11-617E
no.88-02
c. 1

Statistics Canada
Statistique Canada

STATISTICS STATISTIQUE
CANADA CANADA
MAY MAI 19 1998
LIBRARY
BIBLIOTHÈQUE

NOT FOR LOAN
NE S'EMPRUNTE PAS

WORKING PAPER NO. BSMD-88-002E

CAHIER DE TRAVAIL NO. BSMD-88-002E

METHODOLOGY BRANCH

DIRECTION DE LA MÉTHODOLOGIE

STATISTICS CANADA STATISTIQUE CANADA
AUG 8 1996
LIBRARY
BIBLIOTHÈQUE

STATISTICS CANADA'S INTERNATIONAL CONTRIBUTIONS TO INFORMATICS AND METHODOLOGY

by

I. Sande

January 1988

STATISTICS CANADA'S INTERNATIONAL CONTRIBUTIONS TO INFORMATICS AND METHODOLOGY

Prepared by
I.G. Sande
for presentation at
Statistics Canada,
January 1988.

L'apport de Statistique Canada dans le domaine de l'informatique et de la méthodologie sur la scène internationale

Sommaire

Les problèmes qu'éprouve Statistique Canada en matière de collecte, de traitement et de diffusion des données sont les mêmes que ceux des organismes statistiques internationaux. D'autres pays s'intéressent à une bonne partie de nos travaux axés sur la résolution de ces problèmes et ils s'en inspirent. Des exemples de ces problèmes feront l'objet d'une discussion ultérieure. Pour garantir la confidentialité des données, il faut éviter la diffusion des renseignements personnels. Les étapes du contrôle et de l'imputation traitent de la détection et de la correction automatisées des erreurs dans les données. Certaines enquêtes ont des plans de sondage très complexes et les méthodes traditionnelles d'analyse des données ne sont pas valables. La plupart des données recueillies sur une base régulière et plus d'une fois par année sont à l'origine de séries comportant des fluctuations saisonnières qui ont tendance à dissimuler les tendances à long terme et qui doivent être éliminées. Les données utilisées à des fins statistiques ne diffèrent pas des autres types de données, par exemple les enregistrements personnels, mais peuvent être utilisées de manière plus efficace si elles sont remaniées. Pour procéder à la collecte et au traitement de statistiques dans un pays aussi vaste que le Canada, il faut avoir en mains une gamme de cartes, de fichiers et de systèmes géographiques. Il faut disposer les statistiques assemblées à l'échelle régionale de façon à faire ressortir les variables régionales et à mieux évaluer la répartition des caractéristiques dans l'ensemble du pays. Les initiatives de Statistique Canada pour résoudre ces problèmes feront l'objet d'une discussion non technique.

1. INTRODUCTION

The problems that Statistics Canada experiences in the collection, processing and dissemination of its data are shared by the international statistical community. Much of our work in dealing with these problems has attracted attention in other countries and has influenced their own efforts.

This discussion will cover some of the work that has been done at Statistics Canada in statistical theory and methods and computer applications which have gained international recognition. It is not comprehensive. For lack of presentation time, some interesting and important topics have been omitted. It should always be remembered that a great deal of original work goes on at Statistics Canada and that there is a constant exchange of ideas between Statistics Canada and other statistical agencies and institutions. Often very tiny ideas influence thinking in a wide variety of places.

The contributions we will deal with concern

(a) Confidentiality - the problem of preventing disclosure about individuals in the tabulations the Bureau produces.

(b) Edit and imputation - the problem of automated detection and correction of errors in the data.

(c) Analysis of data from surveys with complex designs - the problem of drawing reliable conclusions from the data when the traditional assumptions are violated.

(d) Seasonal adjustment - the problem of removing seasonal patterns from data collected regularly over a long period of time, so that long term trends are revealed.

(e) Statistical database management - the problem of organizing data for statistical purposes.

(f) Geocartographics - the problems of automating the geographic aspects of data collection in Canada, and of displaying data to bring out the variation in characteristics across the country.

This talk is not meant to honour the individuals who were largely responsible for the innovations we are discussing. Rather it is meant to celebrate the work of the Bureau as a whole, for many have contributed in some way. For this reason, personal references have been omitted.

2. PROTECTION OF CONFIDENTIALITY

The confidentiality of respondent data is a prime concern of Statistics Canada and is specifically required by the Statistics Act. Administrative data, which are collected outside the Bureau for other purposes, but acquired by the Bureau to avoid unnecessary contact with respondents, are also protected. There is no doubt that the Bureau's ability to collect data and compile good statistics would be seriously compromised if respondents were not guaranteed that data on individual respondents would not be disclosed, either deliberately to other agencies or inadvertently in published statistics. The responsibility for the protection of the confidentiality of individual data is taken very seriously.

