Manifold’s Methodology for Updating Population Estimates and Projections

Zhen Mei, Ph.D. in Mathematics
Manifold Data Mining Inc.

Demographic data are population statistics collected by Statistics Canada via Census every five years. The most recent Census was conducted in May 2011. The upcoming Census will be conducted in May 2016. There is normally one to two years time lag between collecting and publishing Census data. For example, the first batch of 2011 Census, population and dwelling, were released by Statistics Canada in February 2012. The last batch, income and housing, is scheduled for release in August 2013.

Census normally has a 3~5% under-coverage of population as people might not be at home or not submit the census forms on the census day. For details please refer to Statistics Canada at http://www23.statcan.gc.ca:81/imdb/p2SV.pl?Function=getSurvey&lang=en&db=imdb&adm=8&dis=2&SDDS=3601#a3).

Statistics Canada conducts the Reverse Record Check (RRC) after Census to measure census population under-coverage and adjusts population estimates, e.g.,

http://www23.statcan.gc.ca:81/imdb/p2SV.pl?Function=getSurvey&SDDS=3902&lang=en&db=imdb&adm=8&dis=2

For example, Statistics Canada published total population (33,476,688) of Census 2011 on May 16, 2011 at http://www12.statcan.gc.ca/census-recensement/index-eng.cfm?tz=120308

On the other hand, Statistics Canada has also a population estimate (34,368,053) for June 2011 at http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0510005&paSer=&pattern=&stByVal=2&p1=-1&p2=31&tabMode=dataTable&csid=

The difference is 34,368,053-33,476,688 = 891,365, which is about 2.66% of the total population.

Manifold Data Mining Inc. has been providing current year population estimates since 2001. Below is a brief description of our data sources and methodologies for updating population estimates and projections.

Data Source

Statistics Canada
Health Canada
Regional Health Ministries
Citizenship and Immigration Canada
Regional School Boards
Brisc International Inc.
Flyer Distribution Association
Real Estate Boards/Companies
Canadian Bankers Association
Bank of Canada
Canada Post Corporation
Consumer and business directories books
Publication of hospitals, CMHC, BBM and partners
Proprietary survey and research

Longitude Data

We have been mining historical data to identify patterns in population growth and settlement. This includes the historical Census data 1991, 1996, 2001 and 2006, the yearly immigration statistics from year 1981 to 2011, Royal LePage's quarterly Survey of Canadian Housing Price from 1974 to 2011, publications from Canadian Mortgage and Housing Corporation and Canada Post Corporation.

Key Assumption

At Provincial and Census Division levels we have taken consistent assumptions for each component of population growth (birth, death and migration/immigration) with Statistics Canada. At Sub-Census Division level, we determine the assumption by real estate development and directory books as well as historical trends in the Census and immigration statistics.

Fertility

We estimate age-specific fertility rates by cohorts of women in the reproductive age group 15 to 49 and estimate the number of births each year. The data is based on historical birth rate and statistics from national and regional health offices as well as publications of researchers at health networks around major universities, hospitals and Statistics Canada. The trend is that women are having fewer children and are postponing births.

Mortality

We estimate age-specific mortality rates by population cohorts[1]. The data is based on historical mortality rate and statistics from national and regional health offices as well as publications of researchers at health networks around major universities and hospitals. Life expectancy will increase gradually at a slower pace. Over the last decade, average gains in life expectancy have been in the order of 0.12 year per annum for females and 0.25 year for males. The male life expectancy is expected to progress at a faster pace than female life expectancy.

Immigration

Immigration is a key contributor (>=73%) to the population growth. Asian has been the main source of immigrant population in the last two decades. Changes of global economic environment and shift of federal immigration policy will have significant impact on the trend of immigration in the future. At provincial and Census Metropolitan Area (CMA) Levels we use statistics from Citizenship and Immigration Canada, at sub-CMA level we use surveys, community settlement statistics and directory books to estimate the immigration population. Furthermore, immigrant settlement patterns and their longitude shifts are identified from the historical Census and immigration data for projections in future years. We have been studying distribution of the large pool of foreign students and temporary workers since they will be the growing source of immigrant population and will play an important role in the population projections.

Migration

Based on mover's statistics from past census we estimate migration at Census Sub-Division (CSD) level. Thereafter we use postal code development data from Canada Post Corporation, mover's data from data partners, e.g., directory books and real estate boards, and spatial regression models to project migration at sub-CSD level. We also correlate macro-economic activities with migration of labour force across Canada to establish trend of population growth by economic regions.

Occupation & Labour Force & Household Income

Statistics Canada conducts a Labour Force Survey ("LFS") every month. This survey provides estimates of employment and unemployment which are among the most timely and important measures of performance of the Canadian economy. The LFS also provides employment estimates by industry, occupation, public and private sector, hours worked and much more. Statistics Canada published Labour Force Information every month including cross-tables with a variety of demographic characteristics. We use the most recent LFS, the regional unemployment rates based on the Employment Insurance Program, and the Census 2011 as the foundation for our annual updates of labour force activities. Furthermore, we established associations between labour force statistics in Census and business establishments, immigration statistics and settlement patterns to track trends in the labour force and employment development. This enables us to combine the most recent employment statistics, wages, income and inflation data with the

[1] Ronald D. Lee and Lawrence R. Carter, 1992, "Modeling and Forecasting U.S. Mortality," *Journal of the American Statistical Association* 87(419): 659-671.

demographic information and business establishment data by region, industry and occupation, and estimate the current labour force and occupational activities and income levels.

