Geodemographic Segmentation: The Development of PSYTE Canada

Thomas G. Exter and Ian Mosley
MapInfo Corporation
Toronto, Ontario, Canada

A paper presented to the quadrennial conference of the

International Union for the Scientific Study of Population

Tours, France

2005

Not to be cited without proper attribution.

Geodemographic Segmentation: The Development of PSYTE Canada

By Tom Exter, Chief Demographer
And Ian Mosley, Advisory Data Engineer, MapInfo Corporation

Introduction

Geodemographic segmentation or clustering in the marketing context involves classifying small geographic areas (e.g. in Canada, 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 consumer preferences and behaviors. The development goal of PSYTE Canada Advantage (the second edition of PSYTE Canada) was to classify "neighbourhoods" into meaningful clusters, maintaining reasonable continuity with the original PSYTE Canada system but also detecting and representing new trends and incorporating socio-economic and cultural change where it has occurred.

The basic assumption of geodemographic clustering is that people with similar characteristics, preferences, and consumer behaviors tend to live in like neighbourhoods. However, with social change in Canada as elsewhere neighbourhoods evolve as cultural and economic diversity increases. The extent of diversity along multiple dimensions – whether socio-economic, ethnic, cultural, lifestyle, life-stage, or other – is such that a contemporary segmentation system must take into account unprecedented levels of "within neighbourhood" differences as well as increased diversity overall.

In the development of PSYTE Canada Advantage (a project of the authors as principal investigators within a team[*] at MapInfo Canada) the traditional tools and techniques of geodemography where used in combination with several innovative

[*] The authors wish to thank Chris Michels, Fraser Baldwin, and Kevin Antram of MapInfo Canada for their contributions to this project. Paul Thompson, also of MapInfo Canada, contributed to the case study.

Not to be cited without proper attribution.

processes discussed in this paper. 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 paper describes the process used at MapInfo Canada to produce PSYTE Canada Advantage. In addition, the rationale and methodology for a derivative system – PSYTE Quebec Advantage – is discussed. A final section of the paper illustrates how such "cluster systems" are used in a marketing context through a descriptive case study.

Literature Review

Within the general literature of statistical clustering techniques in a market research and geographic context, reviewed well beginning perhaps with Punj and Stewart (1983) and extending at a minimum to Openshaw and Turton (1996), clustering has challenged researchers in numerous ways. More recently, cluster techniques in the sciences of genetics, artificial intelligence and digital signal processing have made long strides in the ability of analysts to classify complex data into their principal structures.

One set of challenges revolves around two approaches to clustering, hierarchical and non-hierarchical (Hartigan 1975 pp. 84 – 86, Malhotta 1996 pp. 677). The former operates in two modes: agglomeration and divisive (Aldenderfer and Blashfield 1984, pp. 35 – 38, 50, Hartigan 1975, pp. 199 - 200). Agglomeration basically takes *N* observations and iteratively arrives at a single, final cluster by combining pairs of previous clusters. Divisive, on the other hand, starts with one cluster and iteratively splits clusters until *N* observations remain. Hierarchical approaches are generally

Not to be cited without proper attribution.

computationally intense and restricted in use to small data sets. Moreover, as a pairwise approach, hierarchical clustering typically results in less variance being captured when compared with more holistic non-hierarchical techniques. In contrast, non-hierarchical approaches work better with large datasets and generally capture more of the variance in a dataset (Lea, 2003), two characteristics that suggest they are well suited to geodemographic clustering.

First generation classification techniques applied to census data relied heavily on non-hierarchical approaches with the *K-means* type classifications being the most widely used. Basically, a K-means approach starts with *N* random starting positions in *M* dimensional space. The process progressively iterates through its assignments, calculations, and reassignments until all observations in the data space are optimally assigned to one of the *N* starting positions. There are two significant drawbacks to this technique: a) the random starting positions can have a dramatic affect on the final outcome, especially with a small number of clusters, and b) the measures of similarity or proximity between the starting positions (the centroids in multidimensional space) and the observations are not necessarily sensitive to the variance within the data space. (Aldenderfer and Blashfield 1984 pp. 45 – 49, Malhotta 1996 pp. 679, Hartigan 1975, pp. 102)

A "Second Generation" set of tools emerged in the 1990's when research in Artificial Intelligence and Neural Networks relating to pattern recognition and classification gained prominence (Schürmann, 1996). Some applications of these methods involved "machine vision" such as that used in the context of scanning vehicles or finger prints at border crossings (Shenk, 2003). While the datasets involved were not

Not to be cited without proper attribution.

necessarily large, these new algorithms proved to be quite adept at separating the data objects from the "noise." Moreover, for pattern recognition analysts, especially those in a time-sensitive security context, requirements demanded real-time results. Fortunately, since geodemography rarely has a real-time demand, the application of Neural Networks could be applied with longer processing times in exchange for the ability to analyze larger datasets. This opportunity, applied to the present case, offered the distinct advantage of faster processing and, ultimately, greater discrimination among clusters.