The Bureau produces mainly tabulations of various kinds and (sometimes) public-use files or microdata, and there are different protection methods associated with each of these. Statistics Canada has contributed to the development of confidentiality procedures for tabulations.

Tabulations come in two types: frequency and magnitude. In frequency tabulations, we simply count how many members of a population fall into each of a number of categories. For example, a tabulation by sex and income of all the members of particular club might be

		SEX		TOTAL
		M	F	
Income ($,000's)	Less than $25	0	2	2
	$25 - $35	6	0	6
	$35 - $50	15	14	29
	$50 - $70	13	3	16
	Over $70	2	1	3
	TOTAL	36	20	56

Example of Frequency Tabulation

Notice that the highest paid lady is probably readily identifiable and she might not like having her salary disclosed. Likewise, the two lowest paid ladies are also likely to be readily identifiable and wouldn't want their circumstances revealed.

To cope with such problems, the technique of random rounding was originated at Statistics Canada. We are all familiar with systematic rounding - for example, in the calculation of percentages to the nearest whole number. On the other hand, when we give our ages, we usually round down and give our age at the last birthday. In random rounding, numbers are usually rounded to multiples of 5 and an imaginary 5-sided die is tossed to decide whether to round up or down. Thus numbers ending on 0 or 5 stay the same and

When the last digit of the number is	1	it has	4	chances in 5 to be rounded to	0	Otherwise it is rounded up to the next multiple of 5.
	2		3		0	
	3		2		0	
	4		1		0	
	6		4		5	
	7		3		5	
	8		2		5	
	9		1		5	

The statistical validity of this procedure lies in the fact that if we rounded the same table many times, we would, on the average, get the correct table. If we had to publish the same table repeatedly and did the random rounding from scratch every time, some industrious individual could recover the original table by averaging over all the published ones.

A rounded version of the table in the example is

		SEX		TOTAL
		M	F	
Income ($,000's)	Less than $25	0	5	5
	$25 - $35	5	0	5
	$35 - $50	15	15	30
	$50 - $70	15	0	15
	Over $70	5	5	10
	TOTAL	40	25	65

Frequency Table with Random Rounding

Now the one well-paid lady and two underpaid ladies have been well disguised, but the total number of people in the club bears no resemblance to the original 56. We could round the totals separately, and this may be desirable, but then the rows and columns don't add.

For this problem, the technique of controlled rounding has been developed. This rounds up or down to a multiple of 5 and the totals are within 5 units of the original. A version of the original table with controlled rounding is

		SEX M	SEX F	TOTAL
Income ($,000's)	Less than $25	0	5	5
	$25 - $35	5	0	5
	$35 - $50	15	15	30
	$50 - $70	15	0	15
	Over $70	0	0	0
	TOTAL	35	20	55

Frequency Table with Controlled Random Rounding

In fact, this table much closer to the original, not much different from the first rounded version, and more satisfactory when it comes to totals.

The controlled rounding technique, which involves accumulating totals of entries by columns and then rounding and then subtracting successive sums from each other to recover the individual rounded entries, can be extended to larger and more complex tables in 3 or more dimensions; however, in higher dimensions the totals may differ by far more than 5 from the originals.

Statistics Canada is also recognized as a leader in the development of automated techniques for dealing with magnitude tabulations. An example of a magnitude tabulation is the following

	EASTERN CANADA	MIDDLE CANADA	WESTERN CANADA	ALL CANADA
Fishing	66	17	50	133
Hunting	1	25	45	71
TOTAL	67	42	95	204

Total Revenue Millions of Dollars

Example of Magnitude Tabulation

Here we are not counting anything, but giving the total revenue realized by particular industries over all businesses engaged in that industry. Of course, we may very well know how many businesses there are, and that is the problem. Suppose there are only 3 companies engaged in Fishing in Eastern Canada (there are not) and that "market" intelligence tells us that two are very small indeed. Then we know, to a reasonable degree of approximation (say within 10 or 20% of the true value) what the revenue of the largest one is, and this the company in question may not like. The cell is then said to be "sensitive", so we decline to publish it, and propose to publish instead the following table

	EASTERN CANADA	MIDDLE CANADA	WESTERN CANADA	ALL CANADA
Fishing	X	17	50	133
Hunting	1	25	45	71
TOTAL	67	42	95	204

Total Revenue in Millions of Dollars

Table with Sensitive Cell Suppressed

This is obviously useless, since some elementary arithmetic supplies the missing value. The vaiues in some other cells must also be suppressed in order to protect the sensitive cell, such as

	EASTERN CANADA	MIDDLE CANADA	WESTERN CANADA	ALL CANADA
Fishing	X	X	50	133
Hunting	X	X	45	71
TOTAL	67	42	95	204

Total Revenue in Millions of Dollars

Table with Complementary Suppressions

As tables get more complicated, the problem of protection gets more complicated also. There are three elements to the problem: the rules for deciding when a cell is sensitive; the auditing of a publication pattern, to see if a disclosure has taken place; automated determination of which other cells must be suppressed to avoid disclosure, which we call a complementary suppression pattern.

