Panorama: Philosophy and Methodology

Gary Menger, President
Applied Geographic Solutions, Inc.
2016A

The term geodemographic segmentation now dates back several decades, and most people who utilize geographically based business analytics are familiar with the various products available and the range of applications for which they are useful. In other words, people are generally familiar with the concepts and the statistical methodologies employed, perhaps now too familiar. Rather than merely recycling the now ubiquitous and sterile statistical jargon, it is appropriate to discuss segmentation from a broader perspective. While the goal here is, at least in part, to explain the how and what of Panorama, it is in my view equally important for us to step back and ask the annoying why and whatever for questions.

What is Segmentation?

Before we do so, let us begin by reminding ourselves what geodemographic segmentation actually is. Essentially, we are undertaking a classification or taxonomy (the segmentation) of geographic areas (geo) using the attributes of the resident population and the neighborhoods in which they live (demographics).

The purpose of a classification system is to simplify a complex reality – essentially making sure that we can see the forest rather than just the individual trees. We can choose to simplify any particular problem in a myriad of ways, some of which may be useful to us, others not so much. It is therefore vital to determine in advance the purpose of the classification, as that will guide the approach to the problem itself.

By way of example, assume that we have a set of one hundred objects of various shapes, sizes, colors, textures, and uses, and we ask a child to put them into groups. We can imagine that one child would group them according to color, another according to size, and yet another according to shape. Assuming that the child was at least minimally enthused about completing the task, all would be valid and reasonable classifications of the objects, and yet each would be useful for different purposes.

None would be useful for all purposes, as would be readily demonstrated by asking the child who grouped their objects according to shape to eat the round group which regrettably, albeit correctly according to their methodology, includes both an orange and a baseball. Likewise, asking the child who grouped their objects according to color to play with the brown group would possibly find it difficult to simultaneously enjoy playing with the football and the porcupine.

While these might seem silly examples, they nevertheless demonstrate a very important principle of classification, which is that in order to be effective, **the purpose of that classification must be known in advance**. The actual methods used to classify is more or less constant, but we can easily see that if the purpose of the classification is to know what foods to eat that we should not be grouping predominantly on other characteristics.

The method of classification is basically to identify a set of characteristics (e.g. shape, size, color) of a group of objects and to group them according to how similar they are on these features. Statistically, the methods of classification have been with us for decades and while computationally complex, essentially attempt to group together a set of objects in such a way as to maximize the similarity of the objects within any group whilst maximizing the differences between the objects of different groups.

Taking our set of one hundred objects, we can easily imagine grouping by color. This is an easy task until we come across an object which doesn't neatly fit into any particular category, perhaps an object that has polka dots on it. At this point, we could classify it according to its dominant color or we could create a new group for multi-colored objects. The next item, a plaid shirt, poses an additional problem, as it is multi-colored, but has a clear pattern to it. Yet another item may have two colors but no apparent pattern. The problem comes down to how significant the differences need be before we create a new group rather than add the object to an existing group despite its less than perfect fit. At the extremes, we could state that all one hundred objects have sufficient differences such that each should be its own group (in which case we have failed to simplify) or that since all of our objects are indeed all objects that there should be only one group (in which case we have clearly over simplified). In either event, the benefits of classification have been lost.

This leads us to a second essential characteristic of classification, that **the number of groups created is arbitrary**. Indeed, the appropriate number of groups depends upon the purpose of the classification and will be by no means exact. To create a classification based on edibility, we would truly only need two groups. But to make that more useful, we would probably expand this to at least four – harmful if eaten (a broken glass), not edible but not harmful (dirt), edible but nasty (Brussel sprouts), and good to eat (chocolate cake). We can imagine that we could further classify the good to eat group by nutritional value, when it is normally served, whether to cook it or eat it raw, and so forth. Too many groups, and the system becomes unwieldy, however and we might ought to have bothered grouping in the first place.