Unfortunately, the neural network approach presents its own set of methodological challenges. Essentially, a neural network classifier shares some of the characteristics of a K-means classifier: random starting locations, iterative processing, and a centroid-like classification scheme. However, the distinguishing benefit of neural networks relative to K-means is how similarity measures are calculated. Unlike K-means that uses a single measure of similarity (for example, a squared-error function; see Han 2001), in the case of a Kohonen neural net classifier, a multidimensional comparison is made that has greater discriminating power. Furthermore, to appreciate the differences with respect to processing times, when comparing the K-means and Kohonen classifiers, during an initial phase of this project, we found that several months of work with the former could be reduced to just a few weeks with the latter.

Still, issues remain as to how to best utilize this technology in geodemography. Two issues, in particular, make non-hierarchical classifiers less than optimal. In the first place, they still rely on the user providing an initial desired number of clusters. Alternatively, the analyst might want to discover an "optimal" number of clusters within the constraints of a given dataset. Secondly, random starting positions are still used.

Not to be cited without proper attribution.

While these conditions are not "show stopping," they remain detractors from the ultimate goal of developing a fully objective and efficient clustering process using non-hierarchical classifiers alone.

Hierarchical techniques, on the other hand, which have been applied less frequently due to their computational overhead, do not suffer from the requirement of random start locations and a subjective target number of clusters. This discussion may beg the question as to why, if computers are so much faster now, hierarchical classifiers are not used "brute force" to come up with clusters. Two reasons have been suggested that would give pause in this regard. The first is that, as mentioned above, the hierarchical classifiers generally do not capture variance as well as non-hierarchical techniques. The second point against a "brute force" hierarchical process is that it remains difficult to control the final relative size of the clusters given then common objective in geodemographic segmentation of generating clusters whose share of households falls within a relatively tight range.

One emerging line of inquiry that seeks to use a combination of non-hierarchical and hierarchical techniques is the relatively new field in clustering of *auto-clustering*. Additionally, lying somewhere between the classic non-hierarchical techniques and auto-clustering are a set of methods known as two-stage clustering (Parthasarathy, 2003) and hierarchical self-organizing maps or SOMs (Hagenbuchner, et al. 2003). These approaches, which are beyond the scope of this paper, attempt to combine the best features of hierarchical clustering with those of non-hierarchical clustering.

The research and product development process presented here represent a specific attempt to utilize the best available knowledge and techniques with regard to geo-

Not to be cited without proper attribution.

statistical or geodemographic clustering techniques in a "consumer demographics" context. Cluster systems are used generally as tools in site location research, marketing campaigns, and advertising. Because geodemographic clustering generally occurs in a competitive, market-oriented environment, additional constraints (e.g. product positioning vis a vis the cluster systems of other data vendors) are placed on the development process. Specifically, the market expectations of the potential user community need to be taken into account. Nevertheless, a central purpose of the present paper is to present our research findings in a neutral, more academic context and to suggest additional linkages in the ongoing discussions among demographers and geographers on both the applied and theoretical fronts.

Development Assumptions and Process Summary

Geodemographic clustering in the marketing context has traditionally (dating from the 1970's) involved spatial analytics coupled with a subjective process in which the selection of initial variables, the manner of their operationalization, and their purpose-driven weighting heavily influenced the final clusters. However, today's computing environments and new methods of spatial analysis obviate subjective methods to a greater degree. The authors suggest that subjectivity in clustering – including the implementation of predetermined cluster characteristics – is not necessary, indeed, is inhibiting to an optimal cluster solution. One primary tenet in the current research, therefore, is to "let the data speak for itself," and thereby create a more scientifically reliable set of clusters.

In summary, the research team adopted a two-stage clustering process. The first stage involved using a proprietary Kohonen neural net classifier while the second stage

Not to be cited without proper attribution.

used a hierarchical classifier. The objective of the first stage was to develop a set of clusters that captured the essential demographic characteristics of neighborhoods along with their settlement context (defined below). This was accomplished with the Kohonen classifier creating sub-clusters or "atoms." The atoms delineated a neighborhood sub-cluster set that would in turn be the starting point for the second stage of the process. A larger topic, not discussed in this paper, is that the "second stage" process can in fact move in several directions. The software infrastructure developed for this project embodies the vision of permitting "custom clustering solutions" involving the introduction of additional proprietary datasets during the second stage.

The two-stage process relies on several of fundamental assumptions. First, the census-defined Dissemination Area (DA) is the basic geographic building block for the system. Second, demographic and settlement context measures are the only core data used in the first stage. Third, additional datasets including, for example, measures of consumer "lifestyles," can be introduced for the second-stage processing but would not change the fundamental definitions of the neighborhood typology or the initial sub-cluster set. That is, the "atoms" and the initial neighborhood set are sacrosanct. These assumptions are discussed here in turn.