The sensitivity criterion is usually expressed as a concentration rule. For example, a cell is sensitive if the largest 3 respondents contribute more than 80% of the total (this example is artificial). This criterion is itself usually regarded as sensitive by statistical agencies. The idea is that the aggregation with other respondents which are not too different in terms of their contribution helps to protect the confidentiality of the individual respondent data. This type of sensitivity criterion is amenable to mathematical analysis which makes possible the development of automated techniques for the problem.

A suppression pattern is the arrangement of X's in the publication. It may be determined manually or may be the same one as was used last month or last year for the same survey. Under these circumstances, we may want to audit the pattern with the current data to see if there are any approximate disclosures. A programme developed at Statistics Canada does the arithmetic required to determine the smallest and largest values which each suppressed cell could assume, given the remaining data. An audit of the last table with 4 cells suppressed would yield the next table.

	EASTERN CANADA	MIDDLE CANADA	WESTERN CANADA	ALL CANADA
Fishing	41 - 67	16 - 42	50	133
Hunting	0 - 26	0 - 26	45	71
TOTAL	67	42	95	204

Total Revenue in Millions of Dollars

Audited Table with Ranges of Possible Values

Of course, some simple arithmetic can produce the audit here, but in general it is a complex operation. If the bounds coincide, there is disclosure. If they are too close for comfort, say within a few percent, there may be approximate disclosure. We have to assume that there is always some smart and interested individual who is trying to pry disclosures from our tables and who wants the information very badly. Users have to learn to appreciate the skill and expertise involved in withholding information from them.

The determination of a pattern of complementary suppressions required to protect a set of sensitive values is a more difficult problem. For very small tables, we can simply examine all the possibilities. But publications of interest may contain several thousand or tens of thousands of entries, and even automated exhaustive search methods are not practical. Techniques from Operations Research have been adopted and programs developed to deal with the problem of determining the publication pattern.

3. EDIT AND IMPUTATION

The correction of identifiable errors and omissions in survey data can be a very costly process. Returning to a respondent, after the main data collection operation is over, is time-consuming and much more expensive than the initial collection. Re-contacting some respondents often proves impossible. On the other hand, known errors in the data are unacceptable. They seriously impede data analysis, cast doubt on the conclusions, and greatly inconvenience the users.

To deal with the problem of errors in the responses, a strategy of "edit and imputation" is usually adopted. Editing is the process of detecting incorrect, inconsistent or missing data. Imputation is the process of modifying responses where such data occur so that the resulting record is "clean", i.e. it will pass the edit.

The edit is simply a collection of relationships (also - and confusingly - called edits) between items in the response, each of which must be satisfied in order for the response to be considered valid. For example, in a survey of Stat Can employees, some of the edits might be:

Age less than 80

Age greater than 15

If Education = University, then Age greater than 20.

If some of these relationships do not hold, the response fails the edit. It must then be decided which items are to be considered in error, which may or may not be obvious. Erroneous or missing items are imputed in a way that is statistically acceptable. This means that while the imputation may not result in a correct response in individual cases, the overall quality of the tabulated data will be improved.

Why does a survey organization like Statistics Canada adjust the data rather than leave it the way the respondent answered? Adjustment, particularly changing the individual records, appears in some sense to be tampering with the data. There are a number of good reasons why it should be done.

When analysing data, adjustment for missing data is unavoidable. Suppose for example age data is missing for 2% of the respondents on the census. Demographers would have to distribute these people to age groups in order to make any estimates or projections. Whatever they do would be a form of adjustment. If they simply use the good data and reject the missing data, this is effectively an adjustment and a poor one because it will have underestimated the total population. A better, but also simple, approach would have been to use the total count as an estimate of the population size and the clean data as the source of information on the age distribution. This adjustment which is a simple form of weighting is a better than simply throwing the incomplete data away but can still be improved upon.

Adjustment could be applied by each user of the data separately. Such an approach however would be very inefficient. Each user would spend considerable effort doing his or her own correction and each user would make a different correction. In effect, cleaning the data prior to release is a service to the users, saving them the inconvenience of navigating around data problems, and ensuring consistency between users.

The survey organization can in general do a better job of adjustment than the user. The data producer has access to all the confidential information which cannot be released to the user and can develop special expertise in this area of data adjustment techniques.

For these reasons Statistics Canada has a long tradition of cleaning completed questionnaires. Originally this was done clerically and some surveys still do extensive clerical cleaning operations. With the introduction of high speed computers, major surveys automated much of this. As often happens, the initial automation closely followed the clerical procedures. Complex rules were developed to replace the procedures, often incomplete or undocumented, of clerical staff. But complex rules in some instances lead to unmanageably complex computer programs as all possible errors and exceptions were allowed for.

