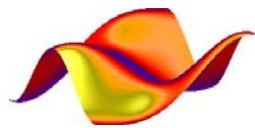


# Manifold Methodologies



*Manifold* Data Mining Inc.

501 Alliance Ave., Suite 205,  
Toronto, ON M6N 2J1  
CANADA  
Tel: 416-760-8828  
Fax: 416-760-8826  
[www.manifolddatamining.com](http://www.manifolddatamining.com)

---

## Micro-Marketing Database at the 6-Digit Postal Code Level

### *Census Data*

Census demographic data are the most complete and comprehensive population statistics from Governments. Statistics Canada publishes census data every five-years at the geographic level of Dissemination Areas (DA) and up. On average a DA contains 250 households that can be quite heterogeneous in demographics. For example, different ethnic groups, apartment dwellers and single detached house owners may reside in the same dissemination area, and they have different consumer behavior and shopping habits. Demographic data only partially explains consumer behavior.

### *Consumer Survey Data*

Survey data, on the other hand, are mostly consumer action and behavior data. These data are normally collected at the household level. A survey may describe certain aspects of consumer behavior, e.g. hobbies, shopping habits and credit card usage. Consumer behavior data are stronger predictors in Customer Relationship Management (CRM) than demographic data. The latter are better in estimating customer's needs. But the survey data are often biased because responders may not be properly representative of the total population.

### *The Right Geography*

DA level data do not have sufficient predictive power for many marketing applications. Central limit theorem tells that the larger the sample size, the more alike the average of samples. Leveraging the high predictive power of household survey data is essential for understanding consumers' spending pattern and purchase decision process, and for the success of marketing. The 6-digit postal codes, with 10-15 households as one unit, provide marketers a pinpointed opportunity in target marketing and customer relationship management. Six-digit postal codes have unique features:

- There is sufficient and sustainable information available;
- More homogeneous than dissemination areas;
- Easy attachment to customer database via addresses;
- No conflict with government privacy rules.

Using advanced data mining and data fusion techniques, we integrated demographic data and consumer survey data at the 6-digit postal code level, and transformed the raw data into powerful predictors for marketers and data modelers.

---

*Manifold Data Sources*

- Statistics Canada
- Canada Post Corporation
- Health Canada
- Industry Canada
- Citizenship and Immigration Canada
- Adhome/Valassis Survey
- BBM Survey
- Directory Books
- Manifold Data Mining Inc.

---

## Manifold Methodology for Creating SuperDemographics

The census is conducted every five years. There is always 2-3 years time lag in collecting and publishing census data. We estimate demographic data between the census years and project for 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:

- $\text{Population}[t+1] = \text{Population}[t] + \text{Natural Increase} + \text{Net Migration}$

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. 2001, 1996, 1991 census;
2. Data sources for natural increase are Health Canada, Statistics Canada and regional health centers;
3. Data sources for migration are Citizenship and Immigration Canada, Canada Post Corporation, and directory books.

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:

- $\text{Birth Rate}[\text{cohort } x] = \text{Births} / \text{Female population at childbearing age};$
- $\text{Survival Rate}[\text{cohort } x] = 1 - (\text{Deaths}[\text{cohort } x] / \text{Population}[\text{cohort } x]).$

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 quite 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 built models with data from Citizenship and Immigration Canada, Canada Post Corporation 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.

We have taken bottom up and top down techniques. Information at sub-DA level was used for predictions and data at sup-DA level were employed for fine adjustment. Directory books, dwelling structure, real estate development and postal code data were indicators for estimating household counts and migrations. Census 1991, 1996 and 2001 were the base and trend for population projection.

a) *Nearest neighborhood and regression techniques*

---

To predict a missing value or an expected event for a new record in a database, e.g., compressed data or new residential areas, we used nearest neighbor and regression techniques, looking for most similar records in the historical database and assigning their value to the new record. We improved the basic nearest neighbor techniques with a multi-level similar measure and an adaptive voting procedure from the K-nearest neighbors for assigning prediction to the new record. The confidence of the improved K-nearest neighbors technique are measured with the following methods.

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

b) *Structural coherence techniques*

---

Multi-collinearity 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<sup>1</sup>.

*c) Transferring data from DA 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 be linked to several dissemination areas. Population within the postal code is split into corresponding portions. Correspondingly, a dissemination area covers many postal codes. 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 leads to a large system of linear systems for proportional populations of all postal codes. This system is over-determined and has more than 700,000 unknowns. Solving such a system for anchor demographic variables and building thousands of predictive models we created 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 all census variables.

*e) Validation and refinement via independent data sources*

---

Our databases have been verified with most recent data from Statistics Canada and survey data from Adhome, BBM and other data vendors.

We update the data on an ongoing basis with most recent data published by governments, private data vendors and in survey results.

---

<sup>1</sup>Zhen Mei: *Numerical Bifurcation Analysis for Reaction-Diffusion Equations*. Springer Series in Computational Mathematics, Vol. 28, Springer-Verlag, Heidelberg, Berlin, New York 2000.

---

## 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. Here are a few examples:

### 1) *Dimension reduction techniques*

---

Dimension reduction is a dynamic process to condense information in large 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
- Center manifold theory<sup>2</sup>.
- Feature selection with clustering techniques<sup>3</sup>.

### 2) *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.

### 3) *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

---

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

<sup>3</sup> Sun H., S. Wang, and Z. Mei: A fuzzy clustering based algorithm for feature selection. *Machine Learning and Cybernetics*, 2002. Page(s): 1993 - 1998 vol.4 4-5 Nov. 2002

---

(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<sup>4</sup> of cluster and optimization of seeds selections.

#### 4) *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.

#### 5) *Product-driven data mining*

---

The purchase behaviour of consumers is influenced by many factors. The consumer's needs and desires are described by factors like the individual's demographics, spending patterns, hobbies and activities, culture, social status, lifestyle and attitudes. Manifold has been cooperating with university researchers on understanding how these complicated and interrelated factors drive consumer purchase behaviors. 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.*

#### 6) *Validation and refinement via independent data sources*

---

We validate the selected and developed techniques with most recent data from Statistics Canada and survey data from A.C. Nielsen, Adhome, BBM and other data vendors. We work closely with our clients to validate theory with their valuable business experience and iteratively improve our techniques.

---

<sup>4</sup> 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.