The first assumption, that DA's provide reliable building blocks, depends on the availability of adequate census data. While census data generally have known deficiencies, other units of analysis such as postal geographies, can introduce compounded deficiencies given the need to translate census demographic characteristics from census units to postal units. The use of DA's as building blocks – the smallest units for which comprehensive census data are published – ensures that reliable and valid data

Not to be cited without proper attribution.

are used, regardless of the application. Reliability is reasonably inherent in the data given the consistent application of census data collection techniques over time. Validity is inherent in the data to the extent that census questionnaires, along with post-censal edits and imputations, have proven valid instruments in the direct collection of basic individual and household information. Furthermore, considering that a geodemographic clustering system is an *a priori* system, as opposed to one based for example on consumer transactional data, statistical reliability is paramount. In this context, the development of PSYTE required that statistical parameters be stable and reliable measures of the underlying demographic characteristics. To that end, census Dissemination Area data were used as the primary unit of observation. In all cases, DA's were the starting points and all non-census data introduced subsequently had to be statistically significant for each DA represented. (N = 52,399)

The second assumption establishes that neighborhoods, census tracts, and dissemination areas are best described in demographic terms. This may seem obvious. However, clustering deals fundamentally with a demographic environment and data related to the demography of the population best describe that environment. It appears untenable to maintain that significant amounts of non-demographic data – consumer purchase behavior, lifestyle indicators, and other non-census-based measures – however well operationalized, can substitute for the basic demographic characteristics of a population for clustering purposes. Moreover, descriptors pertaining to settlement context – population density, proximity to commerce, complexity of street networks, and other measures – complement demographic attributes by providing spatial context to the

Not to be cited without proper attribution.

analysis, thus adding geographic dimensionality and discrimination for the clustering algorithms.

The third assumption stipulates that the operational groupings of the block groups resulting from the first-stage process are, in fact, neighborhood sets. That is, the atoms fairly describe neighbourhood types because they have similar geodemography. Once these sets are defined in the first stage, they are not altered, and they become the multidimensional, operationalized definition of the neighborhood population. Atoms may be further aggregated in subsequent stages depending on various application objectives, but the intent is that they are never disaggregated.

One distinct advantage of the approach to clustering described here is the need to create atoms occurs only once per census period, which in the case of Canada is once every five years. Further, since atoms are immutable there is no need to recreate them for each clustering application, such as an atom-based custom clustering solution. Given that the majority of work in a clustering solution is the collection, preparation and segmentation of atom level data, additional clustering solutions, or "updates" to such solutions, can be achieved with much less repetition of effort. With the basic methodology reviewed, we can now describe how these methods were applied in the development of PSYTE Canada Advantage.

Data Preparation

The development of PSYTE Canada Advantage began with processing and defining Census 2001-based databases. The specific census variables that would go into the process were selected and defined. In a clustering process, the character of the input

Not to be cited without proper attribution.

data determines to a large extent the types of clusters that emerge in the cluster solution. 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 (Aldenderfer and Blashfield 1984 pp. 19 – 22).

Several important statistical issues were kept in mind as the input variables were selected. First, in a clustering technique that is a parametric method (e.g. K-means), all variables should be ratio data. However, in a non-parametric process (e.g. neural networks) nominal, ordinal and interval data can be included.

The second statistical issue is whether the candidate variables are statistically reliable, specifically that they are derived from a sufficient sample base as in the case of census variables derived from long-form questionnaires, and that the inherent variability is sufficient to provide statistical discrimination in the cluster solution. Results will be less reliable to the extent that some or all variables are not significant for each geographic unit being clustered. In general, the Canadian Census is an excellent source of reliable data since the data are collected at 100 percent and 20 percent samples. The same cannot be said for household list data, for example, that is sourced from commercial surveys, subscription lists and product registrations.

Third, every variable selected must also have a corresponding denominator or weighting variable. This requirement allows the data to be normalized with respect to its geographic level and provides for an accurate calculation of weighted means and standard deviations. Since all geographic units are not the same size in terms of area or population, the analyst must account for this by either calculating averages (e.g. income)

Not to be cited without proper attribution.

or percentages (e.g. age cohort composition). Not doing so biases the classification toward grouping geographies together based on their size rather than their true demographic profile.

Finally, considerable thought was applied to how the variables are weighted. For example, K-means clustering can use an explicit weighting scheme whereas neural net techniques generally use an implicit weighting scheme. One advantage of the neural net techniques as used here is they can handle more variables of similar character. Therefore, as described below, several variables were selected to represent each key demographic dimension in the system (Hartigan 1975 pp. 91 – 92, Aldenderfer and Blashfield 1984, pp. 21 – 22).

