

# APPLICATION OF CLUSTER ANALYSIS TOWARDS THE DEVELOPMENT OF HEALTH REGION PEER GROUPS

Larry MacNabb<sup>1</sup>

## ABSTRACT

The inception of the Canadian Community Health Survey in conjunction with the expansion of existing data products for the provision of health region level information has necessitated the need to develop a method of comparing regions with similar socio-economic determinants of health. After the effects of the various social and economic characteristics known to influence health status have been removed it then becomes possible to compare the health status of regions in the same “peer group” and measure the relative effectiveness of the many health promotion and prevention strategies employed across regions. Cluster analysis was used to place 139 health regions into 10 distinct groupings or “peer groups” possessing similar socio-economic characteristics. Health regions were defined using 24 variables chosen to cover as many of the known social and economic determinants of health as possible. This paper will outline the data sources used to define the peer groups as well as the practical constraints and criteria placed on the analysis. Methods used will be outlined and results will be presented along with the obstacles encountered in the actual application of the methodology.

KEY WORDS: Cluster analysis, Health Regions, Peer Groups.

## RÉSUMÉ

Le commencement de l'enquête sur la santé dans les collectivités canadiennes en conjonction avec l'expansion des sources existantes de données pour la santé au niveau régional a rendu nécessaire le développement d'une méthode pour comparer les régions avec des déterminants socio-économiques de la santé. Après que les effets des diverses caractéristiques sociales et économiques connues pour influencer l'état de santé ont été enlevés, il devient alors possible de comparer l'état de santé des régions dans le même groupe de pairs (peer group) et de mesurer l'efficacité relative des nombreuses stratégies de promotion et de prévention de la santé utilisées à travers les régions. L'analyse de regroupement est utilisée pour regrouper 139 régions de santé dans 10 groupes de pairs qui possède des caractéristiques socio-économiques semblables. Les régions de santé sont définies en utilisant 24 variables choisies pour couvrir le plus de causes sociales et économiques connues déterminantes pour la santé. Cette présentation décrit les sources de données utilisées pour définir les groupes de pairs ainsi que les contraintes pratiques et les critères faits sur l'analyse. Les méthodes utilisées sont décrites et les résultats sont présentés avec les obstacles encourus lors de l'application réelle de la méthode.

MOTS CLÉS : L'analyse de regroupement; groupe de pairs; regions;

## 1. INTRODUCTION

In recent years expansion of existing health data products and the inception of the Canadian Community Health Survey (CCHS) have given researchers and planners a wealth of health information at sub-provincial levels of geography. This has led to the subsequent ranking of health regions based on health indicator performance. Health regions are defined by provincial governments as the areas of responsibility for regional health boards (i.e. legislated) or as regions of interest to health care authorities. For the purpose of completeness each northern territory also represents a health region in this analysis. Due to the non-homogeneous nature of health region boundaries it is often not appropriate to paint all regions with one brush in any given analysis. One example of such an inappropriate comparison would be the ranking of the City of Toronto Health Unit with a population of 2.5 million to the Grenfell Regional Health Services Board in the province of Newfoundland and Labrador with a population of only 17,000.

To deal with this issue an analysis was conducted with the main goal of grouping 139 health regions into “peer groups” based on their social and economic characteristics. Development of the final methodology was guided by the “Peer Group

---

<sup>1</sup> Larry MacNabb (larry.macnabb@statcan.ca), Statistics Canada, Health Statistics Division, Main Building, Room 2600, Ottawa, Ontario, K1A 0T6.

*Project Working Group*". This group included representatives from Statistics Canada, provincial ministries of health, regional health authorities, universities and the Canadian Institute for Health Information. At the onset of this initiative an initial set of requirements were defined to direct the analysis. Due to the nature of their intended use the analysis focused solely on variables associated with health outcomes. A further requirement was that the data used had to be available for all Health Regions and final "peer groups" had to be derived using empirical techniques. This all led to the main objective of placing 139 health regions into groups containing 5 to 10 of their closest statistical peers.

## 2. DATA

In order to meet the requirement of using information available for all health regions data were used from a number of sources. These sources included vital statistics, the census and demographic data. The analysis focused on information collected during the 1996 calendar year and included 24 social and economic variables covering most of the known determinants of health.

