Manifold Methodologies

Zhen Mei, Ph.D. in Mathematics

July 2019

220 Duncan Mill Road, Suite 519,
Toronto, ON M3B 3J5
CANADA
Tel: 416-760-8828
Fax: 416-760-8826
www.manifolddatamining.com

1) Manifold's Methodology for Updating Population Estimates and Projections

Census is the official count of population and collection of detail demographic information of individuals. It is conducted by Statistics Canada every five years. The most recent Census was 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 2016 Census, population and dwelling was released by Statistics Canada on 8th February 2017. Additional data was released stepwise in the following schedule:

- May 3, 2017: age and sex, type of dwelling;
- May 10, 2017: Census of Agriculture;
- August 2, 2017: language and families, households and marital status;
- September 13, 2017: income;
- October 25, 2017: immigration and ethnocultural diversity, housing and Aboriginal peoples;
- November 29, 2017: education, labour, journey to work, language of work and mobility and migration.

Statistics Canada publishes data Census Sub-Division (CSD) Level on the release date. Data at the Dissemination Area level was published normally a few months later.

Census population can be often mistaken for the real or actual population. In fact, census population is simply a count of the total number of people enumerated by the census. It is not adjusted for students away from home, people out of the country at the time, people who do not fill in the survey for any reason, including lack of capacity or lack of official language skills, sick or infirm, part of a group that traditionally does not have a high fill rate (First Nations, temporary workers), etc. There is an average undercount of about 3% of the population, but this percentage is highly variable and typically higher in rural areas, First Nations, student residential areas and certain ethnic communities. Some very small communities report a very high percentage of undercount. A few first nation reserves did not participate Census 2016, for example, Six Nations, Chippewas and Oneida. The Census 2016 data for municipalities associated with these first nation reserves undercount the population.

"Population estimates" are different from the "Census population". In "population estimates" we adjust the undercount of Census and estimate the actual population count. The foundation of our estimates is the current and historical Census in the five year interval, plus a number of additional sources, like Canada Revenue Agency tax filings data, and Canada Post data, housing starting statistics from Canadian Mortgage and Housing Corporation, immigration statistics, movers data, birth and mortality rates.

Statistics Canada provides a partial explanation of undercount: https://www12.statcan.gc.ca/census-recensement/2011/ref/estima-eng.cfm**.** Furthermore, in https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501 Statistics Canada shows that undercounts are adjusted for population estimates at Provincial and Canada

levels. On this same page (add data if needed), note the 2nd quarter 2016 population estimate (36,109,487) for Canada, which is 957,759 and more people than the released population count (35,151,728) in the 2016 Census taken that quarter. This is a difference of 2.7%.

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/imdb/p2SV.pl?Function=getSurvey&SDDS=3902&Item_Id=47932

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.

Sample Data Sources:

Statistics Canada
Health Canada
Regional Health Ministries
Citizenship and Immigration Canada
Regional School Boards
BRISC International Inc.
Flyer Distribution Association
Real Estate Boards/Companies
Canada Mortgage and Housing Corporation (CMHC)
Canadian Bankers Association
Building permit statistics from Municipalities
Industry Canada
Numeris Canada
ViviData Canada
Publication of hospitals, government agencies and partners
Open Data Canada, Provinces and Municipalities
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, 2006, 2011 and 2016, the yearly immigration statistics from year 1981 to 2019, Royal LePage's quarterly Survey of Canadian Housing Price from 1974 to 2019, publications from Canadian Mortgage and Housing Corporation and Canada Post Corporation and hundreds of open data sources from municipalities across Canada. Particularly in each Census there are gaps caused by nonparticipation of a few first nation reserves which may vary in the Census cycle. We examine the historical trends of first nation population and utilize research reports from Statistics Canada to fill the gaps.

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, mail and flyer distribution networks 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, although bounce back of birth rate has been observed in some regions. Certain ethnic groups have high birth rates, for example, the fertility rate of First Nations is four times higher than the Canadian average.

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. For example, http://www.statcan.gc.ca/pub/82-625-x/2017001/article/14775-eng.htm. 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.

Interprovincial 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., flyer distribution networks, directory books and real estate boards, and spatial regression models to project migration at sub-

[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.