Cluster Dimensions

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 were used in the first stage. In the second stage, lifestyle and purchase behavior variables were developed and included in the processing. In the end, both census and non-census type variables provided dimensions to the clusters. The non-census variables were normalized to the geography through the calculation of geographic potentials. The primary census-demographic variable sets included: age, dwelling type, family structure, education, employment characteristics, immigration status, ethnicity, ancestry, religion, home language, income, industrial classification, geographic mobility, mode of travel to work, and occupation.

Not to be cited without proper attribution.

The following comments illustrate some of the content and measurement issues the analysts considered in using each of these variable sets:

Income

Four measures of income were included: 1) mean and median household income, 2) income size distributions of households, 3) sources of income expressed as a percentage of total income, and 4) income distributions by householder age.

Education

This category serves two objectives: identify the educational attainment of persons (may correlate with affluence or particular professional occupations) and measure current enrollment levels of the population to distinguish, for example, university neighbourhoods towns from other types of residential areas.

Collective Dwellings

Due to the concentrated nature of these unique populations – military personnel, university students, nursing home residents, and correctional facility inmates – it is important to identify these areas and essentially "set them aside" during the initial clustering process. Later, they can be identified and labeled appropriately.

Dwelling Type

Many personal and family attributes are captured, or at least implied, by dwelling characteristic or housing unit data. The principal ones are: size of dwelling (e.g. number of rooms or units), owner or renter occupancy, vacancy rate, housing vintage, and home value. Such data provide a rich source of cluster dimensionality as well as

Not to be cited without proper attribution.

descriptive attributes for profiles of residential areas as they indicate levels of affluence, predominance of family housing, concentration of apartments, age of settlements, and seasonality of occupancy.

Geographic Mobility

Geographic mobility includes a range of concepts including place (or country) of birth data, internal migration, length of residence, immigration status, and year of arrival. For example, identifying the classic Burgess 'Recharge Zones' helps determine neighborhood evolution or stability. Also, high levels of geographic mobility coupled with economic mobility in urban areas can be an indicator of gentrification.

Place of Work and Commuting

Place of work data determines the nature and extent of commuting for an urban or metropolitan area. This helps to characterize commuting flows, commuting times, methods of transportation, and patterns such as inter-urban, intra-urban or extra-urban transit. This is particularly important for distinguishing new suburban areas adjacent to older towns. Residents of the new suburbs are more likely to commute longer distances than residents of the older towns.

Mode of Travel

Not only is the mode of travel interesting in itself, but this concept provides important insights into settlement context. For example, walking to work may indicate mixed zoning (residential and business) or a higher level of urbanity when combined with the presence of rapid transit systems.

Not to be cited without proper attribution.

Employment

Three general statistics are covered by this category: percentage of persons employed or in the work force, the number of hours worked per week, and the number of weeks worked per year. Thus, not only does such data indicate the general employment level in an area, but the data are also indicative of the extent of full-time, part-time and seasonal employment.

Industrial Classification

The distribution of workers by the standard industrial classification of their employer provides insight into the "industry" or type of work in which persons are employed. This is one key discriminator for affluence, but also describes the economic structure of an area and the work interests of the population.

Occupation

Combined with industry, occupation indicates the range of skills and general compensation levels for the working population. Due to the large number of occupational categories available for analysis, only major occupational groups (15-20 occupations) were used. Occupations were also summarized into four categories: white collar, grey collar, blue collar, and services.

Age

The variables in this group provide essential cohort compositional indicators and permit insight into the age profiles of a neighbourhood. Age distributions can also indicate family structure, the presence of children, and the number of generations present in the community.

Not to be cited without proper attribution.

Immigration and Ancestry

These variables provide key information about the extent to which immigration, both recent and historical, help characterize a neighbourhood. Period of immigration provides insight into the "age" of ethnic neighborhoods and whether they are still being "recharged." Ancestry and country of birth provide additional cultural information.

Home Language

While home language can be seen to duplicate the statistical discrimination of neighborhoods offered by immigration and ancestry variables, it provides an important additional descriptor: cultural assimilation, both in terms of knowledge of official languages, and retention of traditional languages at home.

Household Structure and Family Status

Capturing data about the number of families per housing unit, family structure, marital status and presence of children provides a set of powerful indicators that relate to consumption behavior as well as to the dominant household composition in the neighbourhood.

The Clustering Process – Stage One

Once the database is set up and normalized, the actual clustering process can begin. MapInfo analysts used the two-stage methodology as described above. The first stage involved the application of proprietary neural network geo-statistical techniques to classify the 52,399 dissemination areas with 200+ census variables. In general, neural

Not to be cited without proper attribution.

network techniques, which involve pattern recognition in ways that mimic the human brain, proved to have outstanding capabilities for identifying patterns in socio-economic data.

