

TECHNICAL DOCUMENTATION

PRIZM5 2018

APRIL 2018

The creation of PRIZM5—the latest edition of Environics Analytics’ groundbreaking national PRIZM_{CE}/PRIZM_{C2} segmentation systems—involved more than a year of planning and development. It is a system that categorizes every Canadian neighbourhood and postal code into one of 68 distinct lifestyle segments based on the characteristics of households. PRIZM5 examined 10 years of neighbourhood formation, change and, in numerous cases, stability to better understand the demographic, socioeconomic and behavioural characteristics of consumers.

PRIZM5 is developed by a team of experts familiar with the demographics and geography of Canada at both the city and regional levels. Using authoritative data from recognized suppliers like Statistics Canada, Canada Post, Canada Revenue Agency, Equifax, TomTom and others, the system is built from the ground up.

The vast majority of targeting and marketing in Canada is still done on the basis of age, sex and income. Examples of typical target groups using these methods are:

- Women aged 18-34 for fashion (clothing)
- Men aged 35-50 for golf and tennis clubs
- Households with income over \$150,000 for prestige vehicles

Geodemographic systems permit a refinement of this typical targeting method. Small neighbourhoods are assigned to segments based on similarities of demographic attributes and general lifestyle behaviours. This approach is a well-established method for segmenting and identifying target groups for numerous products and industries.

DATA

The primary foundation for any geodemographic system is high-quality demographic data. The decision regarding which variables to include in the final clustering is a matter of science and art. The science element captures variables that are significant in their ability to differentiate neighbourhoods in ways that are important for marketing. The art component falls into two categories: a) understanding that users of the final system expect specific variables to be included, and b) determining the final selection of weights for each of the variables. When going to market, customer expectation and demand also play an important role. At the same time, there is a need to respect and adhere to thorough scientific methods. Successfully balancing these many requirements makes a product credible and effective.

The number of variables in the final clustering should be kept to a minimum (for scientific parsimony) yet, at the same time, must include all important demographic characteristics. There is no known acceptable number of variables to include in clustering. It is up to the experienced analyst to select the optimal set of variables and experiment with weights to use in the clustering process.

Creating PRIZM5 begins with CensusPlus 2016 which is developed by Environics Analytics and based on Statistics Canada’s Census. The core data are available at the dissemination area (DA) level, the smallest unit of geography for which any significant demographic and socioeconomic data are released. There are 56,590 DAs in Canada. CensusPlus contains nearly 1,000 variables for each of these DAs covering numerous themes.

We take CensusPlus 2016 and combine it with DemoStats 2018, EA's estimates of current year demographics and socioeconomic characteristics at the neighbourhood level¹, to select the final demographic and socioeconomic variables. The variables are categorized in the following 18 themes:

Age	Dwelling Value
Household Size	Mother Tongue/Home Language
Marital Status	Ethnic Origin
Households with Children	Visible Minority
Migration	Aboriginal Identity
Immigration	Education
Dwelling Type	Labour Force/Occupation/Work Place
Dwelling Tenure	Mode of Transport
Dwelling Period of Construction	Income

In addition to the core demographic and socioeconomic data, other data are used as inputs into PRIZM5's creation. Data describing the settlement context—the geographic location of neighbourhoods—is fundamental to understanding where the resulting segments are situated geographically. Are the segments predominantly found in large urban centres, small suburban towns or sparse rural communities? Proximity to major retail centres is also used in our research as a measure to further classify established urban cores differently from suburban and exurban neighbourhoods.

Another important source of input data is supplied by our sister company, Environics Research. Researchers there conduct an annual survey asking Canadians their opinions, views and “social values” regarding numerous societal issues. Using the data they collect, they create “social constructs” that identify common trends in views and attitudes among Canadians.

In addition, we use aggregated, anonymized small-area credit data from Equifax Canada. These data measure credit worthiness, credit usage and credit default rates. Additional variables capturing the financial theme are also included in creating PRIZM5.

From the above data, more than 80 variables are selected, with at least one variable for each of the above themes.

¹ For more information on the development of DemoStats 2018, refer to the DemoStats 2018 Methodology.

PROCESS

At the outset, EA analysts determined that the PRIZM5 segmentation system was going to consist of three different products. The segments had to be available for populated residential DAs and postal codes, and they had to operate at two levels: one with a conventional number of segments between 60 and 70 and a more detailed version with more than 120 segments.

This approach was divided into three stages. Stage one involved the creation of a segmentation system consisting of 150 very focused segments that we called “atoms.” These atoms were produced using a combination of all the data excluding Social Values data. Stage two focused on the addition of Social Values data to the 150 atoms and the aggregation of 150 atoms to the final number of segments (the end result was 68). Below is a general overview of the stages with a brief description of each step.

STAGE ONE

The initial phase of stage one was the creation of a set of variables that captured the settlement context of the DA. Settlement context is a scaled measure of urbanity, from the dense urban downtowns of large cities to the most sparse, uninhabited rural parts of the countryside. These variables serve as key variables in the initial segmentation process.

