PrimeSeg:
Development of a Geodemographic Classification System of Older Americans

Thomas E. Godfrey
Stephen J. Tordella
Decision Demographics
Arlington VA

Prepared for presentation at the 2004 Annual Population Association of America Meetings-
Boston, Massachusetts, April-3, 2004

Introduction

Decision Demographics has developed a national tract-level geodemographic segmentation system classifying persons age 60 and older into 18 distinct segments. While geodemographic segmentation systems have been commercially available for nearly 30 years, this paper shares the story, an on-going story, of Decision Demographics' recent foray into developing a new segmentation system. Our goal is to share some of our experiences, findings and challenges we have faced throughout the process to this point.

Background

Like many commercial endeavors, the idea of developing a specialized segmentation system came about somewhat unexpectedly. Several years ago we were conducting a series of interviews with market researchers in industries targeting older markets. Many of these interviews were with researchers in the senior housing industry, nursing home and Medicare supplement insurance industries. Among our findings was the fact that the majority of researchers who ran more than a rudimentary research and analysis department all used geodemographic segmentation systems. Typical uses of such a system were described. They told us how they could winnow a list of potential sites locations or strengthen response to their direct mail campaigns. While a common theme was a litany of frustrations in essentially using income as a proxy for wealth, in hindsight we discovered another: the inability of their segmentation systems to target their populations. We realized this when asking one researcher an open-ended question soliciting their ideal data tool for their job. This respondent replied that what she needed was a segmentation system specific to older people.

This comment sparked our interest and led to Decision Demographics applying for and being granted a Phase I Small Business Innovation Research Grant through the U.S. Department of Health and Human Services, Public Heath Service. In the SBIR grant program, a Phase I grant is essentially an idea feasibility study grant. This is a report on some of our findings of this feasibility study.

The Geodemographic Segmentation System Marketplace and Current Targeting of Seniors

Geodemographics came into being in the early 1970s with the advent of CACI's ACORN system in Great Britain and Claritas' PRIZM system in the United States. Since then, such tools have become standard since they can be used to identify a specific population (in a somewhat generic way) and target, track, and serve its needs. While the recent explosion of geographic information systems has given the term "geodemographics" many meanings, in the demographic industry it refers to applying neighborhood classification systems to client service and marketing activities. Indeed, ACORN, the first such system's acronym, stands for A Classification of Residential Neighborhoods.

As long as ten years ago Brad Edmondson was reporting in *The Number News* that the choice among segmentation systems in the U.S. was declining due to significant merger and acquisition activity. VNU, a Dutch firm, owns both Claritas' PRIZM system and National Decision Systems' MicroVisioin, while CACI (now ESRI Business Information Solutions) maintains the

ACORN system. Because these systems were developed as tools to reduce mass markets to identifiable and targetable units, their focus has been to describe all residents in each neighborhood. Clusters are typically identified from several characteristics, such as level of urbanization and socioeconomic profile. Neighborhoods are then classified by a few predominant traits that have been chosen to drive overall classification.

In all clustering systems, this approach has identified a number of emerging clusters with large older populations. In CACI's ACORN system, 7 of the 43 clusters could be considered "senior" clusters. These include retirement communities, active senior singles, prosperous older couple, wealthiest seniors, rural resort dwellers, senior sun seekers, and social security dependents.

In Claritas' dominant PRIZM system, 11 out of the 62 cluster have disproportionate shares of older residents. Several examples of these eleven include clusters such as:

- Gray Power: Affluent Retirees in Sunbelt Cities
- Gray Collars: Aging Couple in Inner Suburbs
- Sunset City Blues: Empty Nests in Aging Industrial Cities
- Hometown Retired: Low-Income, Older Singles and Couples
- Back Country Folks: Remote Rural/Town Families
- Scrub Pine Flats: Older African-American Farm Families
- Hard Scrabble: Older Families in Poor Isolated Areas

Besides the eleven clusters with disproportionate shares of older residents, twelve additional clusters list the 65 and over age group as making up a significant share of the cluster's population. However, these clusters also include a significant population younger than 65.

Likewise, in NDS' MicroVision system, 7 or 49 clusters could be considered senior clusters and an additional six clusters are mixed clusters with by-modal age groups. These have a significant share of young people and older people with a dearth of middle-aged residents. This may be an artifact of the classification system, which focuses on items such as income and family/household composition and size. On these dimensions, the youngest and oldest households may appear similar.

Geodemographic Marketplace Developments

Several trends suggest that development of a new, specialized cluster system for the older population is not currently imminent. In the geodemographic industry, product and service development is largely focused on vertically-integrated markets defined by industries such as banking or retailing.

