ENVIRONICS
ANALYTICS

TECHNICAL DOCUMENTATION
DEMOSTATS 2018

APRIL 2018

WHAT IT IS

DemoStats is a database of estimates and projections for a comprehensive set of demographic and socioeconomic attributes about the Canadian population. Built with the industry's most comprehensive set of data sources and leading modelling techniques, it consists of 760 variables across 42 demographic and socioeconomic categories. DemoStats provides estimates and projections for 2013, 2018, 2021, 2023 and 2028. The reference date for DemoStats is July 1, meaning that all statistics are estimates as of that date of the relevant year (which is consistent with the reference date used by Statistics Canada for their series of inter- and post-censal population estimates and projections).

DemoStats variables are available at the six-digit postal code level (FSALDUs) for current-year estimates and at the dissemination area (DA) level for future-year projections. DemoStats also provides historical estimates from five years ago based on the same methodologies to ensure accurate trend analysis. DemoStats is created using innovative methods that combine econometric, demographic and geographic models. And it employs a variety of data sources, including the latest and historical Census data, current economic indicators, post-censal estimates from federal and provincial governments, immigration statistics and economic data such as building permits. DemoStats features variables on population, family structure, household size and type, ethnic diversity, labour force participation and income—including both averages and distributions.

The list of postal codes used for this release of DemoStats is the roster of valid residential postal codes as of June 2017. This postal code roster was the most recent available at the time this vintage of DemoStats was developed. We produce DemoStats for the geographies listed in Table 1.

Table 1: Release geographies

Geography	Geo Abr.	Geography Count
National	CAN	1
Provinces / Territories	PR	13
Census Divisions	PRCD	293
Census Subdivisions	PRCDCSD	5,162
Census Metropolitan / Census Agglomeration Areas	CMACA	157
Census Tracts	CMACT	5,721
Aggregate Dissemination Area	PRCDADA	5,386
Dissemination Areas	PRCDDA	56,590
Federal Electoral Districts	PRFED13	338
Forward Sortation Areas (TomTom Q4 2017)	FSAQ417	1,648
FSALDU - Residential only	FSALDU	758,623

DemoStats 2018 contains all 760 variables available for the current year for all levels of geography listed in Table 1. A smaller subset, 486 variables, is available for historical and projection years. These variables cover 22 of the demographic and socioeconomic categories, and they represent the core dimensions of the Canadian population. The 20 non-projected demographic categories are not projected into the future because we do not believe that there is enough data available to reliably project these variables beyond the current year.

Table 2: DemoStats categories by count of variables and whether the theme is projected

Category	Count	Projected
Code	1	
Basics	19	Y
Total Population by Age	21	Y
Male Population by Age	21	Y
Female Population by Age	21	Y
Total Household Population by Age	21	Y
Male Household Population by Age	21	Y
Female Household Population by Age	21	Y
Households by Maintainer Age	10	Y
Households by Size of Household	9	Y
Households by Household Type	8	Y
Population 15 Years or Over by Marital Status	9	Y
Census Families by Family Structure	32	Y
Census Family Households by Family Structure	31	Y
Total Children At Home by Age	10	Y
Household Population by 5-Year Mobility	3	N
Occupied Private Dwellings by Tenure	4	Y
Occupied Private Dwellings by Period of Construction	8	N
Occupied Private Dwellings by Structure Type	12	Y
Occupied Private Dwellings by Condo Status, Tenure and Structure	16	N
Households by Income (Constant Year)	21	Y
Households by Income (Current Year)	21	Y
Household Population by Income (Current Year)	4	N
Household Population 15 Years or Over by Educational Attainment	9	Y
Household Population 25 to 64 Years by Educational Attainment	9	N
Household Population 15 Years or Over by Labour Force Activity	8	Y
Household Population 15 Years or Over by Occupation	14	N
Household Population 15 Years or Over by Industry	24	N
Household Population 15 Years or Over by Place of Work	7	N
Household Population 15 Years or Over by Method of Travel to Work	8	N
Household Population by Religion	13	N
Household Population by Visible Minority Status	15	Y
Household Population by Aboriginal Identity	3	N
Household Population by Knowledge of Official Language	5	N
Household Population by Mother Tongue	44	Y
Household Population by Language Spoken Most Often At Home	44	N
Household Population by Total Immigrants and Place of Birth	98	Y