3.1 Edit and Imputation of Qualitative Data

In the 1971 Census, the extent and complexity of the edit and imputation process seriously delayed the final release of the data.

Responding to this crisis, a standard methodology was developed at Statistics Canada to treat qualitative variables (such as age, education and marital status) that comprise the major part of the census data. This methodology was implemented as the CANEDIT system and was first used in the processing of the 1976 Census. This resulted in the 1976 Census being processed in a considerably shorter time than the 1971 Census, and also in a much better understanding of all the implications of editing and imputing survey data.

The CANEDIT system proved very effective but did not cover all situations. It could not handle quantitative data nor could it efficiently handle qualitative data in a long code list. It also placed overly restrictive conditions on what could be imputed. In some situations the true response is self-evident from the erroneous response.

In 1981 the CANEDIT system was re-used very effectively. This saved substantial developmented resources and resulted in major reduction in the processing time. At the same time a second system, SPIDER, was also developed to handle variables not amenable to treatment by CANEDIT. This system used some of the principles of the methodology first set out. As a system it is more flexible but less rigorous, allowing greater use of judgement in how the imputations take place.

In 1986 both systems were re-used. This re-use has allowed further substantial gains in productivity.

The principles established for imputing data were that

1) The data in each record should be made to satisfy all the edits by changing the minimum amount of information;

2) As far as possible the distributional characteristics of the data should be maintained;

3) The imputation rules should be derived from the edit rules.

The first two of these principles were aimed at optimising the quality of the data while at the same time minimising the extent to which the survey taker was interfering with the data as collected. The third principle was aimed at preserving the consistency of a record and thus avoiding a situation in which one round of modifications to individual records generates a new set of errors, a situation which at worst can create an unending loop and certainly can lead to far greater adjustment than originally intended.

The CANEDIT system is generalized, i.e. not specific to a particular survey. Thus the system is re-used without change from one census to the next, saving considerable developmental resources. Such software is common today with the rapid growth in the use of micro-computers. Fifteen years ago, when this development was initiated, this approach was very forward looking.

The CANEDIT system consists of the following modules:

(a) The Edit Rule Analyser

This module takes the proposed edit rules and tests them to ensure that they are logically consistent and contain no redundancies. It also provides the specifier and later stages of the system with all the edit rules implied by the initial set. This ensures that the edit will operate successfully if some data are missing and allows the specifier to double check the rules. Once checking has been completed the rules are incorporated into the system.

(b) The Edit

The edit step checks whether or not each record passes all the edits and, when a record fails, flags which edits have been failed.

(c) Selecting the Field to Impute

This module identifies which fields (items or variables) need to be changed to make a particular erroneous record clean. The algorithm is designed to find the minimum number of changes required for each record, using the complete set of edits from the Edit Rule Analyser.

(d) Imputation

The imputation is a 'hot-deck' method. this means that the values of the variables to be changed in an erroneous record are replaced by values from a good or "donor" record. The system is designed to find one donor record that matches as far as possible the erroneous record. All the variables to be changed are changed at the same time. The matching process used to find the donor uses all the variables that are implicated in the failed edits but are not being changed. This guarantees that the changed record will pass the edit once imputation is complete.

If a single donor cannot be found the system resorts to imputing each variable separately while still ensuring that the final record will pass edit.

The CANEDIT system laid the ground work for all subsequent development of generalized edit and imputation systems at Statistics Canada and indeed had a major impact on developments in other statistical agencies around the world. Its robustness has been demonstrated by its re-use totally unchanged in both the 1981 and the 1986 Censuses.