The variables used covered all of the major correlates such as population change, demographic structure, social status, economic status, aboriginal status, housing, urban/rural, income inequality and labour market. A complete list of variables and their definitions can be found in MacNabb 2002.

## 3. METHODOLOGY

### 3.1 Clustering Methods

Cluster analysis attempts to organize variables or observations into distinct groups based on a statistical measure of their distance from each other or a distinct point in p-dimensional space, where p represents the number of variables used to describe each group of interest. In this particular instance health region is the unit of study and p represents the 24 social and economic variables used to describe a health region. Two methods of cluster analysis were considered for this analysis: hierarchical and non-hierarchical.

Hierarchical methods take the approach of splitting N observations into a series of m clusters, where m can range from 1 to N. This corresponds to the extreme case of having all observations grouped into one cluster or conversely having each observation in a cluster by itself. The strength of this technique is that it provides the possibility of increasing or decreasing the number of clusters depending on the required level of aggregation. However, one of the weaknesses of this method is its inability to adapt to relationships uncovered at later stages of analysis (Andberg 1973).

Non-hierarchical methods split observations into a pre-defined number of groups using a specified optimization criterion. This approach was best suited to meeting the main objective of grouping 139 health regions into clusters composed of 5 to 10 health regions. This method performs the task of splitting N observations into a predefined set of k clusters, where each cluster k has mean  $\bar{x}_k$  and varying size  $n_k$ . The optimization method used in this analysis and in many of the most commonly used clustering algorithms minimized the within cluster sum-squared error denoted by:

$$E = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad (1)$$

(Everitt, 1993).

### 3.2 K-Means Clustering

The k-means algorithm found in the SAS procedure FASTCLUS was used to perform the actual cluster analysis. The basic steps of the algorithm are detailed below. For a more thorough description of the method and its variants see Andberg 1973.

1. Select  $k$  observations as cluster seeds.
2. Assign all observations to their nearest cluster seed.
3. Replace cluster seeds with the mean of all observations assigned to each cluster.
4. Repeat 1 to 3 until the change in cluster means becomes or approaches 0.
5. Form final clusters by assigning each observation to its nearest cluster using the final cluster means derived from 4.

### 3.3 Number of Clusters

Selecting an appropriate number of clusters as a starting point for the analysis is one of the major challenges associated with non-hierarchical clustering algorithms. Several test statistics have been suggested (Everitt, 1993) to perform this task. From a practical perspective this decision is generally left to the analyst. The original requirements identified a range of 5 to 10 health regions in each cluster. A maximum of 20 clusters was chosen as a starting point for the algorithm which would result in an average cluster membership of 7 health regions.

## 4. RESULTS

### 4.1 Initial Analysis

During the first stage of the analysis the algorithm was instructed to group 139 health regions into 20 clusters. This resulted in 6 regions being placed into a cluster by themselves. This was problematic in that it indicated that there was not enough variability in the data to support 20 clusters as outlined in the original study objectives. The results of this analysis are presented in Figure 4.1 which plots cluster membership against the first 2 principal components of the 24 variables used in the analysis.

The first 2 principal components accounted for roughly 53% of the total variance in the data and served as a useful tool for simplifying the assessment of cluster groupings. From Figure 4.1 it can be clearly observed that the 6 regions residing in their own cluster (identified by a large circle) are located predominantly on the fringes of the large mass of health regions in the middle of the plot.

### 4.2 Secondary Analysis

In order to deal with the situation of having regions in their own cluster the means from clusters having more than one member as identified in section 4.1 were used as starting seeds for a new cluster analysis resulting in a decrease from 20 to 14 clusters overall. To further reduce the impact of outlying health regions plots of the distance between cluster centers and cluster radius were examined and a maximum cluster radius of 5 was imposed upon the algorithm.

In total this resulted in all regions but 3 being placed into a cluster with more than one member. The remaining 3 regions did not meet the strict cluster radius criteria entered into the algorithm.

### 4.3 Final Results

The 3 regions falling outside the maximum cluster radius were added to their nearest cluster. In order to meet the practical objective of having a minimum of 5 comparative health regions in each peer group, clusters with fewer than 5 members were combined with their closest neighbours. The results are presented in figure 4.2. The end result is a total of 10 clusters or “peer groups” with membership ranging from 5 to 34 health regions. Each cluster is composed of health regions from different provincial jurisdictions.