Household Population by Recent Immigrants and Place of Birth	61	N
Household Population by Period of Immigration	8	N
Household Population by Age at Immigration	9	N
Household Population by Generation Status	4	N
Household Population by Citizen Age	17	N
Household Population by Non-Citizen Age	17	N

HOW IT WAS BUILT - KEY DATA SOURCES

Official statistics

Census 2016 for selected demographic categories

Census 2011, NHS 2011

Census 2006

Environics Analytics products

CensusPlus 2016

CensusPlus 2011

AdjustedCensus 2006

ePCCF

Other Statistics Canada products and administrative data

Statistics Canada Inter- and Post-Censal Demographic Estimates

Statistics Canada Components of Demographic Change

Statistics Canada Population Projections

Statistics Canada Labour Force Survey (LFS)

Statistics Canada Survey of Labour and Income Dynamics (SLID)

Canadian Revenue Agency (CRA) tax-filer and tax family data

Canadian Immigration Council (CIC)

Canadian Mortgage and Housing Corporation (CMHC)

Custom-ordered Census, CRA and NHS tables

Other data sources

Bank of Canada

Oxford Economics

Provincial population estimates and projection

Regional and municipal planning documents for select markets

Building permits for select markets

Canada Post Householder Elite (May 2017 – valid for June 2017)

Equifax

A large reliable household survey summarized to the postal code level

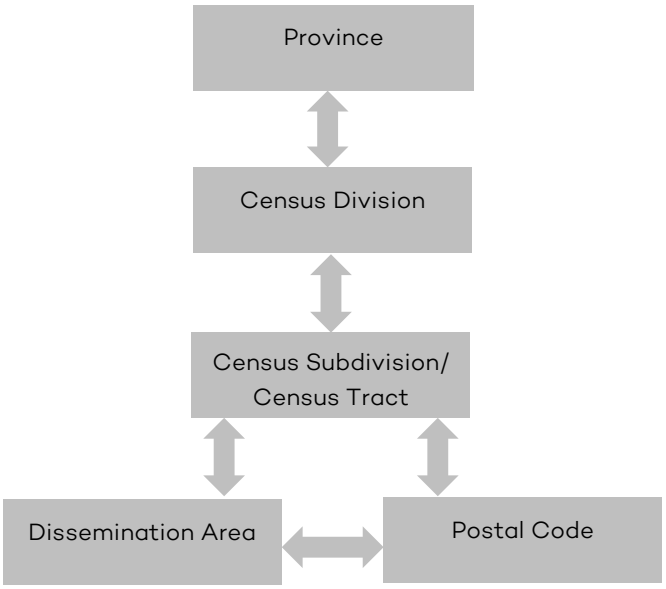
Land use and infrastructure files from various sources

HOW IT WAS BUILT – MODELLING FRAMEWORK:

