Statistics Canada Statistique Canada

Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER

METHODOLOGY BRANCH

METHODOLOGY FOR CV LOOK-UP TABLES

HSMD-98-001E

Owen Phillips and Ritu Kaushal

Survey and Analysis Methods Development

Household Survey Methods Division

Statistics Canada

February 1998

The work presented here is the responsibility of the authors and does not necessarily represent the views or policies of Statistics Canada

METHODOLOGY FOR CV LOOK-UP TABLES.

Owen Phillips and Ritu Kaushal[1]

Survey and Analysis Methods Development
Household Survey Methods Division
Methodology Branch, Statistics Canada

ABSTRACT

Large surveys are faced with the problem of providing users of varying statistical backgrounds and needs with some indication of data quality. Various surveys have addressed this situation by supplying users with look-up tables of standard errors or coefficients of variation (CV's). Characteristics are often grouped by design effect to obtain the approximate measures of sampling error which appear in these tables. A methodology for the construction of such tables was developed based on generalized variance functions (GVF's), using geographic region as a proxy for design effect. This methodology has been applied to the Labour Force Survey (LFS), and an assessment of its performance is reported here.

[1] Owen Phillips and Ritu Kaushal, Survey and Analysis Development Section, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

MÉTHODOLOGIE APPLICABLE AUX TABLES DE RECHERCHE DES CV

Owen Phillips et Ritu Kaushal[2]

Développement des méthodes d'enquêtes et d'analyse
Division des méthodes d'enquêtes-ménages
Direction de la méthodologie, Statistique Canada

RÉSUMÉ

Un problème qui se pose lors des sondages importants consiste à fournir une indication quelconque de la qualité des données à un utilisateur dont les connaissances en statistique et les besoins varient. On s'est efforcé d'y remédier en remettant à l'utilisateur des tables de recherche sur l'erreur-type ou les coefficients de variation (CV). Les caractéristiques sont fréquemment groupées en fonction des effets du plan de sondage, de manière à donner une idée approximative de l'erreur d'échantillonnage qui apparaît dans les tables en question. On a mis au point une méthode pour concevoir de telles tables à partir de fonctions généralisées de la variance (FGV), en prenant les unités géographiques comme approximation des effets du plan de sondage. Cette méthode a été appliquée à l'Enquête sur la population active (EPA), puis on en a évalué l'utilité.

[2] Owen Phillips et Ritu Kaushal, Section du Développement des méthodes d'enquêtes et d'analyse, Division des méthodes d'enquêtes des ménages, Statistique Canada, Ottawa, Ontario, Canada, K1A 0T6.

1. INTRODUCTION

Most large surveys compile statistics for general release in a variety of different formats: for example aggregate or microdata, in either published or electronic form. With such releases, it is desirable to include data quality indicators or to provide the user with information and techniques to calculate these. From the Agency's perspective, the main challenges in presenting data are ensuring that the information provided is user friendly and does not compromise the confidentiality of respondents.

In the context of users with varying degrees of statistical sophistication, the term "user friendly" becomes difficult to define. Most users do not perform complex statistical analysis with the data but nevertheless need to be informed about the quality of the data. On the other hand, there are users who require accurate information about the quality of the estimates. Often the primary obstacle to obtaining detailed information is the confidentiality of the respondents. In essence, the information provided has to be available in different formats, levels of accuracy and complexity to satisfy all users. In this paper, we address only part of the problem; that is, providing information to most users assuming little knowledge or interest in statistical aspects.

Look-up tables of Coefficients of Variation (CV's), or other measures of sampling variability, are often provided with releases to give users a general idea of sampling error. Many surveys produce tables of approximate CV's or standard errors using a subset of characteristics, grouped according to design effects. The National Population Health Survey (NPHS) and the Survey of Labour and Income Dynamics (SLID) take this approach. Design effects are calculated for a subset of key survey characteristics and a conservative value (e.g. the 75th percentile of all such design effects) is chosen. Approximate CV's are then obtained by calculating the variance under simple random sampling, incorporating this conservative design effect into the calculation.