A perennial issue with geodemographic clustering is the problem of outliers. While satisfactory clusters may be produced, concerns can remain about observations that are significantly different from the mean of the cluster across several dimensions. The issue is: Which is the best (most appropriate) cluster assignment of the outlier geographic unit? The use of "atoms" in the first stage of the clustering process minimized the occurrence of outliers in this case. The creation of several hundred atoms – smaller, preliminary clusters of DA's – effectively reduced the statistical likelihood of outliers.

Another issue that required attention during the first stage of clustering was the issue of homogeneity. In an idyllic world a clustering routine produces highly homogenous clusters in which the basic units are optimally similar to each other while their mutual dissimilarity vis a vis other clusters is maximized. Clearly, the real world is different and ultimately more interesting. PSYTE Canada Advantage is clustering system for neighborhoods, not individuals or households. Neighborhoods, like the people who inhabit them, are inherently heterogeneous. The issue is how to measure neighborhood heterogeneity. Neural network techniques are, in fact, uniquely able to measure not only the degree of homogeneity but also the specific combinations of socio-cultural dimensions that characterize a particular cluster's "heterogeneity." For example, many rural neighborhoods have been transformed by the presence of urban-oriented workers and their families. Likewise, some immigrant neighborhoods are characterized by interactions among families of different ethnicities and countries of origin. Moreover,

Not to be cited without proper attribution.

since social processes are not generally random, there is a significant likelihood that heterogeneous neighborhoods in one region will have characteristics of heterogeneity similar to neighborhoods in other regions. For example, the presence of multiple ancestries, in similar combinations, in high-rise residences is common in several urban neighbourhood clusters in Canada. In this exercise, the authors confirmed that geodemographic clustering is still applicable to the task of grouping neighborhoods by their similar characteristics despite their increasing diversity over time.

Hierarchical Clustering – Stage Two

After the "atoms" were created, based primarily on socio-economic and demographic variables along with selected measures of settlement context, the next stage used hierarchical clustering techniques to group the 200+ atoms into the final 65 clusters. In the second stage, lifestyle indicators from a large omnibus survey were combined with the geodemographic atoms for further clustering The proprietary hierarchical technique (based on Ward's technique, Malhotta 1996. pp. 678 – 679, Aldenderfer and Blashfield 1984. pp. 43 – 45, SPSS 2001a) used provided more precise control over the clustering process compared with straight "out of the box" methods found in statistical packages and allowed researchers to "craft" the clusters in a scientifically reliable way.

Prior to running the hierarchical process, however, a principal components analysis (PCA), a special implementation of factor analysis, was performed (Maxwell 1971, SPSS 2001b). 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. In

Not to be cited without proper attribution.

preparation for the hierarchical process to agglomerate atoms into final clusters, the analysts did not want too many variables to bias the process along certain dimensions. (Aldenderfer and Blashfield 1984. pp. 21). Thus, the PCA was used to provide meaningful components among intentional characteristics without the need for a large number of variables. The final steps of stage two involved running the hierarchical clustering process a number of times, examining results, and re-running in order to satisfy the original project specifications. A reporting mechanism had been developed which permitted extensive evaluation of final clusters with respect to product requirements.

Visioning the Clusters

Once the final 65 neighborhood clusters were established, and the analysts were content with their statistical reliability, the process of "visioning" the clusters began. Visioning is the process of naming and describing the clusters consistent with their underlying characteristics. Cluster names and descriptions must "ring true" for the general characteristics of each neighborhood but also for their unique identifiers. 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 some characteristics similarities with other clusters. Cluster descriptions, including maps, statistical profiles, and anecdotal highlights provide a "vision" of the final clusters.

The clusters are also classified by settlement context groups. These are: Urban, Suburban, Town & Exurban, and Rural. Within each settlement context type the clusters are ranked by household income to form the "major groups." This stage in the development process necessarily contains a large dose of subjectivity as the imaging of

Not to be cited without proper attribution.

the clusters reflects the creativity but also the background knowledge of the developers. The statistics have told their story, now the developers must "post hoc" impose a sense of place and people that tells a story for each neighbourhood that can be envisioned by end users. As discussed, the primary goal of the statistical and clustering process described here was to "let the data speak for itself." That dictum was followed rigorously such that the process could be repeated by others in a similar fashion using the same software infrastructure. However, the naming and visioning of clusters, given the marketing context of their end use, will necessarily be shaped by the team members and their particular creative sense.

PSYTE Canada Advantage provides a multidimensional framework that allows analysts to capture the complexity of Canadian consumer culture without having to manipulate literally thousands of census variables. Over the last half-century long strides have been made regarding the methodologies and technologies used to segment geodemographic data sets. One of the principal goals of this evolution has been the increased rigor with respect to the use of statistical models, thus migrating subjective human understandings to more reliable computational models. Simultaneously, the debate and interplay between hierarchical and non-hierarchical techniques has generated applications and processes that should lead to further advances. One promising advance, alluded to in the literature review, is the area of "auto-clustering." Auto-clustering approaches promise to remove all subjective input to the process and analyze data based strictly on their structure of variance (Rauber, et al. 2002). While clearly in early development, they hold some promise and could eventually relieve the researcher of all tedious decisions except for the most important of all: What data should be used in an a

Not to be cited without proper attribution.

priori segmentation system? In the end, that question is perhaps most influenced by purposes to which the cluster system will be put. As the following case study illustrates, cluster system can have an influence on several aspects of business decision-making in a consumer marketing context.