3.2 Edit and Imputation of Numerical Data

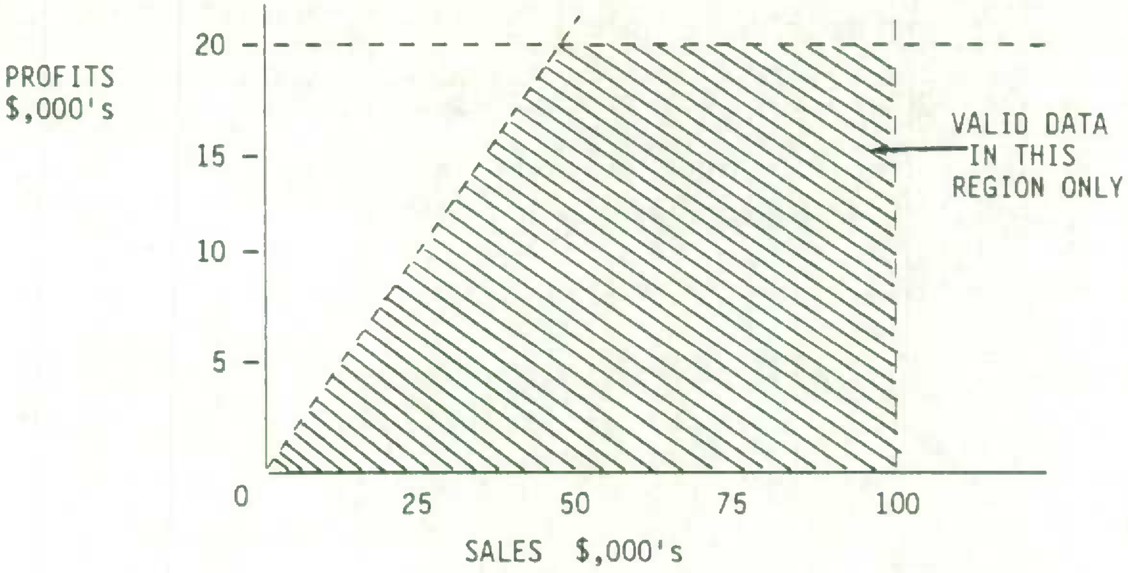
When the response to a question is a quantity, the approach used for qualitative (or categorical) data by the CANEDIT system breaks down. The reason is that quantities cannot, for some purposes, be broken down into manageable numbers of categories. If we are interested in the total revenues of a certain industry we want to add up the revenues of all businesses in that industry, and we cannot add categories.

For quantities, edits tend to be expressed in terms of ranges. To give a very simple example, suppose we are surveying businesses and we want to know for each business: total sales and profits in dollars.

Our edits might look like:

SALES > 0
SALES ≤ $100,000
PROFITS ≥ 0 (we are optimistic)
PROFITS ≤ 40% of SALES
PROFITS ≤ $20,000

We can draw a picture of this:



Example of Numerical Edits

Any combination of profits and sales which do not fall in the hatched area will not pass the edit.

Suppose the data are:

SALES	=	$30,000
PROFITS	=	$15,000

These data would fail the edit PROFITS $\leq$ 40% of SALES. We could examine the survey return to see if the numbers had been captured correctly, or phone the respondent to query the numbers. These two options are quite expensive and take a lot of time, and we don't want to bother the respondent more than we have to; so we decide instead to change either SALES or PROFITS.

Notice that if a large number of responses failed this edit we would be suspicious that something was wrong with the edit or the procedure, and investigate accordingly. However, if the number of edit failures is small, our experience may guide us to make an appropriate "correction" even if such a fix-up is not really correct. The difference to the total value of SALES or PROFITS over all responses may be insignificant.

We can also argue that, if the difference is insignificant, it is hardly worth making a correction. However, the data may be tabulated in many ways and inconsistencies may show up in the tables.

The situation gets a lot more complex when we consider more realistic data. There are many more data items and many more edits. We can no longer draw a picture of the situation. However, the problems remain:

(a) find the responses which fail the edit;

(b) decide which items in a defective response should be changed in order to give a response which does not fail the edits;

(c) decide what values should replace those which we decide to change.

In the late 70's, a prototype system was designed and constructed for the numeric edit and imputation problem, which we refer to as the NEIS (Numeric Edit and Imputation System).

If the edits are linear, i.e. of the form

$$aX + bY + cZ + \ldots\ldots \leq d$$

where X, Y, Z are quantity data items, and a, b and c are constants (i.e. fixed numbers); and if we agree that we want to change as few items as possible when a response is defective, then the system will

(i) analyze the edits, to ensure they are not contradictory or redundant;

(ii) apply the edits to the data;

(iii) decide which items have to be changed using a method which requires only the minimal set of edits;

(iv) search among the good responses for one which is close to the defective response, from which items can be copied to replace the "erroneous" items in the defective record;

The last step is the imputation step. Unlike CANEDIT, the sytem does not make a random choice among matches, but first decides which items to match and then locates the closest 15 (the number can be changed) complete responses. They are then tested one by one as potential donors of required data. The test is very simple: if the fixed-up response passes the edit, the imputation is successfully completed.

The NEIS is quite sophisticated and actually possesses more features than those outlined above. However it is only a prototype and it is clear that, on its own, cannot be used easily as "the" numerial edit and imputation system for any application. Many features have been borrowed for use in other systems, however.

Recently we have embarked on a project to build a Generalized Edit and Imputation System (GEIS). This system will include nearly all the features of the NEIS and in addition will offer more options as regards types of edit and methods of imputation. The price is that the mathematical basis on which the NEIS is built cannot be extended to include many of the desired options. For example, if the edits are not all linear, they cannot readily be analyzed nor do they permit the use of the error localization method of the NEIS. Ad-hoc procedures would need to be defined. Solving these problems poses a considerable challenge for the next few years.