One further point is that the number and type of attributes with which we undertake the classification are arbitrary as well. In our examples above, one child may actually taste each object while another does not. Yet another may systematically drop them to see which break. All, however, are likely to utilize the obvious characteristics – size, color, shape, texture – although perhaps without affording them identical importance (one may do a 'first pass' using color, another using size). Often, we can criticize the classification because it fails to use our particular purpose in its definition – perhaps

something like 'is it good to throw and catch'. While we would be disappointed that the segmentation didn't include this variable as it is after all our primary goal of the segmentation, the reality is that if we were to look at the segments with shapes that are good to throw (spherical seems best), have a good texture (too hard or too soft is not desirable for differing reasons), and so on, we would in all likelihood be able to ascertain the appropriate groups for our particular purpose. All this to say, **the grouping need not be created for your purpose to be useful to you**. An all-purpose system is going to provide utility for most purposes, but, if we are honest, would not be the particular choice to solve any one of those individual purposes. Similar to the handyman that you hire for around your house, he is probably competent at a very wide range of tasks, but highly skilled in but a few of them.

What Do We Do with Segmentation?

One of the original reasons why demographics firms began to develop segmentation systems is that they provide an easy way to instantly convey an area's "feel". Imagine that you have three sites to compare for a potential location, and at least a reasonable formulation of what constitutes a "good" site. That "good" site may even be a composite demographic profile of the set of good sites. So we sit down with four separate reports of detailed, usually multi-page demographics and attempt to do a comparison. This can be overwhelming, and it is very difficult for most of us to consider more than one aspect at a time. For example, we may know that we do well in areas where there are both young families and the adults are well educated. Most analysts can consider one of these factors at a time, and have a difficult time with trading off between nuanced differences between sites.

Enter the segmentation system with its simplicity. Perhaps sixty or seventy segments, each named and described. We can easily fit these four sites on a single page and undertake a comparison based on the segment distribution of each site. The segmentation profiles (e.g. relative distribution of the population over the segments) can be quickly compared, with statistical measures should they be desired.

I have always referred to segmentation based site location research as "the poor man's site model". While less accurate than using the data hungry and sophisticated techniques which have been developed, we can implement such a model at a mere fraction of the cost. This is a **predictive** application of segmentation.

But more than that, we can use the segmentation labels and descriptions to quickly get a relatively accurate mental picture of an unfamiliar area. We find that a site has predominantly households of Panorama group 14, "American Playgrounds". If we have used the system for any period of time, we know pretty much exactly what it means, as we have read the description and are no doubt familiar with some of the locations described – Myrtle Beach, Big Bear and Cambria California – and we have an instant visual image of what our particular site is like. For this segment, the key is that we have a stable, older local population which mixes with a major tourist component. Trying to pull this simple image from a multi-page demographics report is a very tedious task. This is the **descriptive** usage of segmentation.

A third usage comes about through direct marketing. We might know from past mailings which segments have responded best to our advertising, and we can quickly build mailing programs which target those groups. For other forms of advertising, we can look at the profile (distribution of

segments) over the appropriate coverage area – say a radio station reception contour or an IP address location.

Fourth, and less visible to most users, we can use the geographic information in the segmentation system to code small and medium size surveys with Panorama codes, then apply the results nationwide. In effect, we assume that the behaviors represented in the survey are heavily influenced by the demographics which defined the segments. If in the survey, we find that 8% of segment 04 households own a BMW, we can with some assurance apply that statistic to all areas where we find segment 04 households. The most important of these surveys is the GfK MRI survey, which Panorama is linked to.

The Peculiarities of Geodemographic Segmentation

We add to the simple classification problem a major complicating factor. For many of the purposes for which geographic data are used – retail site selection, performance analysis, direct marketing, neighborhood description, and so forth – the ultimate object of interest is not the geographic area itself but rather the individuals who live there. But this is not precise, as the individual who utilizes our product may not be the one who physically purchases it or the one who funds its acquisition. Let us call this a "consumption unit". Its contents are rather annoyingly fluid depending on the particular product(s), but it generally can be considered to be a household, albeit with leakages (e.g. a product is purchased as a gift for someone outside the household unit or, a gift is given to a member of the household).

We could simply choose to group households according to their characteristics, and at first glance, this might seem ideal. However, we often find that we don't know what household is our target, simply that they live within an identifiable geographic area. We can choose any aggregation of geography, perhaps a zip code, but more often than not a block group.