A Case Study: XYZ Company, a retailer specializing in home furnishings

Business issue: XYZ, an established retailer, had created an elaborate strategic plan to expand its business over the next three years. The new plan called for expanding the number of stores by a factor of three, from 25 stores to 75 stores. This plan was developed after an intense planning process involving managers, independent market planning consultants, and principal investors. The key locational question was where to develop the new stores given their current store network and knowledge of their primary customer target markets. XYZ managers had a sense that there were a significant number of potential locations but the key issue was how to make a sequence of site location and development decisions that would lower risk and maximize the probability of successful implementation of the plan.

A secondary but no less important issue was the quality and extent of knowledge the company had about their primary target markets, their best customers, and the strategic direction they had chosen for expansion. Essentially, they needed to understand their customers better in order to create a reasonable estimate of their market potential by region as well as within and beyond their existing store networks. This estimate would also depend on a competitive analysis and information about the plans of key competitors.

Not to be cited without proper attribution.

Plan Execution: A principals team was assembled which included XYZ managers in real estate, marketing, merchandising, and of course, research analysts with expertise in customer segmentation. The team realized that the first step to implementing the plan was to develop extensive information on their current customer base in order to eventually compare existing "best customers" and "best performing stores" with new estimates of market potential across an array of potential store locations. These first steps called for analysts to examine their existing customer research, collect new information as needed, and re-examine prior studies to confirm conclusions. The team examined in-store surveys, customer databases, and sales records by store to determine the extent of their existing information and to reveal any gaps.

The next step was to create a typical customer profile for each store which could be compared to the entire store network or to sub-groups (e.g. English-speaking versus French-speaking areas) within the existing network. For this work the analysts turned to a geodemographic segmentation system – PSYTE Canada Advantage. PSYTE profiles were created by geocoding customer databases for each store and running profile reports for trade areas that provided the absolute number of households with a store customer present, percent distribution by cluster and penetration index by each of 65 neighbourhood clusters. (See Bourgault for a sample online graphic.)

The analysis provided a good sense of which stores represented the best overall penetration of XYZ's traditional target segments as well as which stores attracted new segments identified in the strategic plan. All customer segments could be specifically measured against their distribution across Canada and their higher-than-average presence in selected metropolitan markets. Comparisons were made among stores and across

Not to be cited without proper attribution.

markets that highlighted stores well-positioned within their markets, stores whose sales were lagging relative to potential, and new geographic markets that represented opportunities for the company to expand.

The next step in the plan execution involved answering the questions: Where do we locate our next 50 stores? And, are there selected stores that should close in order to optimize the store network? To answer these questions, retail analysts can make use of site screening models, supportable store models, and market optimization models all based on PSYTE. In this case, a site screening model was implemented that used the results of the customer segmentation profiles and store performance analysis. Neighbourhood types (clusters) that represented the best customers as well as the best-performing stores were ranked. Trade area rules were established that, for example, indicated the typical size and drive-time polygons for the best stores. A database of potential retail sites (the latitude-longitude coordinates of street intersections) were evaluated and ranked based on the suitability for supporting new stores. The top sites were then evaluated on criteria such as traffic patterns, locations of competitors, and overall suitability for development in order to narrow the list down to a set of feasible candidates for further, on-site study.

The segmentation system played an important final role in the development of plans for specific merchandising concepts and "grand openings." For example, the PSYTE profiles used in the real estate screening model to determine core target groups became "creative tools" to generate advertising copy and direct mail content to be sent out prior to the grand openings. (See the Appendix to this paper for the capsule descriptions of each cluster that were used in this task.) Only those PSYTE clusters with

Not to be cited without proper attribution.

significant above-average propensities to shop at XYZ were selected to receive the direct mail piece. XYZ worked with their consultants to select a specific set of postal walks within the trade areas of the new stores that had the highest penetration of the target group. The content of the mailings included information on traditional store merchandise as well as new categories or items that were known to be attractive to the target audience.

Thus, the geodemographic segmentation system proved useful in the multi-stage implementation of XYZ's strategic expansion plan. Current stores were evaluated and ranked. Best customers were identified and quantified. New sites were selected for further study and eventually on-site development. Finally, store merchandising decisions and promotional strategies were implemented in a manner consistent with the original segmentation analysis. The company, in essence, made significant investment decisions while lowering overall risk with a thorough understanding of its current and potential customers.

Not to be cited without proper attribution.