TOP DOWN, BOTTOM UP

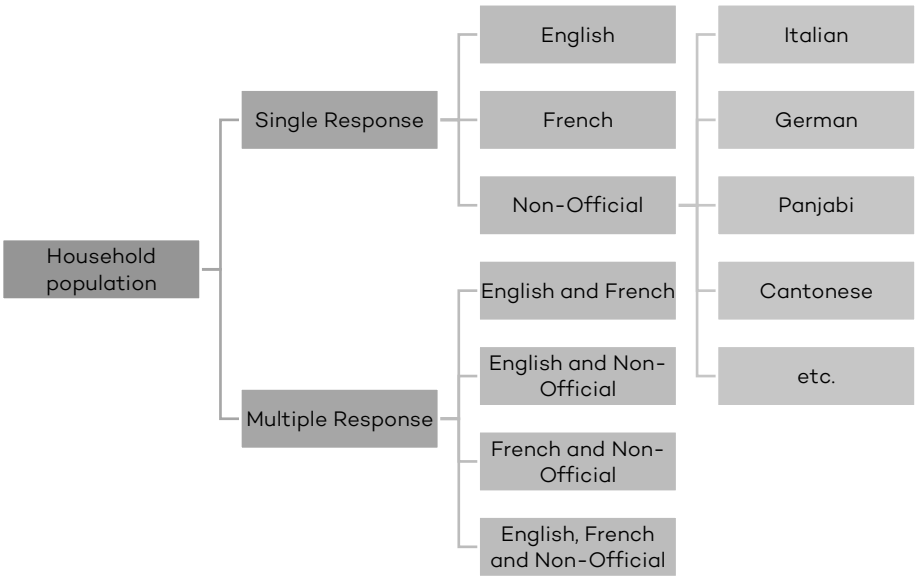
Our modelling framework works from the top down and then from the bottom up. This is true from a geographic perspective and for demographic categories. We start by modelling variables at the provincial level and work our way down to lower levels of geography. As we work our way down, we allow for information at lower levels of geography to affect higher levels of geography (see Figure 1). This feedback loop lets us refine our estimates and projections—and make the data consistent—at all levels of geography. Our estimates also conform to known control totals from authoritative external data sources.

Figure 1: Diagram of top-down, bottom-up approach for geography



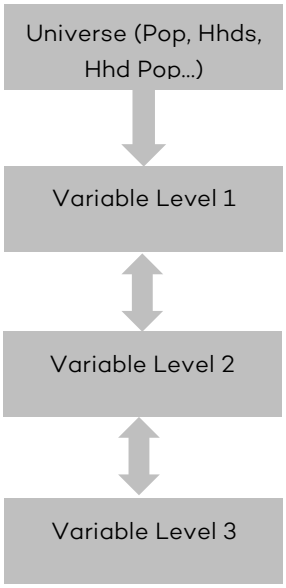
Similarly, many demographic and socioeconomic categories in DemoStats have hierarchical structures. The variable hierarchies define groups of variables that sum to other variables. For example, the theme "household population by mother tongue" has four distinct levels in its variable hierarchy (see Figure 2).

Figure 2: Diagram of household population by mother tongue variable hierarchy



Where these variable hierarchies exist, we use the same top-down, bottom-up approach. We model the highest order variables first. Then we subdivide those high-level variables across more detailed attributes. We allow information from the detailed attributes to feed back into the estimates and projections of the higher-order variables (see Figure 3).

Figure 3: Diagram of top-down, bottom-up approach for a variable hierarchy



KEY UNIVERSE

The most important variables used in developing DemoStats are households and total population. From these two variables, all other attributes are derived. We start building DemoStats by modelling total population by age and sex at the provincial (PR) and census division (CD) levels of geography using a sophisticated demographic method known as cohort component modelling. These cohort component models use the Statistics Canada inter- and post-censal population estimates[1] and components of change as their fundamental input. Using the cohort component framework, we are able to generate multiple scenarios at the PR and CD levels. The methodology is designed such that the CDs sum the PRs. Finalized CD estimates and projections are allocated to the census subdivisions (CSDs) based on historical data and CHMH housing starts, as well as regional and municipal development plans.

CSD population by age is used to calculate household counts. Household counts are calculated based on historical headship rates along with other factors. Headship rates specify the rate at which certain age cohorts form households. From the CSD level of geography, we allocate households to census tracts (CTs) and dissemination areas (DAs). The allocation from CSD to CTs and DAs leverages a variety of inputs: historical census data, geographic potentials, geographic gradient, land use, building permit data and, in some cases, official planning documents. These inputs allow us to triangulate the best areas for growth and allocate households in a reasonable manner.

