PSYTE® Canada Advantage Methodology

Introduction

Geodemographic segmentation or clustering in the marketing context involves classifying small geographic areas (e.g. census dissemination areas or DA's) into relatively homogeneous market segments. The exercise produces a set of clusters or market segments that correlate well with individual preferences and consumer behaviors. The development goal of PSYTE Canada Advantage was to classify "neighbourhoods" (census DA's), into meaningful clusters, maintaining continuity with the original PSYTE Canada system but also detecting trends and incorporating significant socio-economic and cultural change where it has occurred.

The basic assumption of clustering is that people with similar characteristics, preferences, and consumer behaviors tend to live in like neighbourhoods. However, as Canadian society changes and neighbourhoods evolve, cultural and economic diversity increases. The extent of diversity – whether socio-economic, ethnic, cultural, lifestyle, life-stage, or other dimension – is such that the new PSYTE Canada Advantage takes into account unprecedented levels of "within neighbourhood" differences as well as increased diversity overall. Nevertheless, users should discover that the fundamental drivers of consumer behaviors and lifestyles within each cluster are substantially similar.

In the development of PSYTE Canada Advantage, MapInfo Canada has used the latest statistical tools and techniques in geodemography. Advances in spatial analysis, geo-statistical software, and modeling techniques – along with the raw ability of computers to implement new clustering strategies – have opened doors to advanced spatial analytic worlds undreamed of only a few years ago. This methodology statement describes the process used by MapInfo Canada's demographers and geographers to produce PSYTE Canada Advantage.

The Development Process

To some, geodemographic clustering necessarily involves a subjective process, especially in the selection of variables and how each variable is used in the analysis. However, today's computing environment, coupled with new methods of spatial analysis, make subjective methods less necessary. One primary tenet in the development of PSYTE Canada Advantage, therefore, was to "let the data speak for itself," and thereby create a more scientifically reliable set of clusters compared with "first generation" cluster systems.

MapInfo researchers adopted a two-stage clustering process.

- The first stage involved using a proprietary Kohonen classifier, while the second stage used a hierarchical classifier. The objective of the first stage was to develop sub-clusters or "atoms" that capture the essential demographic characteristics of neighbourhoods along with their settlement context. The result of the first stage was a set of 200+ atoms representing similar types of dissemination areas. It is noteworthy that the software infrastructure developed by MapInfo for this project embodies the vision of permitting "custom clustering solutions" involving the introduction of additional proprietary datasets and re-clustering of atoms.
- The second stage of cluster development involved the generation of 65 final clusters from the 200+ atoms using a hierarchical technique. Additional datasets, including measures of consumer behavior and "lifestyles," were introduced for the second-stage processing.

Setting Up the Data

The development of PSYTE Canada Advantage began with processing and defining Census 2001-based databases. Next, specific census variables were selected and defined establishing a variety of demographic dimensions. In any clustering process, the character of the input data determines to a large extent the types of clusters that emerge in the output. For example, if family structure variables are not input, the output clusters will not have a family structure dimension. Likewise, if too many family structure variables are included relative to other variables, then the segmentation system will be predominately family structure clusters.

Since geodemographic clusters are generally used in marketing and site analytic contexts, several sets of socio-economic and cultural variables were selected as primary inputs. Other variable types such as settlement context,

population density, proximity to major retail environments and community services, as well as lifestyle and purchase behavior variables (for the second stage analysis), were established and included in the processing. In the end, both census and non-census type variables were used. The non-census variables were normalized to the geography through the calculation of geographic potentials. The primary census-demographic variable sets included: population age structure, age of household maintainer, dwelling tenure and type, educational attainment, labour force participation, visible minorities, home language, migration experience (both lifetime and 5-year internal as well as international), family type (married couple and lone parent, for example), presence of children, household income (average and distribution), occupation and industry of work, and mode of travel to work. Ultimately, over 200 variables representing these dimensions and others were used in the analysis.

The Clustering Process

Once the databases were set up, the actual clustering process could begin. MapInfo analysts used the two-stage methodology described above. More specifically, the first stage involved the application of neural network geo-statistical techniques to classify the 52,993 dissemination areas using 200+ census and settlement context variables into atoms. The new PSYTE Canada Advantage implemented a proprietary neural network routine developed over several years of testing and research at MapInfo. In general, neural network techniques, which involve pattern recognition in ways that mimic the human brain, have outstanding capabilities for identifying patterns in socio-economic data.

The second stage of the clustering process was executed on the atoms. This stage used hierarchical clustering techniques to group the 200+ atoms into the final 65 clusters. (A 66th cluster designation was used for unclassified areas, essentially unpopulated DA's.) In this second stage of clustering, consumer behavior data (e.g. car purchases and lifestyle indicators) were combined with the geodemographic and settlement context-based atoms. A proprietary hierarchical technique was used providing precise control over the clustering process allowing the researchers to "craft" the clusters in a scientifically reliable way.

For the hierarchical clustering process, Principal Components Analysis (PCA), a special implementation of Factor Analysis, was used. PCA is valuable as a method for its ability to reduce large datasets into their "principal components." Each principal component represents a specific dimension of variance within the database and discards noise, or ineffectual data. PCA was used to provide meaningful components among intentional characteristics without the need for a large number of variables.

Final Steps

Once the final 65 neighbourhood clusters were established, and the analysts were content with their statistical reliability, the process of "visioning" could begin. Cluster descriptions were developed which describe the main characteristics of each neighbourhood type. Often, a unique combination of characteristics informs the "vision" of a cluster. Ultimately, each cluster is distinguished from all other clusters in the system, while simultaneously sharing characteristics similar to other clusters.