Conclusion

Geodemography, as practiced over the past 30+ years, has not faded as some have claimed but has evolved significantly and continues to challenge researchers working in consumer market contexts. In particular, as the statistical tools of applied geography and demography advance, so the challenge of developing reliable cluster systems has moved to a new level. Mountains of data and super-fast computers inexorably require more rigorous models that can capture the underlying structures of consumer characteristics and behavior. Coupled with advances in retail analytics – including site screening models, untapped potential models, and spatial interaction models – geodemographic segmentation will surely remain an important tool. However, further progress will likely depend on multidisciplinary approaches in both the theoretical and applied spheres.

Not to be cited without proper attribution.

References

Aldenderfer, M. S. and R. K. Blashfield. 1984. Series: Quantitative Applications in the Social Sciences. "Cluster Analysis." Beverly Hills: Sage Publications. Pp. 19 – 22, 35 – 38, 43 – 50.

Brown, Nina. 2004 "Robert Park and Ernest Burgess: Urban Ecology Studies, 1925" University of California, Santa Barbara. Available online at http://www.csiss.org/classics/content/26

Hagenbuchner, M., A. Sperduti, and Tsoi Ah Chung. 2003 "A Self-Organizing Map for Adaptive Processing of Structured Data" in *IEEE Transactions on Neural Networks.* 14(3) 491 – 492.

Han, J., M. Kamber and A. K. H. Tung. 2001. "Spatial Clustering Methods in Data Mining." Pp. 191-200 in Miller, H. J. and J. Han. *Geographic Data Mining and Knowledge Discovery*. London: Taylor & Francis.

Hartigan J. A. 1975 *Clustering Algorithms* New York: John Wiley & Sons. pp. 84 – 86, 91 – 92, 102

Maxwell, A. E. 1971 *Factor Analysis As A Statistical Method.* Butterworths. pp. 19 – 20

Not to be cited without proper attribution.

Malhotta, Naresh K. 1996. *Marketing Research: An Applied Orientation.* Toronto: Prentice Hall. pp. 677 – 679.

Openshaw, S. and I. Turton, I. 1996. "A parallel Kohonen algorithm for the classification of large spatial data sets," in *Computers and Geosciences* 22, pp 1019-1026.

Parthasarathy Srinivasan. 2003. "Lecture Notes for CIS 694Z: Introduction to Datamining." Ohio State University. Available online at http://www.cis.ohio-state.edu/~srini/694Z/

Rauber. A., D. Merkl, and M. Dittenbach. 2002. "The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data." in *IEEE Transactions on Neural Networks.* 13(6) Pp. 1331-1333.

Schürmann, J. 1996. Pattern Classification: A Unified View of Statistical and Neural Approaches. New York: John Wiley & Sons, Inc.

Shenk, D. 2003. "Watching You," in *National Geographic Magazine*, November. Available online at http://magma.nationalgeographic.com/ngm/0311/feature1/index.html

Not to be cited without proper attribution.

Simula O., J. Vesanto, and P. Vasara. "Analysis of Industrial Systems Using the Self-Organizing Map" in *Proceedings of the International Conference on Knowledge-based Intelligent Systems (KES'98)* pp. 7. Available online at http://www.cis.hut.fi/projects/ide/publications/fulldetails.html#simula98kes

SPSS Inc. 2001a "Cluster.pdf" in software documentation for SPSS 11.0 for Windows. Chicago. pp 1, 2, 11 – 13

SPSS Inc. 2001b "Factor.pdf" in software documentation for SPSS 11.0 for Windows. Chicago. pp 1, 10 – 12, 19

Yeates, Maurice. 1990 *The North American City.* Harper & Row, Publishers Inc. New York. pp 110 – 111.

Not to be cited without proper attribution.