4. ANALYSIS OF DATA FROM COMPLEX SURVEY DESIGNS

Data analysis is the process of summarizing and drawing conclusions from data. It is one of the prime concerns of theoretical and applied statisticians. Data analysis can tell us about the relationship between smoking and heart disease, under what conditions an industrial process is most efficient, and how age, education and marital status are related to employment.

The methods of data analysis that were developed over the last hundred years or so generally assumed that observations were made completely at random. The units being observed, such as patients in a medical trial, students in a psychology experiment, plots of ground in an agricultural experiment, were all "equally important" as representatives of particular populations and did not influence one another. Thus the statistical techniques could treat them alike.

Sample surveys do not have the luxury of treating all the units they observe alike. Surveys have to get the maximum precision for the least cost. As an example, let us suppose that the Labour Force Survey were carried out in a manner which would make it more amenable to the techniques of classical statistics. First, we would need every month, an up to date list of all people in Canada over the age of 15, and their addresses. Then we would choose a random sample (this means that all samples of the same size are equally likely) of (say)

100,000 of these people, for instance by drawing them one by one out of a hat. Then we would calculate unemployment rates in Canada, in each province, in each age group, etc. and it would be relatively easy to make comparisons and say whether unemployment in Quebec was really any different from unemployment in Ontario.

Unfortunately, this sample would suffer from some obvious defects. Very small provinces (such as Prince Edward Island) would not necessarily be well represented in the sample and we would probably have a hard time making comparisons of the unemployment rates of various age groups and sexes in PEI because there would be so little data. Also, the sample would be very costly since it would entail tracking people down all over Canada.

The actual method of conducting the Labour Force Survey, involves not a list of people, but of areas. Areas are sampled, at different rates for different provinces, these areas are segmented, segments are sampled within areas, and households sampled within segments. This is called a multistage sample. Within a household, all people over 15 are accounted for. By so doing, many operational problems are avoided and costs are reduced. We can ensure that the sample is distributed in such a way as to ensure reasonable estimates of relevant statistics in all provinces. However, it is much more difficult to make comparisons and to analyze the data because the design of the survey is so complex.

Two types of analysis which are common are contingency table analysis and regression. In contingency table analysis, we classify the population by the factors of interest and try to see from the proportion of the population in each cell, what the relationship between the factors might be. An example of a contingency table would be:

		EMPLOYED	UNEMPLOYED	TOTALS	
MALE	UNDER 25	100	20	120	460
	OVER 25	300	40	340	
FEMALE	UNDER 25	150	25	175	570
	OVER 25	350	45	395	
TOTALS		900	130		1030

Example of Contingency Table

This is a "3-way table". If the data were produced by random sampling, standard methods could be used for analyzing this table. If the entries were estimates from a complex sample design, we could not use the same techniques.

In regression analysis, we try to summarize by means of an equation how one variable is related to others. For example

$$IQ = a + b\,(\text{mother's IQ}) + c\,(\text{father's IQ}) + d\,(\text{family income}).$$

To estimate a,b,c,d and to test whether the model is any good, we need to get data from many individuals (and their parents) and then reach into the statistical bag of tricks to get our procedures. Again, if this is a psychology experiment, there is no particular problem; but if the data came from a complex survey, we have trouble.

At Statistics Canada, research into the extension of the classical methods to survey data from complex sample designs has been going on since the late 1970's. A number of methods have been developed. These methods are now being used world-wide for a variety of applications. This has contributed to the overall validity of the conclusions reached by researchers and analysts who were previously using the classical procedures which were not really applicable.

As an example, let us consider the results of an analytical study of the data from the February 1986 Adult Education Survey which was conducted as a supplement to the Labour Force Survey. The purpose of this analysis was to determine which factors contributed to whether or not a given female was taking full-time training courses. Factors included age, education, employment status and marital status.

In this case, the analytical method known as logistic regression was used. This was done in several ways; but we will consider only two cases: in the first, the design was entirely ignored and in the second, the design was completely accounted for.

The model in this case is the following: if P is the probability that a given individual is taking a training course, then P can be expressed in terms of her age, education, employment status and marital status as

$$\log \frac{P}{1-P} = a+b(\text{age group})+c(\text{education group})+d(\text{employment status})+e(\text{marital status}).$$

It is somewhat difficult to write out explicitly, but the analysis results in the various categories of age, education, etc. being associated with contributions to the probability that a female takes training courses. With each estimate of the contribution, an estimate of the sampling error of that estimate is also made. If some contributions appear to be missing, it is because other categories of the same factor are estimated relative to it, i.e. the contributions are relative to a reference category, which is set at zero. Thus for example the contributions of age are measured relative to the reference age group of 18-20 years.