Methodology
As census is conducted every five years and there is a 1-2 years lag in collecting and publishing census data, we estimate demographic data between the census years and project for 1, 5, 10, and 15 years in the future. Our update techniques are based on the following techniques:

- Enhanced cohort survival methods;
- Nearest neighborhood and regression techniques;
- Structural coherence techniques.

Example: Population Forecasting
Population estimation calculates the expected population for the present; population projection calculates the expected population for one or more periods in the future.

The cohort-survival method is the essence of population forecasting:

- Population[t+1] = Population[t] + Natural Increase + Net Migration

This formula states that the population at the next time interval ("t + 1") is equal to the population at the beginning time interval ("t") plus the net natural increase (or decrease) plus the net migration. This is calculated for men and women for each age group.

1. Data source for population at the beginning interval is the Census data from Statistics Canada, e.g. 2011, 2006, 2001, 1996, 1991 census;
2. Data sources for natural increase are Health Canada, Statistics Canada and regional health centers and scientific publications;
3. Data sources for migration are Citizenship and Immigration Canada, Canada Post Corporation, and telephone directories.

Natural increase is the difference between the number of children born and the number of people who die during one time interval. The follow two factors are essential in calculating natural increase:

- Birth Rate[cohort *x*] = Births / Female population at childbearing age;
- Survival Rate[cohort *x*] = 1 - (Deaths[cohort *x*] / Population[cohort *x*]).

Net Migration is the difference between the number of people moving in and the number of people moving out. There are many ways to calculate net migration. Theoretically one can construct complex linear models to predict migration for each cohort. One of the simplest models is based on the assumption that the rate of migration for the next time interval will be the same as the rate of migration for the last time interval for each cohort:

- Migration Rate[t+1] = {(Pop[t] - Pop[t-1])-Natural Increase} / Population[t].

We build models with immigration data from Citizenship and Immigration Canada, new postal information from Canada Post Corporation, labour force survey and macro-economic business activities from Statistics Canada, and directory books.

After population projection we estimate the households and other census data with the following methods:

- Nearest neighborhood techniques;
- Structural coherence techniques.

Income data are projected with current and historical labor force surveys from Statistics Canada. Refinements are performed with the consumer survey data. Labour force data are updated with business establishment data and adjusted with the survey data.

We apply bottom up and top down techniques to population estimates and projections. Information at sub-DA level was used for projections and data at sup-DA level were employed for fine adjustment. Directory books, dwelling structure, real estate development and postal code data are key factors for estimating household counts and migrations. Census 1991, 1996, 2001, 2006 and 2011 were the base and trend for population projection. In the following we summarize the key techniques in creating and updating SuperDemographics.

a) Nearest neighborhood and regression techniques

To estimate population in a new residential area, we use nearest neighbors and regression techniques, looking for most similar records in the historical database and in the neighbourhoods in terms of construction type, year, number of dwelling, phone lines, … and assigning an initial value to the new area. We improved the basic nearest neighbor techniques with a multi-level similarity measure and an adaptive voting procedure from the K-nearest neighbours for assigning prediction to the new record. The confidence of the improved K-nearest neighbours technique are measured as follows.

- The distance to the nearest neighbor provides a level of confidence in accuracy.
- The degree of homogeneity among the prediction within the K-nearest neighbors is an indicator of confidence in coherence.

b) Structural coherence techniques

Multi-colinearity is common in large databases. We use structural coherence to measure robustness of the databases. In the modeling process, we explore structure in data and variables structure and preserve structural coherence of the database.

To preserve the coherence structure of the census data, we have applied the theory of nonlinear dynamic systems developed by Manifold's principal to the spatial and demographic dynamics[2].

c) Transferring data from DA (Dissemination Area) to postal code level via numeric methods

Data at different geographic levels are linked by a large system of linear equations. For example, a 6-digit postal code can run across several dissemination areas. Population within the postal code will be split into different portions corresponding to the dissemination areas. Correspondingly, a dissemination area may cover multiple postal codes. The total population of the dissemination area is equal to the sum of proportional populations of the linked postal codes. Setting up such a linear equation for every dissemination area and postal code in Canada generates to a large system of linear systems for population weight of all postal codes. This system is over-determined and has more than 750,000 unknowns. By solving such a system for anchor demographic variables, e.g., population, dwelling, income, … we obtain the core census data at the 6-digit postal code level.

d) Predictive models for postal code level data

Based on the anchor variables at the 6-digit postal code level, we used spatial linear and nonlinear regression techniques to derive all other demographic variables. Particularly we considered the variation of population density and dwelling values among different postal codes within same dissemination area. Thousands of models were built to predict census attributes to all residential 6-digit postal codes.

e) Validation and refinement via independent data sources

Our databases have been verified with most recent data from Statistics Canada and survey data from our partners, postal information from Canada Postal Corporation, real estate boards, data vendors and online maps.

f) Errors

All regression results were derived within 5% error bounds with 95% confidence level.

Contact

Dr. Zhen Mei or Thomas Ding
Manifold Data Mining Inc.
220 Duncan Mill Road, Suite 519
Toronto, ON M3B 3J5
Canada
T: 416-760-8828
F: 416-760-8826
E: zhen@manifolddatamining.com

[2] Zhen Mei: *Numerical Bifurcation Analysis for Reaction-Diffusion Equations.* Springer Series in Computational Mathematics, Vol. 28, Springer-Verlag, Heidelberg, Berlin, New York 2000.