For geographic areas, the smallest truly practical level is the block group, as this is the smallest scale for which we have consistent, relatively accurate, and comprehensive data. It might be argued that we could take household data and aggregate it to that level, and that this would be a more useful tool overall. For the following reasons, we argue that a general purpose segmentation system is best created at the block group rather than household level:

- **Data Comprehensiveness:** A vast array of data is available at the block group, primarily from the Census Bureau, and with generally known properties. While obviously not perfect, the data is of sufficiently high quality that we do not need to supplement it for the characteristics which it measures. While one could argue that block geography may be appropriate, it should be noted that there is a very narrow range of characteristics at that level and that many characteristics have to be imputed from the block group. Conversely, while household level data has become more complete, accurate, and extensive over the past few decades, many of the characteristics of households are in fact imputed from the geographic data. Many of the attributes of households are not known for many households and neighborhood averages are utilized. Most household level databases are good at identifying the age and sex of the adult members of the household, but are often lacking in details on children. Likewise, it is generally assumed that household level systems have access to the income and credit data which is maintained by the credit bureaus. This, however, is simply not true, as these data are protected

from use or disclosure by federal statutes.

- **Temporal Consistency:** Household level data is extremely fluid and there is no specific date as to which the data clearly reference. The information on the age of the individuals may be current, or it could be several years old, and it is generally not possible to determine which is the case for any particular record. On the other hand, with any Census or American Community Survey (ACS) release, the data is referenced to a specific date.

- **Error Tolerance**: At the household level, any segmentation system is effectively a rules-based system which pre-defines at least the gross characteristics of the groups – age, size and composition of household, imputed income level, and so forth. Any error in the measurement, whether it be due to temporal consistency, comprehensiveness, or imputation, will in all likelihood affect the precise classification of the household itself. On the other hand, the same error would not likely change the classification of the geographic area, as it is based not on the precise characteristics of the individual household, but the overall characteristics of an aggregate of such households – including not just the averages but the actual distributions of values.

- **Spatial Considerations:** We know that if two identical households are located in different types of neighborhoods that their consumption patterns will tend to reflect their neighborhood as well as their individual characteristics. If one household is at the upper end of the income scale of a low income neighborhood and the other at the lower end of the scale in a higher income neighborhood, these two households will have surprisingly different expenditure patterns. Call it the "Keeping up with the Jones'" effect. The locational context has impacts on many things such as travel to work, recreational preferences, hobbies, and even musical tastes. And these form the basis of what makes the feel of, for example, Syracuse NY different from that of Albuquerque NM. While a household based system can obviously include items such as population density, climate, and so forth, they are more suited to geographic systems.

Given these differences, we believe that the following statements can be made:

- For analytical purposes where the individual consumer units are known (e.g. address of the household), or for direct marketing purposes where individual consumers are targeted and their physical address known, a household based segmentation system is preferable.

- For analytical purposes where the main unit of analysis is an aggregate of households (ZIP+4, block groups, drive time studies around sites, etc.), we believe that a household system offers no advantages, despite the rhetoric in the marketplace to the contrary. A geographically based system is more stable and the data more comprehensive, and we believe results in superior analysis.

- For geographic area description, geographically-based segmentation is greatly preferred, as the intent is to classify forests rather than individual trees. It is far too easy in a household based system to lose sight of the visual image of the forest because too many rare tree types are represented. A typical Panorama report for a five mile ring will show far fewer segments

represented than an equivalent area using a household based system

- For survey linkage, there are few differences between the performance of systems provided that they are well constructed.

For these reasons, Panorama has been constructed at its native, or base, level at the block group. Individual census blocks can be modified from the parent block group type on the basis of a subset of demographics for which we have sufficient data. In our view, this reflects the orientation of most of the AGS users towards spatial description and analytics usages rather than household targeted analytics.

Panorama is therefore a truly geographic segmentation system rather than an individual grouping system generalized to geographic areas.