The next phase was the assessment and selection of CensusPlus 2016 and DemoStats 2018 variables from the more than 1,400 variables available to be used in the creation of the atoms. We selected variables that were known from experience or believed to be significant for the purposes of differentiation among the DAs.

The final step involved the selection of the clustering algorithm for creation of the atoms. There are numerous algorithms available for cluster analysis, and the method used was critical to the success of the final result. Several methods were tested and used to generate solutions during this stage.

The algorithm EA selected is K-medians, an iterative partitioning approach commonly used for these types of applications. Its greatest strength is its ability to find similar patterns that maximize within segment uniformity while differentiating between all the identified segments. Additionally, the resulting segments are not as influenced by extreme values (“outliers”) as many other traditional methods.

To determine the best algorithm, we tested thousands of solutions, night after night for a period of weeks. Every run was informed by the previous one, resulting in changes to variables and in variable weights with each subsequent run. The best solution of each run, the one that maximally differentiated the generated segments, was examined and systematically tested. The best solution, and the one that was selected, was the one that offered the greatest discrimination of segments against actual consumer behaviour (more on this later). Finally, we produced what we consider to be an optimal 150-atom segment solution.

STAGE TWO

The focus of this stage was linking Social Values to the atoms and the aggregation of atoms to the final 68 segments with Social Values in them. The Environics Research data were linked to each atom taking into account sample size issues. Using EA's estimated demographic and socioeconomic data, along with settlement context, financial credit data and Social Values, the atoms from the DA system were aggregated to between 60 and 70 larger segments using several clustering algorithms.

Analysts identified many segment "solutions" by applying different weights to a variety of variable subsets, the subsets derived from the sources of data previously mentioned. In aggregating atoms to the final set of segments, we considered solutions that ranged from 60 to 70 segments. We resolved to look for fewer than 70 segments because a larger number of segments would provide diminishing returns to achieving discrimination and marketers have limited ability to deal with large numbers of segments. In reviewing the solutions that were automatically generated through the clustering processes, 68 segments seemed to offer the greatest variety in neighbourhood and Social Values types while meeting our minimum cluster size as represented by the number of Canadian households assigned to it (0.50 percent of Canadian households). The 150 atoms nest perfectly within the final 68 segments.

TESTING SEGMENTS

Each of the solutions is tested using a variety of data supplied by several of our significant data partners and an analysis of key products and services. From the thousands of clustering runs, three solutions emerged as leading contenders.

Key survey providers for the testing exercise were:

- Vividata (formerly Print Measurement Bureau)
- Numeris
- AskingCanadians™ (Social Media, Mobile and eShopper surveys)
- IHS Automotive
- Select client data from different industries

More than 3,000 variables were selected from the various surveys covering such topics as category and product usage, media preferences, leisure activities and attitudes. The survey data were aggregated to the 68 segments and each variable was compared to its Canadian average. A review of each segment's demographics, socioeconomics, settlement context and behavioural survey variables served as the method for analyzing and comparing the numerous cluster solutions.

These solutions were tested against one another to help identify the single best segmentation system solution. In reviewing the solutions we examined the following for each of the 68 segments:

- 1) Demographic reports with more than 400 variables summarized at the segment level showing percent of segment having the attribute and an index showing its relation to the Canadian average.

- 2) Geographic maps showing the segment distribution and whether the segment was concentrated in a few markets or dispersed across the country, including very detailed maps of major cities.
- 3) Survey reports with selected variables with an index against Canada for each segment on a huge selection of category, product, behavioural and attitudinal variables from among the data.

STAGE THREE

With the 68 segments finalized at the DA neighbourhood level, the next task was to assign all residential postal codes to the final solution of 150 atoms and its 68 segments. This stage involved combining the DemoStats 2018 data with the Equifax credit data, all at the postal code level, to complete the assignment.

A set of demographic and socioeconomic variables were selected from the more than 500 available in DemoStats. Added into the mix were the settlement context data that were assigned to all postal codes (based on the DA they fall within) and a small set of variables from the Equifax data. These data were all assembled for the complete roster of residential postal codes.

We then created 150 cluster centroids, the statistics reflecting multi-dimensional segment profiles—the basic building blocks of segments—using the atoms created at the DA neighbourhood level. While we had centroids from the DA already created, we also developed a new version using only postal code-level data. Several versions of the centroids were created and tested to ensure they captured the fundamental characteristics that describe each atom at the DA level. Did family-based segments have the correct ages of children? Were culturally diverse segments showing high concentrations of the relevant groups? Were urban segments found in urban areas and rural segments in rural areas?

One of the many versions of centroids was selected because it depicted the 150 atoms best. Following this, all postal codes were assigned to the closest atom based on statistical proximity. The distance measure used guaranteed the optimal assignment for each postal code using all selected variables.

In addition, there were a few final manual adjustments to the automated cluster solutions. These adjustments were made to preserve, as much as possible, the settlement context structure of Urban, Suburban, Exurban, Town and Rural. Other important considerations in the clean-up phase ensured that the wealthiest segments were captured appropriately, that key culturally diverse segments were identified correctly as speaking dialects of Chinese and South Asian languages, and that francophone segments had a minimum number of French-speaking populations.