CSD level. We also correlate macro-economic activities with migration of labour force across Canada to establish trend of population growth by economic regions.

Ethnicity and International Migration
Our projection of the Chinese, South Asian, East European, Filipino, and Caribbean communities is based on historical Census from 1991 to 2016, immigration statistics 1991 to 2019 from Citizenship and Immigration Canada as well as birth and mortality rates of these communities in Canada.

Immigration is a key contributor (>=75%) to the population growth. Asian has been the main source of immigrant population in the last two decades. 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 the large pool of refugees, foreign students and temporary workers as the increasing source of immigrant population and factoring them into the population projections.

The household spending patterns of these communities are derived from the annual Survey of Household Spending from Statistics Canada. We used predictive models to link the spending data with Census data and extrapolate them to the 6-digit postal code covering whole Canada. We used spending patterns in areas with high concentration the cultural communities to represent cultural spending patterns. Coupled with their settlement patterns and socio-economic data at the 6-digit postal code level we extrapolate their spending patterns across Canada.

Labour Force, Occupation and Income
Household income is closely correlated with labour force and occupation. Statistics Canada conducts monthly a Labour Force Survey ("LFS"). This survey provides statistics of labour force and employment by industry and region, for examples, participation rate, unemployment rate, occupation, public and private sector, hours worked, wages and salaries and so on. Every month Statistics Canada publishes the Labour Force Information along with a variety of demographic characteristics and Consumer Price Index (CPI).

The most recent Census is the foundation for our updates of labour force activities, household and personal income. We have established correlations between labour force statistics in Census and Labour Force Survey, and business establishments, immigration statistics and settlement patterns. We have also an ongoing tracking process of the trends of the labour force and employment development. Using the most recent LFS and CPI statistics, and the regional unemployment rates from the Employment Insurance Program with wages, salaries and inflation data, Census and business establishment as input variables we build predictive models to estimate the current labour force and occupational activities, and income levels. We also incorporate the taxfiler data (employment and investment income) and the macro economic data, e.g., monthly CPI

data and publications from Bank of Canada into our forecast models for refining the estimates of the personal and household income.

Dwelling Value
We start with the self-reported dwelling value in the current Census and adjust it to the market value with real estate statistics and surveys from Canadian and regional Real Estate Associations and companies, for example, home listing prices, House Price Survey from Royal Lepage, Market Watch from Toronto Real Estate Board. We incorporate also estimates and trend analysis from Canadian Housing and Mortgage Corporation (CMHC) to reflect the geographic patterns and historical development of the real estate market. In addition, we consider influence of immigrant and aging population, and macroeconomic changes in our forecast models and validate our models with the sales reports from various real estate boards. However, in certain markets like Vancouver, Toronto and recently Montreal, the housing price can vary irrationally and may not be captured completely in our models.

Methodology
As census is conducted every five years and there is a 1-2 years lag in processing 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, collaborative filtering 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. 2016, 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, Real Estate Boards/Companies 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 assumes 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, statistics from various Real Estate Boards/Companies and directory books.

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

- Collaborative filtering 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 adjustments. Directory books, dwelling structure, real estate development and postal code data are key factors for estimating household counts and migrations. Historical Census from1991 to 2016 were the base and trend for population projection. The new release of Census 2016 is being incorporated into our year 2019 population estimates and projections. In the following we summarize the key techniques in creating and updating our population estimates and projections.

a) Nearest neighborhood and regression techniques
To estimate population in a new residential area, we use nearest neighbors and spatial 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 790,000 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, which incorporates population density, dwelling types, real estate development and building permits where available, patterns and trends in population settlements, business establishments and economic developments

We create the system of equations linking population from postal code level to DA level by geo-coding the 6-digit postal codes with high precision[3]. We validate the geo-coding precision with Google map, Bing and open street maps.