Statistical Methods and Considerations

The vast majority of segmentation exercises use what is known as k-means clustering. Essentially, the analyst states up front the number of segments to be created (N). Most algorithms begin by taking the first N records and setting them as the initial cluster centers based on the values of the variables. Each subsequent record is then compared to each segment using the difference between its values and the cluster center, assigning the record to its nearest (minimized difference) segment. The segment centroid (or average of our variables) is then recomputed and the next record processed. At the end of all records, we then have an initial cluster solution. Since our cluster centers have changed during processing, we take a second pass through the data, changing the segment assignment and the cluster averages on the fly. We continue to iterate through the data until a specified number of passes have been made or no further changes occur.

This method is highly dependent upon the starting points, and for geographic data, we are actually talking about Autauga County, Alabama block groups serving as the cluster center basis. This is clearly not the best approach.

Over the course of working with a number of segmentation systems since the early 1980's, we have learned that it is preferable to begin with a more deliberate starting solution.

We utilized the two main dimensions of our Demographic Dimensions product, Affluence and Family Status as the starting point. We selected 8 target values from each dimension, then located for each combination the block group most closely resembling it nationwide. These then served as the starting segment "seeds". The original k-means solutions were based on an eight by eight starting point, or 64 segments.

The initial segment allocation of sixty-four eventually became sixty-eight as our goal was to ensure that segments included both a minimum and maximum population count. Variables were added and deleted from the analysis in order to ensure that the solution selected was stable, as we have found from past experience that small segments have a tendency to be underrepresented in syndicated surveys (thereby resulting in less reliable data) and that if segments are allowed to be too large, that an insufficient level of generalization occurs.

We also employed a weighting technique which allowed us to systematically modify the relative importance of each variable group on the final allocation. Modifying the weighting of variable groups is an especially important technique in splitting large groups in a meaningful fashion.

Variable Groups and Sets

The following sets of variables (e.g. a table of population by age) were used and grouped as below, as weights were employed at the group level to ensure that the number of variables in a particular set did not adversely affect the results. By way of example, using nineteen age groups (e.g. 0-4, 5-9…) and two sex groups (male, female) would result in the overall weighting of age to be much greater than sex in the results. Weights are therefore applied to sets of variables.

- Population
  - Age
  - Median Age
  - Sex
  - Race and Hispanic Origin
  - Educational Attainment (high school, some college, etc.)
  - Marital status

- Group Quarters Population
  - By type (correctional institutions, college dormitories, etc.)

- Households
  - Household structure (married couple with children, lone parent male)
  - Age of head of household
  - Vehicle Status (0, 1, 2+)
  - Household Size
  - Family Household Size
  - Linguistic isolation (English speaking, non-isolated Spanish, isolated Spanish)

- Labor Force
  - Status (employed, unemployed, armed forces, not in labor force)
  - Labor force participation rate
  - Occupation by industry (mining, retail trade, etc.)
  - Class of worker (private for profit, non-profit, unpaid family, government)
  - Historical unemployment change (1980-2000, 2000-2010, 2010-2016)

- Income
  - Median Household and Family
  - Per Capita
  - Households by income range (e.g. <$10,000, $10-$15000, etc.)
  - Historical income change (1980-2000, 2000-2010, 2010-2016)

- Housing Occupancy

        - Tenure (own, rent, vacant)
        - Vacancy reason (seasonal, for migrant workers, for rent/sale, long term vacant)

    - Housing Characteristics
        - Value of owner occupied dwellings
        - Median home value
        - Monthly rental costs
        - Median rent
        - Units in structure (single family detached, apts with > 50 units, mobile homes)
        - Year of construction (e.g. 1970-1979, 2000-2005)

    - Locational Context
        - Population density (block group)
        - Contextual density (within 5 and 10 miles)
        - Historical population growth (1980-2010, 2000-2010, 2010-2016)

Overall, several hundred variables were used in the initial analysis, with the list being narrowed to approximately one hundred and fifty, as it became apparent that many of the variables did not contribute anything substantive to the differentiation between segments (usually because of auto-correlation issues).

It should be noted that the longitudinal variables were chosen to track relative changes in population, income, and employment over time with the aim of understanding the differences between neighborhoods which are on the rise and those on the decline.