Factor	Category	Estimates			
		Design accounted for		Design not accounted for	
		Estimate	Standard Error	Estimate	Standard Error
Age	18-20				
	21-24	.27	.14	.11	.10
	25-29	-.08	.18	-.26	.12
	30-34	-.25	.16	-.54	.13
	35-39	-.60	.19	-.62	.15
	40-45	-.62	.24	-.82	.16
Education	Primary	-1.43	.29	-1.38	.21
	Some Sec.	-1.22	.18	-1.16	.14
	Secondary	-1.08	.14	-.84	.12
	Some Postsec	.23	.13	.37	.12
	College	-.31	.13	-.10	.12
	University				
Employment Status	Employed				
	Unemployed	1.44	.12	1.38	.09
	Part Time	.46	.14	.32	.09
	Not in LF	.34	.11	.21	.09
Marital Status	Married	-.91	.14	-1.07	.12
	Single	-.30	.15	-.34	.13
	Sep/Div/W.				
Constant		-2.28	.23	-2.02	.19

Comparison of estimates when the design is/is not accounted for.
Females taking full-time training.

Example of Analysis of Data from Complex Surveys

In this example we can see some fairly large differences in the estimates, e.g. in the age categories 21-24, 25-29, 30-34. We cannot call these differences "statistically significant" because the standard errors are quite substantial too. In fact in this particular instance we learn that the design is to a large degree independent of the variables under study, which is good to know, but cannot be assumed and is certainly not always true.

5. SEASONAL ADJUSTMENT OF TIME SERIES

When statistics on economic or social characteristics, such as the Consumer Price Index or the unemployment rate, are collected regularly over a long period of time, we usually find a strong seasonal pattern appearing in the data. This seasonal pattern is associated with more or less predictable changes during the course of a year and gets in the way of studying the long term trends or cycles and making short term comparisons between periods.

People who concern themselves with analyzing time series usually regard the series as being made up of three components: the trend/cycle component, the seasonal component and the irregular component which is unpredictable and represents isolated events or random variation at each period.

To be able to make meaningful comparisons between periods, analysts want to remove the seasonal component. This process is called seasonal adjustment and the development and improvement of methods of seasonal adjustment has been and continues to be the subject of considerable research.

A widely used method of seasonal adjustment, called X-11, was developed at the US Census Bureau in the 1960's. This method was not without problems, notably it did a better job of adjusting in the middle of the series than at the ends. We know this, because when successive years are added to a series and the seasonal adjustment is repeated, the changes in the adjusted series are greatest at the oldest and newest values.

To solve this problem, a significant modification of the X-11 method was developed at Statistics Canada, called X-11-ARIMA. This method first extrapolates the series at both ends using a statistical model known as ARIMA (Autoregressive, Integrated Moving Average), and then applies the X-11 method.

As an example, Graph 1 shows, the number of unemployed men, aged 25 years and over monthly from January 1975 to December 1986 (the dashed line). This series is extrapolated foward by one year (the stars) by an ARIMA model and then seasonally adjusted (the continuous line) by X-11-ARIMA to give the trend combined with the irregular components.

Graph 2 shows the "final" seasonally adjusted values (obtained at the end of 1986) for the period June 1982 to June 1983 compared to the preliminary values originally obtained by X-11-ARIMA and X-11 in 1983 (when the next 3 years of data were not yet available). It is easy to see that the changes in going from preliminary to final estimates were far greater for X-11 than for X-11-ARIMA, indicating that the X-11-ARIMA adjustment is a far better predictor than the X-11 adjustment.

Graph 3 shows the differences between the final and preliminary series for X-11-ARIMA and X-11. Not only are the discrepancies for X-11-ARIMA smaller, but they are less systematic. The discrepancy in the X-11 adjustment would lead one to suspect some seasonal component might still be present in the X-11 seasonally adjusted series.

Software based on the X-11-ARIMA method has been developed at Statistics Canada. It performs four basic functions: forecasting; estimation of time series components, including trading-day variations and Easter effects; aggregation of original and seasonally adjusted data; the production of statistics relevant to the quality of the estimates.

X-11-ARIMA has been distributed to more than 5000 customers all over the world. Consulting services have been given to a wide range of institutions.

X-11-ARIMA was originally developed for main-frame computers, but is now available as a SAS procedure and in a micro-computer version.

6. STATISTICAL DATABASE DESIGN

6.1 Efficient Storage of Statistical Data

There are two significant differences in the requirements for the efficient storage of statistical data, in comparison with the storage of data that is to be used for operational or administrative purposes. Firstly, the data often becomes available in a relatively error-free form early in its lifetime, and is then subject to a very large number of queries on only a few variables for the purpose of analysis, without the need for it to be further modified. By contrast, a banking system would often require access to all the information about an individual account, when processing a transaction against that account.

There are significant benefits to be gained by organizing statistical data in a way that optimizes performance for usage that is typical of such data.