Approximate standard errors for selected Current Population Survey (CPS) estimates appearing in *Employment and Earnings* (U.S. Bureau of Labor Statistics, 1995) are also based on groupings of certain characteristics. Tables of standard errors, obtained using Generalized Variance Functions (GVF's), are given for major labour force characteristics by age, sex and, in some instances, race groupings. Also provided are the parameter estimates for regression models, allowing the user to compute standard errors based on the size of the estimate without interpolation. The Australian and Canadian Labour Force Surveys present similar information in the form of CV look-up tables.

Allowing the user to compute exact variances by including the necessary information has been proposed for more sophisticated users. However, this may require the release of sensitive information, which might in turn lead to record identification. Due to confidentiality restrictions, variations of this approach, such as the collapsing of strata (Mayda, Mohl, Tambay, 1996) and the inclusion of bootstrap weights on microdata files (Yung, 1997), have been considered. These methods are, however, geared toward more complex analysis and sophisticated users.

In this paper, an approach using generalized variance functions is described. Section two describes a simple methodology for CV look-up tables based on GVF's. This approach is evaluated using data from the Labour Force Survey (LFS) in section three. Section four concludes the paper.

2. GENERALIZED VARIANCE FUNCTION APPROACH

This section introduces GVF's and outlines a methodology for fitting these to data to produce CV look-up tables. The GVF is a model-based approach that draws on the relationship between estimates and their associated standard errors.

2.1 Generalized Variance Functions

The notion of a simple mathematical relationship between an estimate and its CV is given by Wolter (1985). He proposes five models appropriate to describing this relationship, and points to other models that might be considered. For the purpose of this discussion only two of these models will be considered with particular attention given to the first:

$$\log(CV) = \alpha + \beta \log(X) \tag{2.1.1}$$

$$CV^2 = \alpha + \beta /X \tag{2.1.2}$$

where X is the estimate of the number of individuals in the population possessing a particular characteristic and α and β are regression parameters to be estimated.

Valliant (1987) examines to some extent the theoretical properties of (2.1.1) and (2.1.2), presenting a stronger argument in the case of (2.1.2). He looks at some commonly used estimators in stratified two-stage sampling and shows that the relative variances of the resulting estimates are of the form given in (2.1.2). Valliant also offers the following justification of models (2.1.1) and (2.1.2):

If $P = X/N$ is the proportion of the population possessing some characteristic and F is the design effect associated with that characteristic under the particular sampling scheme, then the relative variance of p, the estimate of P, is approximated by

$$CV^2 = F(1 - P)/(nP) = -F/n + NF/nX$$

which is of the form in (2.1.2). This expression may be rewritten as

$$CV^2 = NF(1 - P)/nX$$

which, when P is small, is $CV^2 = NF/nX$. Taking the log of both sides results in an expression of the form in (2.1.1).

$$\log(CV) = \frac{1}{2}\log(NF/n) - \frac{1}{2}\log(X).$$

2.2 Methodology

The methodology described here borrows from those outlined by Wolter (1985) and Ghangurde (1981). Both authors suggest that, having chosen a suitable subset of characteristics upon which the GVF's are to be modelled, characteristics be grouped according to some defining measure, such as similar design effects. The goal here is to group characteristics in such a way that they will follow a common model. In most surveys, design differences between geographic areas result in these areas being confounded with the design effect. As design effect may be a foreign concept to the non-statistical user, the use of geographic areas as grouping criterion adds to the desired simplicity of the tables.

For example, in the case of the Labour Force Survey provinces are strongly correlated to design effects. While resulting fits may not be as good, the use of provinces avoids the cumbersome use of two sets of tables; one for the design effects and the other for the CV's themselves. *The Labour Force Australia* (Australian Bureau of Statistics, 1995) also uses geography based GVF's to provide CV lookup tables.

The choice of models and regression techniques to be employed in fitting these models should be considered. Different GVF's should be applied to the data in order to obtain the best fit. Wolter points out that despite the lack of theoretical underpinnings, GVF's have been used successfully by various surveys for many years.

