The 1871 National Microdata Census Sample of Canada

Gordon Darroch and Michael Ornstein
York University
Toronto, Ontario

Overview: Canadian Historical Mobility Project and Class, Household and Mobility Project.

Gordon Darroch and Michael Ornstein, Department of Sociology, York University, June 1986.

The Context

The studies have been conducted in two phases though they are aspects of a continuing project. The first was focused on all four provinces of Canada in 1871 (Ontario, Quebec, New Brunswick and Nova Scotia). The second phase is focused on a large region of Central Ontario in the 1861-1871 decade (a wedge of counties stretching from the middle of Lake Erie to the lower shore of Lake Huron on the west, and on the east, from about one third of the way from Toronto to Kingston, at Port Hope, north to the southern tip of Georgian Bay). The studies are based on samples taken from the nominal data given on the census manuscripts of 1861 and 1871. Nominal data means simply the records of the individuals and households recorded on the original folios by the nineteenth-century census enumerators. At the time of writing they are available, in varying quality, on microfilm from 1851 to 1881 for Canada.

Both phases of the data collection have unique elements. The first phase created a representative national sample of households for 1871 that allows detailed analysis for a variety of variables. We have reported some results in historical journals (for example, Darroch and Ornstein, 1983; 1984). The Ontario phase has two unique elements. First, it is based on record linkage of very

large and unusual samples of individuals drawn from the census manuscripts of 1861 and 1871. Second, we created records for these individuals that nearly exhausted the information from all schedules of the censuses of those years, including household information, farm tenancy and productivity, real estate and the data of the manufacturing censuses (though the latter are very problematic). In some cases this required additional linkage procedures to attach information from more than one schedule to the same purported individual. This report provides an account of the methodology involved and of the nature and limitations of the data files.

The studies were undertaken in the context of two types of historical analysis that emerged in the 1960s and early 1970s. One was the breakthrough in demographic studies represented by the development and spread of family reconstitution using parish records in pre-census times. The other was the work on social mobility in past time, largely stimulated in North America by Stephan Thernstrom early study Poverty and Progress (1964). Though in many respects a limited work, Thernstrom first study nonetheless had a pervasive influence on the writing of social history after the mid-sixties (see, Social Science History, special issue, spring, 1986:1-44).

Two aspects of these analyses influenced the design of the current studies. First, they showed that a systematic social history could be built up from unconventional historical sources for the great majority of people who left no intentional traces or

records. Second, and more specifically, each was confronted by a serious problem of design by the facts of migration. The problem of migration was simply that there was a great deal more of it everywhere, in every era, than historians had usually imagined. The "discovery" of the volume of migration in the past deeply complicated the new historical methodologies, which were founded on the capacity to build limited biographies for historical individuals by linking nominal records with some fidelity. In other words, only the stable population for given geographic areas under study were "at risk" in this crucial linkage methodology; the surprisingly large numbers of movers simply escaped the analytic net.

In part, of course, the problem stems from the arbitrary nature of the civil or administrative units most often adopted as convenient sites for study, a small town, a parish or two, a city or possibly a county or département. These units bear only limited correspondence to meaningful social structural or individual spaces, in the past, as in the present.

The problems presented by migration and the limits of civil units to tracing individuals through historical records were probably exaggerated in early studies (Thernstrom, 1964). Still, recent historical studies of migration underscore the general difficulty, since they are based on rare historical sources, such as continuous population registers (Kertzer and Hogan, 1985; Hochstadt, 1986), on the unique U.S. Soundex indexes of surnames (Stephenson et al., 1978) or on formidably tedious procedures of

