

**Review of proposed Data Selection
Methodology for the Consumers
Association of Canada 2004 Automobile
Insurance Study
12 October 2004**

Prepared by:
Carl James Schwarz, P.Stat.
Statistical Consulting Service
Statistics and Actuarial Science
Simon Fraser University
8888 University Drive
Burnaby, BC V5A 1S6

Table of Contents

1	Executive Summary	3
2	Introduction.....	3
3	Methodology Description	4
3.1	Description of Sampling Design for Profiles.....	4
3.2	Description of Sampling for Postal Codes.....	5
4	Methodology Review and Assessment	5
4.1	Review of Profile Selection	5
4.2	Assessment of Profile Selection.....	6
4.3	Review and Assessment of Postal Code Selection	6
5	Closing Statements.....	6
6	Curriculum Vitae for Reviewer: Carl James Schwarz, P.Stat.	7

1 Executive Summary

Having reviewed the quota-sampling scheme used for this study, I am satisfied on the following points:

- The quota-sampling scheme used to select driver profiles is an excellent method of ensuring that the sample reflects the correct gender/age/claims history composition, and cannot be further improved, given the available data and the structure of the study.
- The probability proportional to size methodology used to select postal codes is a well-established method that is well suited to the structure of this study.
- The profile selection and postal code selection methodologies were performed correctly.
- The study method – comparing fixed profiles across provinces – provides a fairer comparison than using actual provincial statistics. Further, the quota-sampling and postal code selection methodologies used are very appropriate for this type of study.
- Any remaining bias is a result of limitations on available data (e.g. Canada-wide claims histories not being available), which cannot be further reduced without more data.
- I am comfortable advising that the study proceed using this methodology.

2 Introduction

The goal of the study is a comparison of automobile insurance rates across provinces and territories in Canada. These rates are set by considering a wide number of factors such as the number of drivers that will be using a vehicle, the location where the vehicle is registered, the gender of the drivers, the claims histories of the drivers, and other factors some of which are proprietary to insurers. A particular set of drivers, vehicle types, claims history, and other factors is known as a profile. Comparing the “average cost of insurance” is a difficult task because of differences in the mix of profiles across provinces.

For example, a direct comparison across provinces of the average premium paid per driver is computed by taking the total premiums paid in that province divided by the total number of drivers in that province. This would be affected by many factors such as vehicle mix (some provinces have a higher proportion of light trucks vs. cars), urban vs. rural split (some provinces are more highly urbanized than other provinces), and the average number of vehicles per driver (some provinces have higher rates of vehicle ownership).

An alternate way to compare rates would be to take a set of representative profiles and find the average cost to insure this group across the nation. This is the method used in this study. By using a consistent profile set, differences in the average insurance cost become free of effects of profile differences among the provinces and territories.

This study has three components:

1. Obtain a representative set of profiles.

2. Select representative locations from each province.
3. Cost the set of profiles at each location and compute a province wide average. Comparisons will also be made across subsets of profiles (e.g. claim free drivers vs. high risk drivers) and across subsets of locations (e.g. urban vs. rural).

This review focuses on the first two components. It is based on a description of a sampling methodology provided as an MSWord document on 6 October 2004 and responses to emailed questions.

3 Methodology Description

3.1 Description of Sampling Design for Profiles

The study has access to 10,801 profiles from an online auto-insurance quote service. Each profile contains driver and vehicle characteristics such the number of drivers, the driver ages, the vehicle types, the miles driven, a record of driving offenses, and the claims history etc. No comprehensive list of profiles is available Canada-wide as it is often proprietary information.

This set of profiles could not be considered to be a random sample from profiles across Canada for a number of reasons. First, the set of 10,801 profiles likely suffers from self-selection biases. For example, drivers that have large insurance premiums may be more likely to go to an online quote service to obtain the best possible premium compared to claim-free drivers. Second, some provinces use a public insurance model where there is no need to obtain comparative quotes.

In order to mitigate some of the self-selection and other biases in these 10,801 profiles, a quota-sampling scheme was implemented to balance the profiles selected with respect to gender, age, and claims history, i.e. select profiles to ensure that the gender ratio, the age composition, and claims history match Canadian ratios.

Actual Canada-wide gender and age ratios were obtained, and approximately 700,000 claims-history records from a public-insurance provider were accessed which contained information on claims experience and other factors. No Canada-wide information on the claims experience of drivers is available. It was decided to subdivide this data into classes based on combinations of gender, age class (e.g. 16-19), and claims history (claims within the last year, claims in years 2 to 6 previously, and other histories). The proportion of all drivers in each class was found.

A set of 300 profiles was selected using quotas based on the proportions found previously. A preliminary analysis showed that approximately 400 drivers would be contained in 300 profiles, so the initial quotas were based on dividing just under 400 drivers into the gender/age/claims classes (e.g. the initial quota might be 2 male drivers in the 16-19 age class with claims in the last year).