RESULT

PRIZM5 is a whole new set of geodemographic segments for Canada reflecting the most recent and reliable data. There are 68 segments, made up of 150 atoms, which capture all aspects and dimensions of the Canadian landscape. It is available for both dissemination areas and postal codes.

SOCIOECONOMIC STATUS INDICATOR (SESI)

With the final segmentation system created, we had to decide how to number and rank the segments. A proprietary score was developed to characterize each segment using a Socioeconomic Status Indicator (SESI). This SESI score reflects a variety of factors such as average household income, discretionary income, education attainment level, value of private dwellings, average net worth and household size.

As a result, a blue-collar, high school-educated segment whose residents earn above-average incomes may rank lower on the SESI ladder than an educated, up-and-coming young segment whose residents have just average household incomes. A segment with an older population, many of whom are on fixed incomes, may rise in the ranking if their net worth is significant. And a segment earning \$120,000 on average will rise or fall in the ranking depending on whether the household is composed of dual-income couples or families with several young children.

The 68 segments have been ranked from 1 to 68 on the SESI scale, with 1 classified as the highest. Because this ranking reflects more than income alone, most of the segments have a SESI score that is different from their average household income ranking.

SOCIAL AND LIFESTAGE GROUPS

The 68 PRIZM5 segments were assigned to one of 19 Social Groups and 14 Lifestage Groups. The Social Groups consider the urban-rural context, home language (English, French and non-official), affluence, family status, age of maintainer and ethnicity. Each segment was assigned to one and only one Social Group. The Social Groups reflect various groupings, patterns and trends. A critical issue concerned the urban-rural dimension, which is neither linear nor one-dimensional. Each segment was assigned to one of five settlement types for the purpose of forming the Social Groups: Urban, Suburban, Exurban, Town or Rural. In general, urban segments are found in large- and medium-sized cities. Suburban segments tend to consist of communities located on the outskirts of cities and can often be found as the main section of smaller cities and larger towns. Exurban areas are most commonly found surrounding the suburbs of the eight largest cities in Canada. They include new subdivisions, low-density environments and towns within the outer commuter zones of cities. Town neighbourhoods are found in smaller towns across the country, except those towns that are located in exurban regions. Rural neighbourhoods reflect areas that are smaller than towns and include very small towns, villages, hamlets and rural farms and isolated areas.

The final set of segments had many francophone-based segments, a variety of culturally diverse segments and many segments that represent important combinations of age, lifestage and family status—from young singles living on their own up to widowed seniors in apartments. These were essential inputs into the creation of the Social Groups. The ranking of Social Groups is based on average income (not a SESI ranking). Groups have a letter and number combination. The letters U, S, E, T or R, represent Urban, Suburban, Exurban, Town or Rural, while the numbers indicate income, with 1 being the highest average income for the groups and 7 being the lowest.

The Lifestage Groups categorize household composition by assessing the presence of singles, couples and families. The major grouping divides the 68 segments into Young, Family and Mature. These groups are then

further subdivided by analyzing the commonality among the segments. The Young group is divided into two based on singles versus starter families. Families are split into four sets based on the age of children: the very young, tweens, teens and twenty-somethings. These groups are then further subdivided based on ethnicity, age of parents, language and affluence. This classification results in nine Family groups. The Mature group is divided into three based on age and geography.

ANNUAL UPDATE

Each year, when DemoStats is completed, PRIZM5 is updated. The update reflects the most recent estimates of Canada's demographic and socioeconomic characteristics, along with updated Social Values data and financial credit data from Equifax. Both postal code and dissemination area PRIZM5 assignments are reassessed given the up-to-date estimates.

We recognize that changes and shifts occur in Canada's demographic landscape on a daily basis. However, we are only concerned with large systematic change that can be measured. During the development of DemoStats, we examined the themes that are most important, such as age structure, income distribution, housing stock and tenure, diverse populations, family structure and the continuing urbanization of Canada.

For the 2018 update, our objective was to retain existing PRIZM5 segment assignments, except in cases where solid empirical evidence indicated a neighbourhood has changed significantly. We wrote routines to look at data compiled for each postal code and DA, calculating the "fit" to all segments to confirm any significant change. For those areas whose previous year's segment assignment was no longer the best fit, we changed the assignment. And, of course, new postal codes were assigned to the segment that was the best fit based on measuring distance to all segments.

In the 2018 update, 21 percent of postal codes underwent a change in their PRIZM5 assignment—a shift that affected 21 percent of households. The three most stable segments—that is, those that experienced the least geographic change—were Traditional Town Living (47), Exurban Homesteaders (36) and Country Acres (35). The three segments that experienced the greatest geographic change were First Nations Families (57), Jeunes et Actifs (44) and Enclaves Multiethniques (49). Meanwhile, the three segments that experienced the highest growth in households were Traditional Town Living (47), Our Time (39) and Exurban Homesteaders (36); the three segments that declined the most in households were First Nations Families (57), Wide Open Spaces (40) and Heritage Hubs (15). But it is important to note that, even with these changes, PRIZM5 can be described as a very stable segmentation system, reflecting an expected level of geodemographic change over the last year.