As to the actual fitting of said models, Wolter suggests that Ordinary Least Squares (OLS) may not be appropriate for linear models as it attributes too much weight to smaller estimates. Instead, he suggests weighted or iterative methods should be used to arrive at the parameter estimates. Valliant's empirical study supports this notion in the case of (2.1.2), favouring a weighted regression with weights inversely proportional to the estimated CV, but finds little difference between the fitting of (2.1.1) using OLS and an iterative fit of its non-linear counterpart. Ghangurde shows results similar to Valliant's in applying (2.1.1) to LFS data.

Generally, conservative estimates of CV's are sought. Often the target is to produce CV's that are conservative for 75% of all estimates. In an effort to produce conservative CV's, all of those observations for which the CV of the estimate is less than the fitted CV are dropped, and the model is refit using the remaining data to obtain the final model.

The resulting model can be used for the purpose of producing CV look-up tables and/or software. For many electronic products it is possible to incorporate functions that calculate the approximate CV for a given estimate. It is desirable to provide quality indicators that can be associated with an estimate. However, a compromise has to be made between useful information provided and the cost and space required to store the extra information. In such cases, GVF is a useful technique because a single function using two parameters can be used for a geographic region without storing any additional information.

To produce the CV look-up table, select those CV's which delineate the ranges of estimates (e.g. 1%, 2.5%, 5%, 10%, 15%, 20%, ...), and calculate the size of the corresponding estimate using the final model. A graph illustrating this procedure is given in Appendix B. The body of the resulting table consists of the ranges of estimates. Column headings are the CV's which cut off the given ranges under the final model. Each line of the table presents the estimate ranges for a particular geographic region, and hence represents estimates resulting from different models. The table defines a step function, where the approximate CV of a given estimate is equal to that corresponding to the lower value of the range into which it falls. An example is given in Appendix C and is discussed further in the next section.

3. APPLICATION TO LFS DATA

CV look-up tables for monthly totals and annual averages, at the Canada and province level, are included in LFS products. Estimates and CV's used to fit GVF's were those for the period of October 1995 to September 1996 (the most recent twelve months for which data was available). These included: labour force characteristics by province, age and sex; employment by province, sex and industry; employment by province, sex and occupation; and employment by province, age, sex and full-time/part-time status. Labour force characteristics include the number of people in the labour force, the number of employed and the number of unemployed.

CV's were modeled as a function of the estimate with which they are associated using (2.1.1) and (2.1.2). The two GVF's were fit using OLS regression. Model (2.1.1) gave better fits in comparison to (2.1.2) based on R-square values and diagnostics.

Initial fits for the ten Provinces and Canada level estimates gave R-square values of approximately 0.9 using (2.1.1). In an effort to produce CV's that were conservative, observations for which the jackknife CV (CV's for LFS estimates are obtained using the jackknife variance) was lower than the value obtained using the regression coefficients were dropped, and the model was fit to the remaining data. R-square values for the new fits ranged from .9513 for Saskatchewan to .9776 for Canada. Model fittings are summarized in Appendix A.

The resulting CV table was constructed by evaluating the eleven models (Canada and the Provinces) for CV's of 1%, 2.5%, 5%, 10%, 20%, 30% and 50%, to establish approximate ranges of estimates. The table for CV's for monthly totals is given in Appendix C.

Interest in the table for monthly totals has increased as it is now being published in the monthly LFS press release. In order to assess the performance of this table, LFS data for the five month period following the construction of the table was used. The assessment includes estimates and CV's for selected labour force characteristics and full-time/part-time employment, by age and sex, at the Canada and province level, as well as employment by industry and employment by class of worker at the Canada level. The goal was to produce CV's which would be conservative in 75% of the cases, i.e. the table would result in a CV that was greater than the jackknife CV for 75% of all estimates. As CV's for estimates of rates and month-to-month change are often obtained using the table in conjunction with some formula, an assessment was done on the performance of the table in relation to jackknife CV's for monthly totals as well as rates and month-to-month changes. However, overestimation may be due, in part, to the approximations given in the formulae. Hence, the interplay of the tables and the formulae is unclear. For this reason, we restrict our attention to the assessment of CV's of monthly totals. The percent overestimates by geographic region and by characteristic were calculated for the five month period, and are given in Table 3.1 and Table 3.2.