Independent of DA households, our team creates household estimates at the postal code level using Canada Post HouseElite deliverable address counts along with data from Equifax. Equifax provides a household concept based on credit files for a given geography. Once the DA and postal code household counts are created for the current year, we use a proprietary fuzzy match-up file to compare the independent household estimates. The fuzzy match-up file provides a robust and more detailed set of linkages between DA geography and postal codes than a PCCF. The fuzzy match approach assumes[2] that many-to-many relationships are possible for DAs and postal codes in both urban and rural areas. A sophisticated set of simulation and optimization algorithms is used to determine the best household estimate for DAs and postal codes. The resulting household counts for the DA and postal code are consistent when translated through the fuzzy match-up. This process does not favor either household estimate. It seeks to find the best possible solution given both estimates and all other available information. The finalized, current-year household estimates are used as the base for the 3-, 5- and 10-year projections.

Given the final households counts for DAs, we estimate the population by age, sex and all other key universe variables. The estimates of these universes use CensusPlus 2016 as fundamental inputs to estimate the 2016 DemoStats equivalent. Various demographic and statistical techniques are used to model these universe

[1] The post-censal estimates provided by Statistics Canada use the 2011 Census as their fundamental benchmark. These statistics are adjusted for the net-undercoverage rates for the 2011 Census, are brought forward to the July 1, 2011, reference date and include imputes for non-responding Indian reserves. Estimates for inter-censal years are created based on the change between censuses, net-undercoverage rates and administrative data sources. Post-censal estimates are created based on administrative data and can be one of three types: first preliminary, second preliminary and final. The first preliminary estimate is a first estimate for the most recent year in the data series. The second preliminary estimate is a revised estimate based on additional data for the second most recent year in the data series. All other years in the data series are final estimates and are not revised until new census data become available. It is important to note that Statistics Canada revises its estimates before finalizing them.

[2] Historically, the standard practice in the industry is to assume that a single postal code can only link to one DA.

variables into the future. In the case of population by age and sex, the higher level estimates and projections created earlier in the development process are used as controls for the small area-level data.

DEMOGRAPHIC CATEGORIES

To estimate and project the 42 demographic categories, we use a wide array of methodologies, including: historical trend projection, various regression techniques, mathematical optimization and machine learning. For each category, we use the best approach available given our data sources and our understanding of the dynamics that drive the category. Additionally, we modify our methodological approach for different levels of geography. Methods are adjusted based on level of geography for two reasons:

1) Most external data sources are only available at higher levels of geography;
2) It is well documented that the correlation and variance structures observed for a phenomenon at one level of geography may be quite different at other levels of geography. This is known as the modifiable area unit problem (MAUP).

We use our CensusPlus 2016 or CensusPlus 2011 database as our fundamental starting point for all categories. Where authoritative data sources are available, we use them to bring the Census data forward to the most recent year available from the data sources. We also use external data sources to calibrate models and build relationships between variables and categories in the DemoStats database. Regardless of the approach, all DemoStats categories use the top-down, bottom-up approach described above.

KEY THEMES FROM CENSUSPLUS 2016 USED IN 2018 DEMOSTATS

- Pop by Age and Sex
- Household Population by Age and Sex
- Household Maintainers by Age
- Dwellings by Structural Type
- Dwellings by Tenure
- Dwellings by Period of Construction
- Household Size
- Marital Status
- Households by Type
- Families by Family Structure
- Children at Home by Age
- All of other demographic categories will be transitioned to using 2016 Census data as in the DemoStats 2019 release.

POSTAL CODE LEVEL DEMOGRAPHICS

Modelling data to the postal code level is a challenging task. There are no authoritative data sources at the postal code-level equivalent to the Census. The best and most comprehensive data source at the postal code level is the Canada Post HouseElite file. This file is released monthly and contains the counts of deliverable addresses, along with other information like delivery mode. The deliverable address counts are classified into

four types: house, apartment, farm and business. The deliverable addresses classified as house, apartment and farm are considered residential units. Residential deliverable address counts by type are fundamental inputs used to generate all postal code-level demographics. Additionally, Equifax and an unnamed large, reliable and comprehensive household survey (aggregated to the postal code level) are used to model demographic categories at the postal code geography.