Appendix A – PSYTE Canada Advantage Cluster Descriptions

PSYTE Code	Cluster Name	settlement context	urban core/fringe	2004 average household income	Income Grou	Percent of Households
1	**Canadian Elite**	Urban	fringe	$ 250,776	Elite	0.7%
4	**Professional Duets**	Urban	core	$ 131,638	Elite	0.7%
12	**Urban Gentry**	Urban	fringe	$ 95,587	Upscale	2.1%
2	**Suburban Affluence**	Suburban		$ 165,704	Elite	0.5%
5	**Family Comfort**	Suburban		$ 120,189	Upscale	1.6%
7	**Euro Traditionals**	Suburban		$ 100,717	Upscale	0.7%
10	**Suburban Growth**	Suburban		$ 97,344	Upscale	1.2%
11	**Asian Heights**	Suburban		$ 96,369	Upscale	0.9%
3	**Exurban Estates**	Town and Exurban		$ 145,148	Elite	0.4%
6	**Commuter Homesteads**	Town and Exurban		$ 104,787	Upscale	0.9%
9	**Towns with Tempo**	Town and Exurban		$ 98,538	Upscale	1.1%
13	**Kindergarten Boom**	Town and Exurban		$ 94,669	Upscale	0.6%
16	**Bicycles and Bookbags**	Town and Exurban		$ 90,857	Upscale	1.2%
14	**Cruising Commuters**	Suburban		$ 94,047	Upscale	1.7%
15	**Quebec Upscale**	Suburban		$ 93,193	Upscale	1.8%
19	**Family Crossroads**	Suburban		$ 78,404	Upper Middle	1.5%
20	**Row House Streets**	Suburban		$ 74,905	Upper Middle	0.6%
24	**Satellite Suburbs**	Suburban		$ 71,596	Upper Middle	1.9%
18	**Exurban Wave**	Town and Exurban		$ 82,025	Upper Middle	1.7%
23	**Town and Country**	Town and Exurban		$ 72,445	Upper Middle	2.3%
17	**Young Technocrats**	Urban	core	$ 82,336	Upper Middle	1.1%
21	**University Enclaves**	Urban	fringe	$ 73,459	Upper Middle	2.3%
22	**Upbeat Blues**	Urban	fringe	$ 72,723	Upper Middle	0.9%
25	**Urban Promise**	Urban	fringe	$ 70,690	Upper Middle	2.2%
26	**South Asian Corners**	Suburban		$ 70,277	Upper Middle	0.3%
27	**Quebec Melange**	Suburban		$ 70,252	Upper Middle	1.7%
28	**Conservative Homebodies**	Suburban		$ 69,134	Middle	2.0%

Not to be cited without proper attribution.

31	**Quebec Rows**	Suburban		$	64,753	Middle	2.0%
8	**Primary Pursuits**	Rural		$	99,126	Upscale	0.8%
29	**Agrarian Heartland**	Rural		$	66,888	Middle	2.0%
30	**Northern Lights**	Rural		$	66,080	Middle	0.5%
33	**Village Views**	Rural		$	63,288	Middle	2.0%
35	**New Frontier Families**	Rural		$	62,611	Middle	1.7%
40	**Quebec Farm Families**	Rural		$	59,716	Middle	3.8%
32	**Quebec Urbanites**	Urban	fringe	$	63,670	Middle	1.0%
34	**Workers' Landing**	Urban	fringe	$	63,023	Middle	0.6%
36	**Pacific Fusion**	Urban	core	$	61,516	Middle	0.7%
38	**Sushi and Shiraz**	Urban	core	$	61,118	Middle	1.8%
39	**Hi-Rise Sunsets**	Urban	core	$	60,646	Middle	1.7%
41	**New Canada Neighbours**	Urban	core	$	58,654	Middle	0.6%
45	**Urban Vibe**	Urban	fringe	$	51,057	Middle	0.7%
37	**Village Blues**	Town and Exurban		$	61,127	Middle	3.9%
42	**Senior Town**	Town and Exurban		$	52,919	Middle	0.4%
47	**Middletown Mix**	Town and Exurban		$	50,499	Middle	2.6%
43	**Suburban Hi-Rise**	Suburban		$	52,404	Middle	0.8%
55	**Asian Mosaic**	Suburban		$	44,779	Lower Middle	0.5%
44	**Highland Havens**	Rural		$	52,222	Middle	2.3%
46	**Open Country**	Rural		$	50,872	Middle	1.0%
48	**Cabins and Cottages**	Rural		$	50,321	Middle	1.9%
51	**Peaceful Pastures**	Rural		$	46,582	Lower Middle	3.7%
54	**Quebec Rural Blues**	Rural		$	45,957	Lower Middle	4.2%
61	**First Peoples**	Rural		$	39,769	Low	0.6%
49	**Blue Collar Stride**	Urban	fringe	$	49,382	Lower Middle	2.4%
50	**Euro Quebec**	Urban	core	$	49,128	Lower Middle	1.1%
53	**Urban Bohemia**	Urban	fringe	$	46,158	Lower Middle	1.0%
56	**Quebec Walk-Ups**	Urban	core	$	44,754	Lower Middle	2.0%
57	**Hi-Rise Melting Pot**	Urban	core	$	44,410	Lower Middle	0.8%
59	**Service Crew**	Urban	core	$	43,458	Lower Middle	2.5%

Not to be cited without proper attribution.

52	**Elder Harbour**	Town and Exurban		$	46,314	Lower Middle	3.0%
58	**Second City Renters**	Town and Exurban		$	43,517	Lower Middle	1.9%
60	**Quebec Town Elders**	Town and Exurban		$	43,269	Lower Middle	2.5%
62	**Blues in Motion**	Urban	fringe	$	36,080	Low	2.6%
63	**Quebec Seniors**	Urban	core	$	34,660	Low	0.8%
64	**Metro Medley**	Urban	core	$	33,485	Low	0.8%
65	**Quebec Urban Stress**	Urban	fringe	$	32,078	Low	2.3%

Not to be cited without proper attribution.