Table 3.1 Overestimates by Geographic Region

Geographic Region	Number of Estimates	Number of CV's Overestimated by look-up tables	Percent Overestimated
Canada	435	404	92.87
Newfoundland	300	290	96.67
PEI	300	289	96.33
Nova Scotia	300	285	95
New Brunswick	300	290	96.67
Québec	300	290	96.67
Ontario	300	297	99
Manitoba	300	297	99
Saskatchewan	300	288	96
Alberta	300	298	99.33
British Columbia	300	294	98

Table 3.2 Overestimates by Characteristic

Type of Estimate	Number of Estimates	Number of CV's Overestimated by look-up tables	Percent Overestimated
In labour force	660	660	100
Employment	660	658	99.7
Unemployment	660	603	91.36
Full-time	660	657	99.55
Part-time	660	630	95.45
Employment by industry	115	94	81.74
Employment by class of worker	20	20	100

As is evidenced in Table 3.1 and Table 3.2, the methodology employed here has resulted in CV's that are too conservative. The CV look-up tables, when applied to recent data, have produced overestimates for a far greater percentage of estimates than the desired 75%. The percentage of overestimates are closer to the goal for smaller estimates like unemployment and industry. The problem results from the step function. While the line given by the second fit results in conservative estimates, the construction of the step function from that line results in more overestimation than anticipated. If only the function is being provided, the second fit may suffice. Applying it to the table format results in overestimation. Defining smaller steps or using 100% of the data to determine the regression line are possible options.

Varying the size and range of steps for different geographic regions might be useful. The table given in Appendix C excludes some estimates for the smaller provinces at 1% CV's because the population is smaller than the size of an estimate of a quality better than or equal to 1%. The size of the steps could be broken down further in the ranges corresponding to the smaller CV's for the larger provinces and Canada, and values exceeding the size of the labour force of a given province should be omitted from that line of the table. A more detailed breakdown of the ranges corresponding to larger CV's could be considered for the smaller provinces.

The absolute differences between the CV's obtained from the jackknife method and those obtained using the CV look-up table quantifies the accuracy of the look-up tables. Table 3.3 shows the distribution of the absolute difference for Canada and the provinces. Note that for Canada, where the CV's tend to be smaller relative to the provinces, a high proportion of the table CV's have a small difference from the jackknife value. Among the provinces, the smaller provinces have a higher proportion of CV's with larger differences, however CV's at this level tend to be large and a difference of 4 or 5 for a provincial level estimate does not have the same impact as it might at the Canada level. Overall, about 80.0% of the CV's obtained from the tables are within 3 percentage points of the jackknife CV and about 94.4% are within 5 percentage points. Given that most of the differences are the result of overestimation, these differences tend toward the conservative side.

Table 3.3 Cumulative percentage of CV's by absolute difference and geographic region

Geographic Region			$\lvert CV_{table}-CV_{jackknife} \rvert$		
	< 1	< 2	< 3	< 4	< 5
Canada	67.4	89.7	97.2	97.5	98.9
Newfoundland	17.7	36.7	70.7	87	94.7
PEI	22	42.3	66	85	92.7
Nova Scotia	27.3	61	74.3	83	91
New Brunswick	27.7	59.7	72.7	80.3	92.7
Québec	28.7	78	87	92	98
Ontario	40.3	72.7	88	92.3	97
Manitoba	12	65.3	79.7	86	94
Saskatchewan	12.7	61.3	75.3	81.3	87.3
Alberta	15.7	68	79.3	85.3	94
British Columbia	14.3	69.3	82.3	90.3	95.7
Overall	27.7	65	80	87.7	94.4

4. CONCLUSIONS

The GVF technique is simple, easy and useful. In effect, it smooths the variance function and provides a good approximation. The need to estimate design effects is eliminated; variances for a subset of estimates are sufficient. The resulting tables and/or function can accompany both aggregate or microdata in electronic or paper form.