Programs based on individual customer data records (so-called database marketing) have become increasingly common. While targeting individuals is efficient in telemarketing and direct mail, this may not be the preferred method for social services, an important potential use of a segmentation of older persons. Delivery of these services often involves identifying a suitable site location and developing a service territory. Understanding neighborhood characteristics still reigns as a dominant need in this research.

In many individually-based marketing programs, data quality remains an issue. Unlike Census data, where collection, analysis, and reporting are coordinated in one central agency, individual-level data are captured by various methods with no independent oversight. Data may come from diverse sources as warranty cards, credit reports, coupon incentives, and lengthy product-use surveys. In addition, local telephone, driver's license, and other administrative records are used. Data quality also decays as these secondary uses of the data more quickly become less accurate over time than the generally relatively slow evolution of neighborhood characteristics. In addition, availability of some of this information is being threatened by new privacy laws.

Consolidation of the geodemographic industry made clear the need for niche products in marketing information. Indeed, Barbara Clarke-O'Hare had noted this as early as 1994 in *Marketing Tools* magazine. Industry-specific cluster systems are available for the financial services and automotive industries, and some system focus on specific consumer characteristics or purchase behaviors. These may cover various population such as Hispanics, television viewers, or magazine readers.

A direct result of industry consolidation is the emergence of Applied Geographic Solutions, Inc., which incorporated early in 1997. Their marketing literature touts their formation in "response to the rapidly decreasing number of reliable demographic data providers in the United States". This firm supplies Experian's Mosaic lifestyle segmentation data, which represents the globalization of geodemographic systems with systems for the United States and 17 other countries. Moreover, this system, which classifies nearly three-quarters of a billion consumers, has only 14 underlying lifestyle types in common across the 18 countries covers. Experian has come onto the scene touting globalization and improved data reduction while claiming to perform comparably to other commercial U.S. segmentation offerings. Experian also represents another player using individual-level data drawing from their vast credit card information resources.

Segmentation of the 60 and Older Population

In our feasibility stage, we decided to start our segmentation development at the census tract level. This was driven primarily by two factors. Compared to the quarter-million block groups across the nation in 1990, we would be working with only 62,000 census tracts. This would reduce both computing demands as well as potential issues related to small sample sizes.

For data resources we were aware of and then confirmed that the 1990 Census Special Tabulation on Aging, STP-14 data set fit our needs quite well. This special tabulation for the Agency on Aging essentially is a national SF-3 file repeated for three age groups—those age 60-64, 65-74 and 75 and over. For these three age groups the majority of long-form socio-economic and housing items are tabulated.

Actual clustering was performed by SAS' Fastclus Procedure. This procedure is specifically designed for the clustering of large data sets. It selects observations for cluster seeds and assigns each case to the nearest seed based on computed Euclidean distances. After each observation is assigned, the cluster seeds are replaced by the cluster means. This step is repeated until the

changes in the cluster seeds become small. Finally, clusters are formed by assigning each observation to the nearest cluster seed.

Prior clustering experience using a client's survey data taught us the importance of scaling data before clustering. Our client's surveys typically ask either binary yes-no questions or 5-point scale questions. Analysis of clusters would often suggest that the 5-point scale questions were largely driving the cluster assignments due to their higher standard deviation. One solution to this is standardizing variance. We found these adjustments, while solving the problem, presented another more practical hurdle; adjustment of variables back to original metric for interpretation. Interpretation of clustering output is largely a subjective process and we found evidence on both sides of the camp as far as the necessity and advisability of standardizing input data. As a way around this issue, we decided to convert our data items to a percentage share of a tract's population. For example, for an education measure we might compute the percentage of a tract's population age 65 to 74 who were college educated.

At this point we carried out some initial clustering to see how the data would fall. Since our clustering routine requires the resulting number of clusters to be predetermined, we produced cluster solutions all the way from a two-cluster solution up to a fifteen-cluster solution. We then examined the means for all variables by cluster for each of these fourteen solutions. Results were promising as it was evident that clusters had logical commonalities largely based on socio-economic factors and then these socio-economic factors would divide based on regional differentiators as more clusters were created.

One of the first refinements introduced was some data reduction. Factor analysis help guide us in the detection of variables based on the same construct . While some would argue we should use the factors as the input for clustering, we chose to use the factor analysis to guide removal of related categories. This decision was driven by a commercial reality. When clusters are determined with factors alone, it is nearly impossible to explain to a client how a cluster is derived. With actual variables it is easier to develop a post hoc cluster assignment equation. Also, actual variables would make cluster interpretation and evaluation somewhat easier.

Another finding from the initial clustering was that a number of tracts were dominated by or completely occupied by group-quarters persons—not overly surprising considering the population, these group quarters were nursing homes. These tracts were set aside.