Figure 4.1 – Initial cluster results versus first 2 principal components.

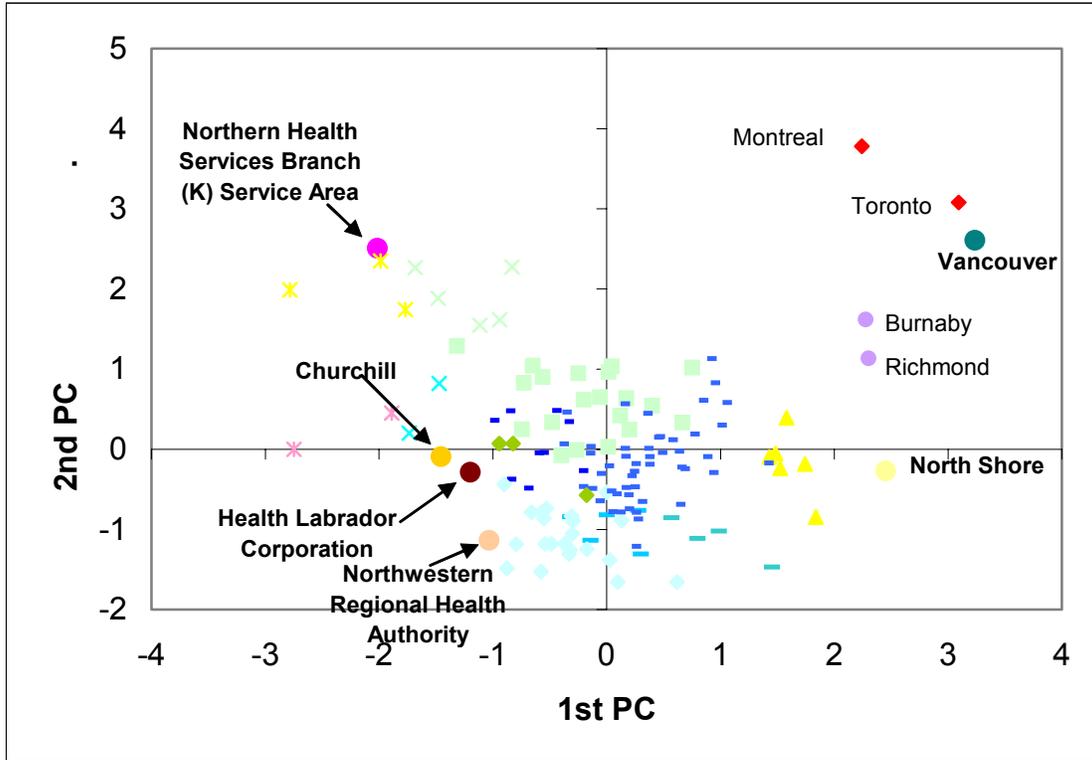
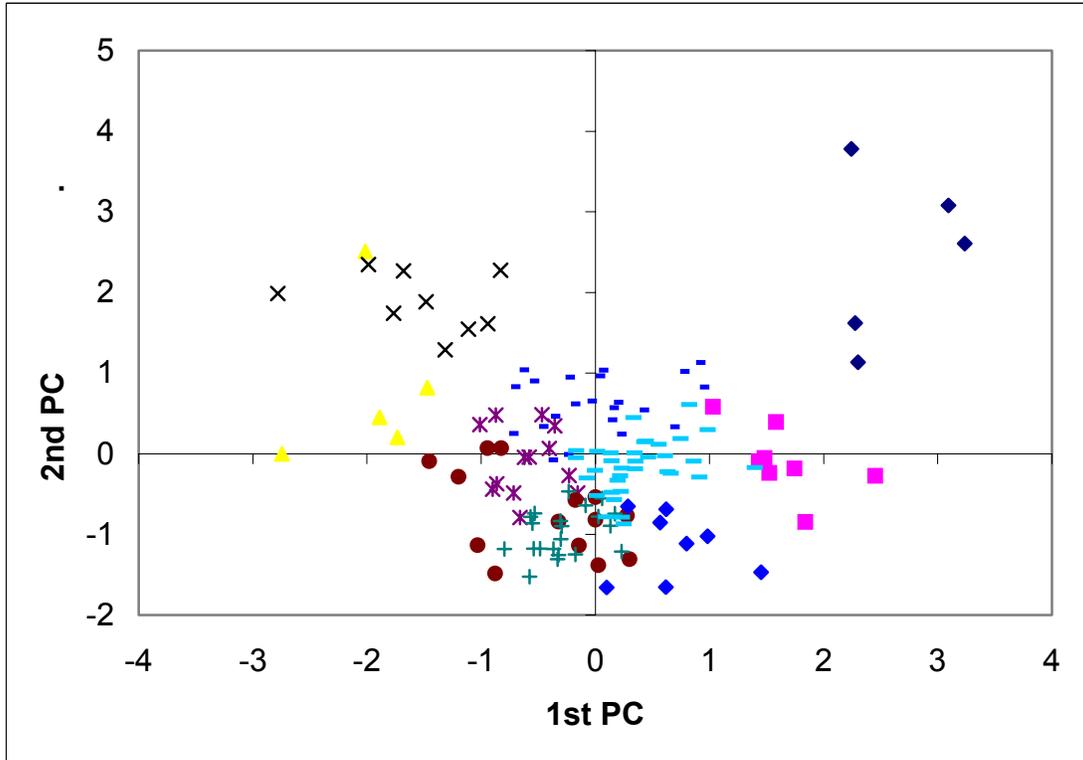