CV look-up tables or GVF's may be easier for users when common geographic delineations, like provinces, are used. Simple formats might encourage users to be better informed on the quality of data they are using. The methodology presented here addresses the need for simplicity, but perhaps at the cost of accuracy. With some modification, the needs of both the agency and a majority of its clients could be met.

ACKNOWLEDGMENTS

The authors would like to thank Wesley Yung and Joseph Duggan for their helpful comments.

REFERENCES

AUSTRALIAN BUREAU OF STATISTICS (1995). *The Labour Force Australia - January 1995.*

GHANGURDE, P.D. (1981). Models for estimation of sampling errors. *Survey Methodology*, 7, 177-191.

MAYDA, J.E., MOHL, C., TAMBAY, J.-L. (1996). Variance estimation and confidentiality: they are related! *The Statistical Society of Canada: Proceedings of the Survey Methods Section, 1996*, 135-141.

U.S. BUREAU OF LABOR STATISTICS (1995). *Employment and Earnings - May 1995.*

VALLIANT, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82, 499-508.

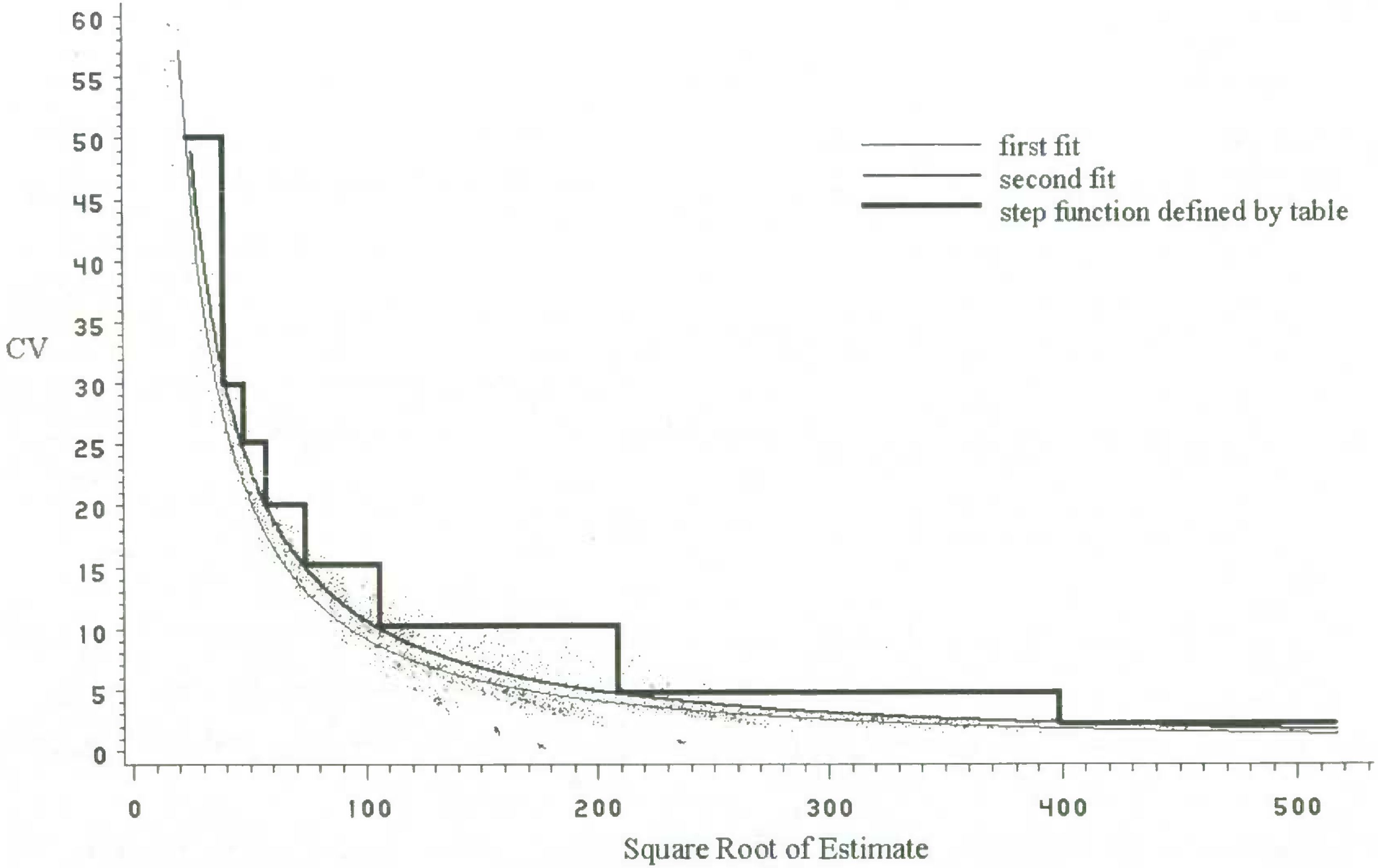
WOLTER, K.M. (1985). *Introduction to Variance Estimation*, New York: Springer-Verlag.