The last change we made was the introduction of a priori breakouts. These breakouts serve as the backbone of our segmentation. While some of today's major clustering systems have a backbone based on urbanization and income, we explored other options reflecting our unique population such as age distribution and share of the age 65 and older population that was retired. Little or insufficient differentiation was apparent with these options. In the end we divided all census tracts into four groups based on median income quartiles for the age 60 and older population. Other factors considered such as retirement levels did not provide sufficient differentiation across all parts of the nation. Complex measures of income such as a combination of several census income measures was considered and tested. However, these measures were only marginally different from simple median incomes. Again, simplicity of explanation to the end user combined with little gain resulted in the simpler approach to win out. While the initial

clustering suggested a strong urbanization component underlying the clusters, we chose not to make an a priori breakout on this factor largely because resulting group sizes would be relatively small before they were even introduced to the clustering algorithm.

Taking each of the four groupings of tracts based on median income, we produced several cluster solutions for analysis. After examining the differentiation among clusters with standard census data , we turned to a third-party data source. All developers of geo-demographic segmentation systems that we contacted emphasized the importance of using actual product/service consumption data for model development and testing. Following this industry standard practice, we acquired data from a national consumer consumption survey. These consumer surveys are massive—paid respondents fill out questionnaires that are hundreds of pages long eliciting responses about nearly every imaginable product that could possibly be purchased by a person for household or personal consumption. Besides the consumption items, several batteries of questions about attitudes are also included. Not only are these surveys long, sample sizes are impressive with responses from tens of thousands respondents each year. We chose to use the detailed purchase and attitude questions but not the extremely detailed brand and frequency of use information. This approach still resulted in a core of nearly 900 items available for profiling by cluster.

Despite the large initial sample size, restricting respondents to those age 65 and older eliminated three-quarters of respondents. Then due to problems in appending respondent census tract codes, our data supplier lost an additional one-third of the sample. In the end we were able to link census tracts for 4,100 respondents. With 18 segments this resulted in an average of 229 responses per segment. While the sample size was lower than we had hoped, significant consumption pattern differences were evident across segments. Patterns quickly emerged that were often corroborated and supported with the full census data for the segment.

At this point the process became quite subjective. Some clusters solutions displayed little differentiation while some resulted in a few smaller highly differentiated segments and a few large “leftover” segments. The final clusters chosen were arrived at from a variety of approaches. For example, a four cluster solution in a given quartile would be chosen but then an individual cluster among these four might have been further broken out into another two or three clusters. In many cases further breakouts were attempted but then rejected because little meaningful differentiation was apparent. In other cases different consumption and attitudinal patterns were readily apparent and the additional breakout was retained.

With cluster solutions chosen, the final data task was cluster assignments for the tracts that had been eliminated. Besides dropping the nursing home tracts mentioned earlier, several hundred tracts were eliminated because of missing data due to low population size. Following typical industry practices, we turned to Classification and Regression Tree procedures (C&RT). This procedure finds highly related surrogate measures to use as a replacement when an item is missing data. Despite these procedures, the data for tracts with very low population sizes is not very robust. Nevertheless, the reality of a geo-demographic system means customers expect complete coverage and an assignment based on some data remains better than no assignment. The group-quarters tracts were left unassigned.

Bringing the Segmentation System to Market

A segmentation system needs to be accompanied by short and evocative descriptions of each segment. These thumbnail descriptions are not only a valuable marketing tool but a very useful way to quickly understand and grasp each dimension of the segmentation system. Researchers using segmentation systems also find these descriptions useful in conveying their findings to clients and management. Short, lively descriptive summaries of a segment are far more easily remembered and understood than a series of segment numbers or letters.

Finally, and most importantly to bring a segmentation system to market it must be known to work and provide value. While the use of consumer consumption data gives confidence that the system will explain and predict consumption, demonstration of utility using real-life marketing campaigns is far more powerful. This is the stage we are presently at. We have found some partners who have an older constituents base and were willing to share their data. By geocoding the addresses on their mailing lists we are able to then analyze their targets by cluster assignment. Initial results have been somewhat promising though we still need additional research. One partner has tried using one of the well-known segmentation system on the market and found mixed results. Our results have been mixed too. Currently, we are sorting out whether this is a function of our segmentation system's performance or the nature of the client's population.

Future plans include further comparative tests using actual marketing data using this 1990 census based system. Ultimately, we plan to recast and recluster using the 2000 Census data. Mostly likely we will move toward assigning each block group a cluster. While this will present additional challenges due to the small size of the age 60 and over population, we feel the market demands it, since all commercial systems deliver data down to the block group. Besides the data efforts in creating clusters and cluster assignments, much work will be invested in researching and developing client delivery mechanisms that meet client needs and expectations.