We use dwelling type and home listing prices at the 6-digit postal code level in the last 8 seven years to refine the estimates derived from Census 2016.

d) Predictive models for postal code level data

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

[3]D. Li, S. Wang, W. Cai, and Z. Mei: *A Navigation Assistance Agent: Mobile LBS Web Service*, SAE Technical Paper 2007-01-1107, 2007

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) Consumer product usage, purchase behaviour, shopping pattern, media usage, financial and psychographic data products

Since 2007 we have been providing the Canadian marketplace with a dozen of data products on consumer product usages, purchase behaviours, shopping patterns, media usages, financial and psychographic patterns. We developed these data products based on the Return-To-Sample Survey from the NUMERIS (formerly Bureau of Broadcast and Measurement BBM RTS). This survey is conducted twice a year and the sample size was over 63,000 till 2013 and around 42,000 afterward in each wave till 2019. We have licensed over 11 years of the survey data. Totally we have over 1.4 million responders in our database and they are stratified properly by geography and demographics. They represent Canadian consumers across the country. For over 15 years NUMERIS RTS survey data has been widely by Canadian media operators, agencies and advertisers.

Applying data fusion and deep learning techniques to the NUMERIS RTS responders' level data and our 6-digit postal code level demographic and spending data we developed thousands of predictive models to estimate propensity of consumer purchase behaviour, consumption and psychographics for all 6-digit residential postal codes across Canada. The propensity score measures how likely consumers in a 6-digit postal code purchase and use certain products and services, how often they may shop at certain stores and what do they think about certain things. We optimize the propensity score with random forest techniques.

For new postal codes which have no or very limited data available, we use the nearest neighborhood techniques or the collaborative filtering techniques to impute the data from the nearest postal codes. Alternatively, average value at a high geographic level can be used as the initial estimate which will be refined at a later stage when more data becomes available. For example, when residents move into a new condominium building, few data is available except geo-code, home listing price, number of units. We'll use data of nearby and relatively new condominiums to estimate the data for this new condominium. Refinement will be done when more information is available, latest till next Census.

f) Validation and refinement via independent data sources

Our data products 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.

We check validity of the data by examining whether there are missing values or data points outside of an allowed range of values. For example, household size should be lies between 0 and 15, unless in a band housing arrangement.

We verify data consistency by checking relationships between variables. Consistency can be based on logical, legal, accounting or structural relationships between the variables. The

relationship between population age less than 15 year old and marital status is one example where the consistency is reflected through: “Total population less than 15 years of age should not be greater than the total population of never married.”

We perform distribution analysis to identify records that are outliers with respect to the distribution of the data. For example, Gini coefficient and income inequality, dwelling value and real estate market in major cities.

In-phase and out-phase validation is also used to verify accuracy of our estimates and predictions.

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

2) Manifold Methodologies for Data Mining

At Manifold we develop and apply innovative and efficient data mining techniques to help clients achieve their marketing objectives. We employ both the well-established statistical methods and the newest data-driven technologies to custom solutions for our clients. We have active joint research projects with university professors (Sherbrook, York) on innovative data mining algorithms and big data analytics. These initiatives supported and endorsed by Natural Science and Engineering Research Council of Canada (NSERC). Here are a few examples:

a) Dimension reduction techniques

Dimension reduction is a process to condense information in big and potentially sparse database into low dimensional manifolds with the following features:

- They share all essential attributes with the original database;
- They are suitable for efficient campaign management, analytics and data mining, as well as Ad Hoc query and reporting.

We used the following proven methods and proprietary technologies:

- Correlation analysis
- Variable clustering
- Principal component analysis
- Factor analysis
- Discriminate analysis
- Regression analysis
- Feature selection with clustering techniques[4,5]
- Projective adaptive resonance theory
- Bayesian neural network[6]
- Collaborative filtering and machine learning[7].

Many machine learning techniques are results of our joint research projects with university researchers and proprietary development. Our research projects have been endorsed by The Natural Sciences and Engineering Research Council of Canada (NSERC).

[4] H. Sun, S. Wang, and Z. Mei: A fuzzy clustering based algorithm for feature selection. Machine Learning and Cybernetics, 2002. Pages: 1993 - 1998 Vol. 4-5 Nov. 2002

[5] P. Lasek and Z. Mei: Clustering and visualization of a high-dimensional diabetes dataset. Procedia Computer Science, Vol. 159, Pages: 2179-2188. 2019

[6] J. Xue, Y.Liu, X. Zeng, W. Zhang and Z. Mei: A Bayesian network model for predicting type 2 diabetes risk based on electronic health records. Modern Physics Letters B. Vol 31, Issue 19-21, World Scientific Publishing Company, 2017