YUNG, W. (1997). Variance estimation for public use files under confidentiality constraints. (To appear in) *The American Statistical Association: Proceedings of the Survey Research Methods Section, 1997.*

APPENDIX A: Parameter Estimates and R-square Values for Model Fittings

Region	Parameter/R-square	First Fit	Second Fit
Canada	α	8.659632	8.455907
	β	-0.602375	-0.572652
	R-square	0.9151	0.9754
Newfoundland	α	7.521956	7.546088
	β	-0.602375	-0.553143
	R-square	0.8867	0.9776
PEI	α	7.394402	7.250339
	β	-0.643917	-0.603692
	R-square	0.8938	0.9361
Nova Scotia	α	8.031839	8.028516
	β	-0.618634	-0.601018
	R-square	0.9155	0.9765
New Brunswick	α	7.806593	7.0779354
	β	-0.613005	-0.592926
	R-square	0.9155	0.9744
Québec	α	8.854297	8.712533
	β	-0.616769	-0.590838
	R-square	0.9196	0.9771
Ontario	α	8.836005	8.668418
	β	0.616813	-0.588071
	R-square	0.915	0.9725
Manitoba	α	8.382124	8.294125
	β	-0.651971	-0.622892
	R-square	0.906	0.9593
Saskatchewan	α	8.108821	7.958403
	β	-0.627208	-0.592335
	R-square	0.8963	0.9513
Alberta	α	9.01314	8.849388
	β	-0.661196	-0.627566
	R-square	0.905	0.9586
British Columbia	α	8.801921	8.685786
	β	-0.62895	-0.602645
	R-square	0.9142	0.9736

APPENDIX B: CV vs Square Root of Estimate - Newfoundland



APPENDIX C: CV's for Estimates of Monthly Totals for Canada and the Provinces

	1%	2.5%	5%	10%	15%	20%	25%	30%	50%
Canada	2587.6	522.4	155.7	46.4	22.9	13.8	9.4	6.8	2.8
Newfoundland	840.9	160.4	45.8	13.1	6.3	3.7	2.5	1.8	0.7
PEI	164.4	36	11.4	3.6	1.9	1.2	0.8	0.6	0.3
Nova Scotia	633	137.8	43.5	13.7	7	4.3	3	2.2	0.9
New Brunswick	499	106.4	33.1	10.3	5.2	3.2	2.2	1.6	0.7
Quebec	2535.9	537.8	166.4	51.5	25.9	15.9	10.9	8	3.4
Ontario	2521.7	530.9	163.3	50.3	25.2	15.5	10.6	7.8	3.3
Manitoba	606.5	139.3	45.8	15	7.8	4.9	3.5	2.6	1.1
Saskatchewan	684	145.6	45.2	14	7.1	4.4	3	2.2	0.9
Alberta	1330.6	309	102.4	33.9	17.8	11.2	7.9	5.9	2.6
British Columbia	1817.1	397.3	125.8	39.8	20.3	12.6	8.7	6.4	2.8