapted to our unique needs. Having had more experience than most in conducting such analysis (which is not saying very much), we have devised some heuristics and identified some practical problems that deserve to be examined and tested by others as well.

The practice of defining "strong" and "modal" embodiments of each archetype seems to be instructive. The cut-off points are of course arbitrary. While this practice is something of a compromise in deference those who must have "groups" to look at, it still reinforces the reality that a large number of people simply don't fit in one box. And that is a reality.

If a respondent is neither fish nor fowl (i.e. in one segment or another), how can their data even be tabulated? Our solution with the Boomer data was to create four cases from each original case, weighted by each of the four archetypal weights or loadings. The resulting N was still 2,000. As long as the analyst remembers that she is profiling, say, the New Ager DNA and not New Ager_s, good insights can be found. It is admittedly an abstraction to be comparing bodies of DNA rather than buckets of people in a crosstabulation.

The logic would be that if you found a particular archetype to be over-represented in your target audience, then identifying values or needs strongly tied to that archetype is still useful. You might just elect to profile your target audience's values and needs, and not worry about archetypes. However, there is something of a *Gestalt* to them that can be put to work.

Researchers customarily use discriminant analysis to reproduce segment classifications from a smaller number of variables. This is not possible with archetypal analysis, but there is still a need to "score" fresh samples in terms of archetype weights. Though imperfect, tobit probability models for each archetype work well in practice. Trial-and-error is needed to identify subsets of predictors, and \hat{y} values need to be truncated to limits of 0.0 and 1.0. In our Boomer data, we achieved correlations between the original archetype weights and the modeled weights of +0.70 or better, with some over +0.90.

Concluding Note

Archetypal analysis has had a long, slow gestation in marketing research. What it will grow into—if it grows at all—is difficult to predict. As opposed to cluster-based segmentation, it honestly portrays many/most consumers as the blends they are rather than falling into a few tidy categories. Understanding the pure or extreme archetypes is especially relevant in the area of communications. On the down side, the application of archetypal analysis will be hindered by its unfamiliarity and the absence of off-the-shelf software solutions. Even so, we hope to have piqued the interest of others in exploring its value as a form of segmentation and possibly other applications.

##

References

Cutler, Adele (1993), "A Branch and Bound Algorithm for Constrained Least Squares," *Communications in Statistics - Simulation and Computation*, 22, 305-321.

Cutler, Adele and Leo Breiman (1994), "Archetypal Analysis," *Technometrics*, 36, 338-347.

Louviere, Jordan and Richard Carson (1998), "Archetypal Analysis and Segmentation," presented at the Advanced Research Techniques Forum, American Marketing Association, Keystone, Colorado.