6.2 The Relational Model

Much of the theoretical work done on the design of databases during the 1970s concentrated on what were known as 'relational' databases. The main concept was that data should be organized as two-dimensional tables, where each contained a number of variables (known as column or domains), and a number of rows, representing different observations, or logical records. It was shown that more complex data organizations could be reduced to this form by means of an appropriate set of transformations, and that there were significant benefits to be obtained in terms of simplicity. In addition, such relational designs were more flexible, in that linkages were not built in to the data, but could be redefined to suit a particular set of query or update requirements.

At the same time, it became possible to define a set of standard operations on these tables, that could be used to manipulate the data they contained in a variety of ways. For example, the selection of a subset of the rows and columns of any relation could be accomplished by a standard software component, as could the logical joining of any two relations based on an arbitrary linkage key.

The language SQL (Structured Query Language) has emerged as an implementation of these concepts, and it is rapidly becoming accepted as a standard for the manipulation of data, in the same way that various programming languages have been accepted as standards for defining algorithms.

6.3 RAPID

In the early 1970s, Statistics Canada decided that a re-implementation of the systems that support the Census of Population was required. The decision to adopt database technology was made, and a review of the commercially available software was conducted. No package was at that time capable of processing files as large as were required for the Census, and an internal development program was undertaken. The result was RAPID, a general purpose database system that had the following principal characteristics:

(a) It supported files of relational design, with a very high limit on the number of rows and columns permitted in any one relation.

(b) Rows and columns could be added to or removed from any relation without the need to re-organize it.

(c) Data was stored by column. (This is known in the literature as 'transposed' organization.) It has the benefit that all the values of a particular subset of variables can be accessed without the need to transfer others into high-speed memory. This feature is especially suitable for statistical data, as explained above.

(d) A balance was achieved between very compact physical storage of data, and high performance in CPU terms.

RAPID was used successfully in the 1976, 1981 and 1986 Censuses of Population, and will be used in 1991. It has also been used in a number of other projects at Statistics Canada.

6.4 International Cooperation

A number of other organizations expressed an interest in RAPID, and it was supplied to over 30 installations. Many of these were other national statistical offices, some of which have used it extensively. Sweden, Hungary and the GDR are principal examples.

The software received the official support of the Statistical Computing Project (which is sponsored by the United Nations Development Programme/Economic Commission for Europe), which put together a collection of statistical software for use in developing countries. The member countries of the project have extended RAPID by providing a set of 'base operators', which closely resemble the SQL commands, for extracting data from a database that consists of a collection of RAPID relations.

7. GEOCARTOGRAPHICS

7.1 Georeferencing

Prior to the 1971 Census, there was no system which could support the retrieval of small area census data by user-specified zones. Users had to approximate their zones of interest by groups of enumeration areas or larger standard geostatistical areas such as census tracts or municipalities. Furthermore, these 'standard' zones often changed from one census to another. To address these problems, Statistics Canada developed the Geographically Referenced Data Storage and Retrieval System (GRDSR), which allows the census data to be stored at a very low geographic level, the block face, in large urban areas. With this system, users can receive data customized to their query zones, across several censuses.

This system and the street network file that supports it, the Area Master File (AMF), have received international attention and growing use within Canada, both in government and the private sector for planning, mapping, and vehicle dispatch applications.

7.2 Census Collection Mapping

One of the problems facing any agency conducting a national census of population and housing is to have a complete, easy-to-read, and up-to-date set of maps for use by enumerators dropping off questionnaires on census day. For the 1986 Census, a production system for computer-assisted collection mapping was developed to meet the needs of the census.

The system was quite successful: 1200 census tract maps were produced from the Geography AMF on schedule and on budget. The project resulted in a number of benefits to the department such as better consistency between the various geographic data bases and reduced production costs for new areas. For the 1991 Census, census tract maps will be produced by computer wherever the input from the AMF exists.

This project has been followed very closely by the U.S. Bureau of the Census. In fact, there have been regular technical exchange visits, and sharing of algorithms for automated census mapping.

7.3 Statistical Mapping

Statistics Canada has a wealth of data which often has significant regional variation. Statistical maps can be effective for communicating spatial trends and the distribution of characteristics across the country. Over the years, a range of cartographic techniques have been developed or acquired, installed, tested, and put into practical application.

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA
1010221860

c. 1

8. CONCLUSION

It should be clear from these examples that Statistics Canada has made significant international contributions to statistical methodology and computational techniques in the production of official statistics. It is essential for an agency such as Statistics Canada to maintain an active research and development programme in order to continue to meet new challenges and to develop state-of-the-art solutions to the problems it faces which it will share with the international community. There is no doubt that we will continue to make significant and interesting contributions to statistics and computing.

This presentation would not have been possible without the contributions of several people. Special thanks go to C. Hill, D. Binder, M. Jeays, E. Dagum, J. Yan and G. Sande for their co-operation in providing the material on which this talk is based.