Profiles were randomly selected without replacement. A profile was retained if the quotas for all drivers in the profile were not exceeded. For example, if a selected profile had two drivers, one a male 28 years of age with no claims, and the second a female 38 years of age with a claim in the last year, then it would be retained only if

both classes of gender/age/claim-history were below the quotas. When the number of drivers in a gender/age/claim class reached its quota, no further profiles are retained that would “add” drivers to any class that has reached its quota. For example, if the quota for males, aged 16-19, with claims in the last year was two, and two such drivers had been selected, then no further profiles would be retained that had such a driver.

Some minor adjustment was needed near the end of the sampling to ensure that exactly 300 profiles were selected (i.e. some minor changes to the quotas based on the total drivers in the selected 300 profiles).

3.2 Description of Sampling for Postal Codes

A master list of postal codes and the population within these codes was obtained. The list of postal codes was sorted alphabetically, and then a systematic sample of codes was selected by accumulating the population numbers and selecting postal codes whenever the accumulated population exceeded a multiple of 30,000. If a postal code had a population greater than 30,000 it could be selected multiple times.

4 Methodology Review and Assessment

4.1 Review of Profile Selection

The quota-sampling scheme was motivated by the large potential for bias caused by self-selection in the list of 10,801 profiles. The quote scheme as described was implemented correctly and will mitigate this bias over the gender/age/claim-history classes.

Limitations on the available data translated into the following limitations on the methodology:

- Even if the quota-sampling scheme were perfect, a residual bias would still exist because consumers in provinces with public insurance or who are not motivated to use the online-quote system are not included in the list.
- Similarly, the use of the information from a public-insurance provider to set quotas on claims history may not be strictly representative across Canada. This could be corrected if claims history data were available from all provinces, and included in the methodology.
- While the quota system will balance the profile selection over the gender/age/claims-history class, there is no guarantee that it will also balance the profiles over other factors such as vehicle type. Again, detailed data on vehicle type would be needed to correct this limitation.

This methodology also balanced vehicle type, but data was only available one-dimensionally (that is, the total balance of cars vs. trucks was matched to the Canadian ratio, but not also by age or gender). If more detailed Canadian data were available, the sampled profiles should be further tabulated by other factors (such as vehicle type) and the selected profiles compared against Canadian values. It should be

noted that even if one-dimensional tabulations do not indicate any major problems, there is no guarantee that higher-dimensional tabulations will match. For example, if one-dimensional tabulations by vehicle type and vehicle age seem reasonable, it is possible that a two-dimensional tabulation by vehicle type and vehicle age (together) may not match Canadian values.

4.2 Assessment of Profile Selection

This sampling methodology should be sufficient to derive a set of fixed profiles to be compared across provinces, despite the theoretical potential for residual biases to be present in the selected profiles. The key result of having a representative fixed set of profiles to be used for cross-Canada comparisons will have been obtained. It seems unlikely that the selected profiles will match some peculiarity in a province that would place it an extreme advantage or disadvantage relative to the other provinces.

4.3 Review and Assessment of Postal Code Selection

The described method for selecting postal codes is a well known method for selecting objects using a probability proportional to size (PPS) methodology. This ensures that postal codes with larger populations have a larger chance of being selected. It is important that if postal code is selected more than once, that it is included multiple times in the analysis when average costs across a province are computed.

The described methodology for selecting postal codes will have more locations selected in provinces with larger populations. This is not a requirement for a valid comparison across provinces – an alternate methodology would be select the same number of locations from each province which would require a different cutoff when looking at the cumulative population by postal code (i.e. some provinces would use a value of 30,000 while other provinces would use a value of 50,000). The advantage of the current methodology is that the set of locations are self-weighted so combined results over provinces (e.g. into regions such as Atlantic Canada) is easily done by merging the selected postal codes. If different selection intervals were used (e.g. 30,000 vs. 50,000), then a weighted average would need to be used if provinces are combined into regions.

5 Closing Statements

Based upon the documentation provided, the quota sampling scheme used for this study appears to have been implemented correctly and should remove large portions of a gender/age/claims history bias. There is still potential for residual bias because the available claims data were not from across Canada, but the main goal of obtaining a reasonable set of profiles for cross Canada comparison should be met.

The methodology for selecting postal codes is appropriate to obtain a representative set of locations within each province.

It should be noted that by using a fixed of profiles across Canada, that the “average” insurance cost obtained should not be interpreted as the average cost to the resident of that province. It is a synthetic average that deliberately avoids differences among provinces in their profiles. An analogy would be the fixed basket of goods used to construct consumer price indices across Canada – such a fixed basket of goods ignores any regional differences in what is actually consumed (e.g. some provinces may favor chicken rather than beef etc). Such a fixed basket (and fixed set of profiles) is useful in comparing costs among provinces.

I am comfortable advising that the study proceed using this methodology.

6 Curriculum Vitae for Reviewer: Carl James Schwarz, P.Stat.

Attached following.