One final note on variable selection is important. The range of variables was limited to those available from the census and not including a wide range of AGS data which in and of itself relies upon those same demographics in their modeling. These instead were used during the validation process and are reported in the various AGS methodology documents. These include:

- CrimeRisk
- Assets, Debts, and Net Worth
- Consumer Expenditures
- Quality of Life
- Demographic Dimensions

Discussion of Results

The solution selected includes sixty-eight segments which range in population from just under one million to just under ten million, avoiding the problems associated with having one or more highly unusual segments.

The geographic nature of the segmentation more than emerges from the segments, as a number of segments are highly geographically concentrated in certain areas of the nation, in certain urban contexts (such as the classic beltway suburbs), and in particular environments (resort towns). The interested reader should review the full Panorama segment descriptions which include the distributions of each segment nationwide.

The segments are as follows:

01	One Percenters
02	Peak Performers
03	Second City Moguls
04	Sprawl Success
05	Transitioning Affluent Families
06	Best of Both Worlds
07	Upscale Diversity
08	Living the Dream
09	Successful Urban Refugees
10	Emerging Leaders
11	Affluent Newcomers
12	Mainstream Established Suburbs
13	Cowboy Country
14	American Playgrounds
15	Comfortable Retirement
16	Spacious Suburbs
17	New American Dreams
18	Small Town Middle Managers
19	Outer Suburban Affluence
20	Rugged Individualists
21	New Suburban Style
22	Up and Coming Suburban Diversity
23	Enduring Heartland
24	Isolated Hispanic Neighborhoods
25	Hipsters and Geeks
26	High Density Diversity
27	Young Coastal Technocrats
28	Asian-Hispanic Fusion
29	Big Apple Dreamers
30	True Grit
31	Working Hispania
32	Struggling Singles
33	Nor'Easters
34	Midwestern Comforts
35	Generational Dreams
36	Olde New England
37	Faded Industrial Dreams
38	Failing Prospects
39	Second City Beginnings
40	Beltway Commuters
41	Garden Variety Suburbia
42	Rising Fortunes
43	Classic Interstate Suburbia
44	Pacific Second City
45	Northern Blues
46	Recessive Singles
47	Simply Southern
48	Tex-Mex
49	Sierra Siesta
50	Great Plains, Great Struggles
51	Boots and Brews
52	Great Open Country
53	Classic Dixie
54	Off the Beaten Path
55	Hollows and Hills
56	Gospel and Guns
57	Cap and Gown
58	Marking Time
59	Hispanic Working Poor
60	Bordertown Blues
61	Communal Living
62	Living Here in Allentown
63	Southern Small City Blues
64	Struggling Southerners
65	Forgotten Towns
66	Post Industrial Trauma
67	Starting Out
68	Rust Belt Poverty

Where are the groups?

Most segmentation systems include a second order of classification, that of the group. The purpose of the group is to further simplify the segmentation and provide additional context by putting similar segments into each group.

These groups are typically based primarily on income, perhaps modified by urbanicity (inner city, suburban, rural). There is usually a group for those segments which don't seem to fit into any particular group, such as the group quarters dominated groups. All too often, in our experience, users simply use the groups as given, and the results are less than optimal.

Any number of groups could be created depending on what particular characteristics are used to lead the process. A retailer specializing in children's toys would find a grouping based on the age and number of children to be more useful than one based on income, for example.

For this reason, Panorama does not include segment groupings. It is our goal that end users create their own groupings on the basis of their particular circumstances and customer data (or at least surrogates for it using the GfK MRI data).

myPanorama

myPanorama is a system under development which will provide a systematic means for end users to appropriately group Panorama segments according to their own data and particular requirements. A number of "stock" groupings will be provided, including:

- Income
- Diversity
- Child Oriented
- Elderly Oriented
- Urban Context

A full range of MRI variables will be available to aid in groupings, as will be the raw demographics and additional AGS datasets such as demographic dimensions and CrimeRisk. User profiles will be importable in order to ensure that actual customer data can be used to guide group determination.

The then simplified system, myPanorama, will be a customized version of the segmentation which marries the simplicity of the group system (usually 6-12 groups) with the complexity of the individual end user environment.