A machine learning technique known as K-Nearest Neighbour (K-NN) is used to generate all initial values for demographic attributes at the postal code level, except for household counts, household income and labour force statistics. The K-NN approach requires that a set of anchor variables exist for "target" and "training" datasets. The target dataset consists of the 758,623 residential postal codes valid for June 2017. The training dataset contains in-house historical data at the postal code level. A set of 20 anchor variables was created for the target and training datasets using the data sources listed above. The anchor variables captured the following dimensions: dwelling structural type, wealth, age, marital status, family structure, ethnicity, geographic context and geographic proximity. It is worth noting that all anchor variables are standardized.

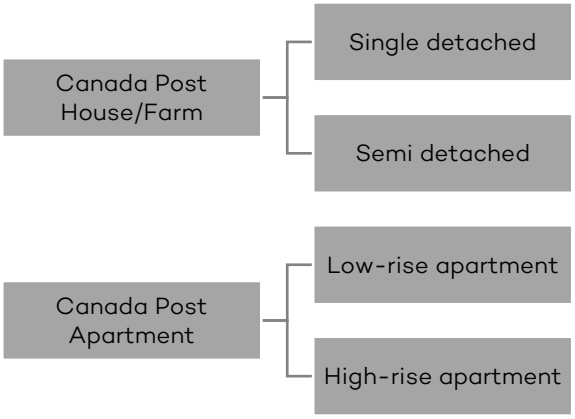
K-NN works by selecting the K most similar observations from the training dataset for each target observation. This is done by using various metrics of similarity based on mathematical distance between observations in multi-dimensional space. K can be set to any value between 1 and the count of training observations. In practice, K most commonly ranges between 5 and 50. We use cross-validation techniques to determine the optimal value for K and weighting schema for anchor variables. From the selected set of K training observations, we create a weighted average ratio value for each variable. The ratio for each variable is appended to the target postal code.

The resulting ratios are multiplied out against their relevant universe based on a strict order of operations. These resulting postal code-level "seed" values are subsequently run through a set of rescaling and optimization procedures. The rescaling procedures adjust the seed values so that they add up to controls derived for census geographies and to the appropriate universe total for each postal code. The controls in this rescaling procedure are created using the process discussed in the section "Demographic Categories."

SPECIAL NOTE ABOUT CANADA POST AND CENSUS STRUCTURAL TYPE DEFINITIONS

It is important to understand the distinction between Canada Post and Census structural types. We use both concepts, and they are critical to producing postal code-level demographics and PRIZM5 assignments. The Census captures eight dwelling structure types: single detached, semi-detached, row house, duplex, low-rise apartment, high-rise apartment, moveable and other dwelling. Canada Post captures three residential structural types: house, farm and apartment. In most circumstance the two classifications of structural type are reasonably consistent (see Figure 4):

Figure 4: Mapping of Canada Post to Census structural types



However, row house, duplex, moveable and other structural types, as defined by the Census, can be classified as either house or apartment depending on different circumstances. These definitions can be particularly confusing in Montreal. Because of the different in dwelling type definitions between Canada Post and Census data, it can be very challenging to determine the correct mix of structural types for a postal code.

To create DemoStats, we made the following adjustments to CensusPlus 2016 in order to create our 2016 base year:

1) DemoStats accounts for the 2.4? percent net-undercoverage rate for the 2016 Census;
2) DemoStats uses the reference date of July 1st, while the Census uses the reference date of May 10th;
3) DemoStats imputes for non-reporting Indian reserves.

REFERENCE DOCUMENTS

DemoStats Release Notes:	www.environicsanalytics.com/en-ca/release-notes
DemoStats Variables List:	www.environicsanalytics.com/en-ca/variables
DemoStats Metadata:	www. environicsanalytics.com/en-ca/metadata