[7] Y. Chen, M. Yann, H. Davoudi, J. Choi, A. An and Z. Mei: Contrast pattern based collaborative behavior recommendation system for life improvement. Advances in Knowledge Discovery and Data Mining, Pages: 106-118, Springer Verlag, 2017

b) Resample techniques

Survey data are mostly collected at the household level. These data may describe accurately certain aspects of consumption behavior of the responders. However, the sample size is often too small and the sample is biased because responders may not represent the total population properly. We developed stratified sample techniques to improve the efficiency of survey data.

c) Cluster analysis

A process clustering objects in a database into different groups so that:

- Objects in the same group are as similar as possible (Homogeneity);
- Objects in different groups are as different as possible (Heterogeneity).

Here the measure for similarity is crucial. Particularly for categorical variables, there are many ways to define a similarity matrix. For the interval scale variable, we use Euclid or Mahalanobis distance.

We have enhanced the K-means clustering techniques with the identification of a local optimal number of cluster and optimization of seeds selections. Our algorithms have been published in scientific conference proceedings and journals[8] Using Projective Adaptive Resonance Theory (PART) we developed neural networks to perform clustering on ordinal and categorical datasets of consumer surveys. To cluster text data, we apply Latent Dirichlet Allocation (LDA) model to discover automatically topics in the consumer survey and convert the text into vectors which can be clustered directly by the PART algorithms. This enables us to accurately estimate the dwelling values and to validate the CanaCode lifestyle clusters with Numeris RTS data.

d) Data fusion

Data fusion with stratified sampling techniques. Stratum is the key to link survey data at the household level with census data at the level of dissemination areas. We used a multi-staged and adaptive nonlinear method to reduce the dimension of the database. We defined effective statistical distance functions and measured structural coherence in selecting the geographic level and integration of demographic, expenditure and behavior databases.

e) Product-driven data mining

The behaviour of consumers is influenced by many factors. The consumer's needs and

[8] Sun H., S. Wang, and Q. Jiang: A New Validation Index for Determining the Number of Clusters in a Data Set. Proceeding of INNS-IEEE Conference on Neural Networks'01 (Washington DC), pp.1852-1857, 2001.
Sun H., S. Wang, and Q. Jiang: FCM-based Model Selection Algorithms for Determining the Number of Clusters. Pattern Recognition, 2003.

desires are described by factors like the individual's demographics, spending patterns, hobbies and activities, culture, social status, lifestyle and attitudes, beliefs and motivations, trust and relationship with others. Manifold has been cooperating with university researchers on understanding how these complicated and interrelated factors drive consumer purchase behaviors.

Based on Bayesian networks, we developed machine learning techniques to predict and simulate nonlinear consumer behaviours. Our results are published in:

R. Aggarwala, C.S. Bohun, R. Kuske, G. Labute, W. Lu, N. Nigam and F.M. Youbissi: Product-Driven Data Mining. Proceeding of the Seventh PIMS-IMA Industrial Problem Solving Workshop, 2003 and CANADIAN APPLIED MATHEMATICS QUARTERLY Volume 12, Number 1, Spring 2004
http://www.math.ualberta.ca/ami/CAMQ/pdf_files/vol_12/12_1/CAMQinfo.pdf

Yan Chen, Margot Lisa-Jing Yann, Heidar Davoudi, Joy Choi, Aijun An and Zhen Mei: Contrast Pattern based Collaborative Behavior Recommendation System for Life Improvement. In Advances in Knowledge Discovery and Data Mining, pp 106-118, Springer Verlag, 2017

Jiang Xie, Yan Liu, Xu Zeng, Zhen Mei: A Bayesian network model for predicting type 2 diabetes risk based on electronic health records. International Journal of Modern Physics B, World Scientific Publishing Company, Vol 31, Issue 19-21 2017

f) Validation and refinement via independent data sources

We validate the selected and developed techniques with most recent data from Statistics Canada and survey data and publications from A.C. Nielsen, Ipsos Reid, Adhome, NUMERIS and other data vendors.

We work with our clients and market research partners to validate theory and algorithm with their valuable business experience and best practice. We have been improving iteratively our techniques. We welcome your comments and suggestions.

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