Figure 4.2 – Final cluster groups versus first 2 principal components.



## 5. DISCUSSION

### 5.1 Strongest Predictors

In order to identify which variables were driving the final groupings a stepwise discriminant analysis was performed on the final cluster assignments against the full set of 24 variables used in the analysis. Partial R-SQ statistics were set at 0.15 for both the entry and removal of variables from the analysis. Any variable which had an R-SQ of 0.5 or higher when regressed against a variable included in the model was removed. A summary of the key variables and their relevant Partial R-SQ statistics is presented in Table 5.1.

The percentage of Aboriginals and Visible Minorities in each health region played a strong role in defining the final peer groupings. Unemployment Rate, 1996 Population and proportion of population over the age of 65 also played a key role in the groupings. Finally income inequality and migration mobility also helped to further refine the clusters.

**Table 5.1 – Stepwise discriminant analysis of final health region groupings on all 24 variables.**

Step	Variable	Partial R-SQ
1	Aboriginal Percentage	0.9075
2	Percent Visible Minority	0.8952
3	Unemployment Rate	0.8127
4	1996 Population	0.7369
5	Proportion of Population Age 65+	0.6093
6	Income Inequality	0.5121
7	Migration Mobility	0.2803

### 5.2 Limitations

One of the major limitations of non-hierarchical cluster analysis is that it is an averaging technique. As such many of the health regions which lie on the boundaries of a specific peer group could in actuality be paired with regions in a neighbouring cluster. In total 85% of all health regions were grouped into a cluster with at least 2 of their 5 closest peers indicating that overall the methodology performed well and is meeting the original study objectives.

From a practical perspective it was revealed that many health regions are not that different. Evidence from this can be seen in the fact that 3 of the 10 peer groups had membership of more than 20 health regions. Splitting the data further would amount to an exercise in splitting hairs.

Finally health region geography is defined by provincial governments resulting in non-standard boundary definitions limiting the effectiveness of the analysis.

### 5.3 Advantages

Effectively the analysis removed the impact of socio-economic influences in the comparison of health region data. This now give researcher the ability to perform appropriate “apples to apples” comparisons among health regions. This has the added advantage that researchers can focus their attention on searching for real differences between regions as opposed to those caused by subtle changes in their social and economic compositions. It is hoped that this will lead to the identification of best practices in health promotion and planning strategies amongst health region planners.

### 5.4 Next Steps

Future work includes the development of updated peer groups using 2001 census data and revised health region boundaries. This work will also attempt to follow the stability of the boundaries over time.

## REFERENCES

Andberg, MR. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Everitt, BS. (1993). *Cluster Analysis* (3<sup>rd</sup> Edition). Toronto: Halsted Press.

MacNabb, L. (2002). *Health Region Peer Groups*. Statistics Canada Internet Publication: 82-221-XIE.

SAS Institute Inc. (1999). SAS OnlineDoc® (Version 8). Cary, NC: SAS Institute Inc.