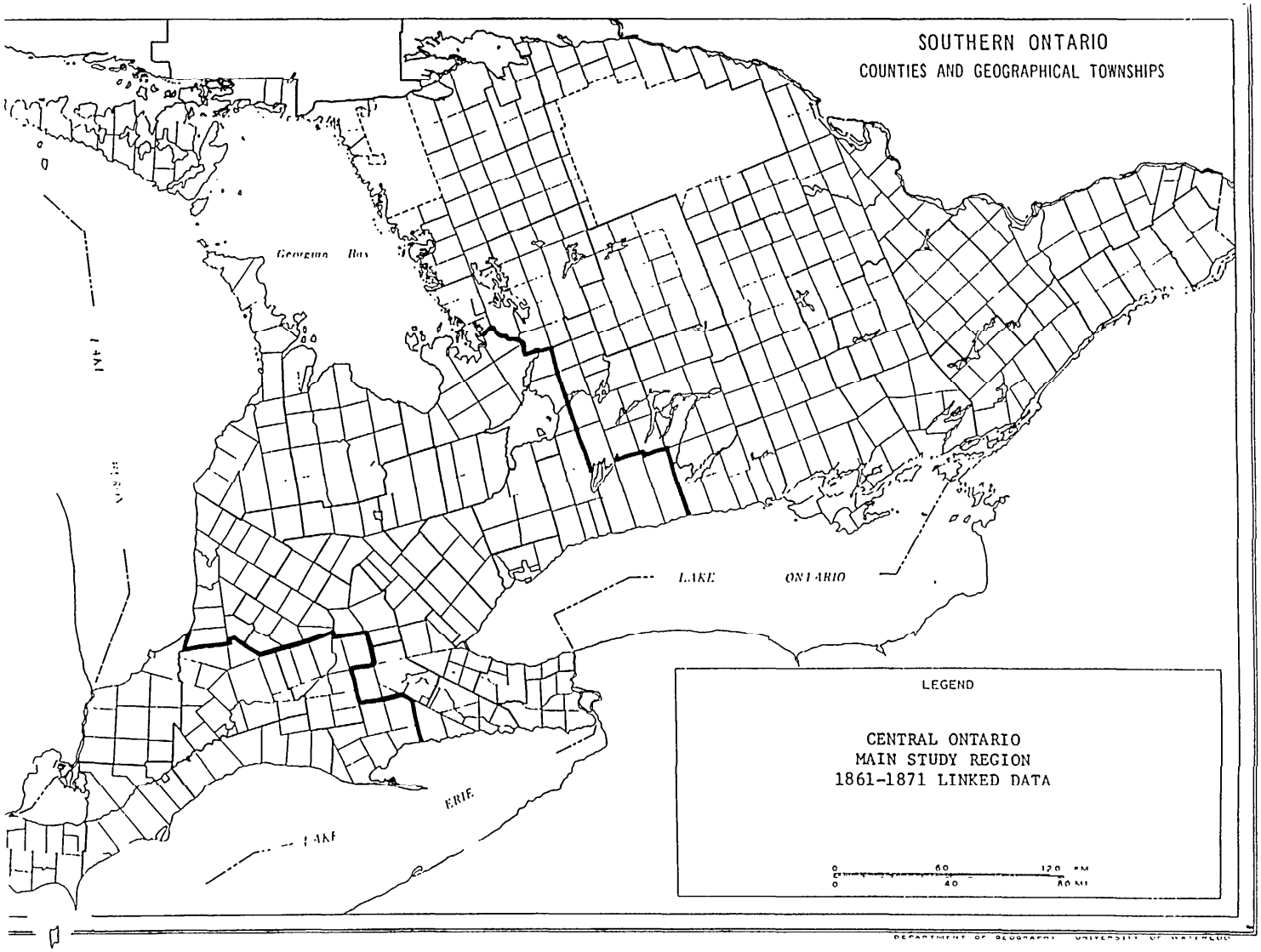
tracking individuals through innumerable discrete records (Knights, 1971).

The Sampling Method

Migration was only one among several, related concerns of our study centering on social class formation, mobility and the household economy after mid-Century in Canada. Some solution to the methodological dilemma was necessary in order not to vitiate other inquiries. Our solution was to combine a methodological sledgehammer with a methodological scalpel. The sledgehammer was simply to expand the area under study sufficiently to capture the large component of total migration made up of local and circular moves (despite the heavy flow of outmigration to the U.S. in nineteenth-century Canada). Initially, we envisaged a study of the population of all four provinces at Confederation. In principle such a study is possible, though only the first stage of the current study has such broad scope. The more intensive, second stage focuses on the large area of Ontario described above and outlined in Figure 1.

Figure 1 about here

The scalpel was a sample design and necessary complement to the sledgehammer. Clearly, some sampling strategy is required by an historical study that breaks out of the confines of local administrative units and attempts to construct a "collective biography" of thousands of ordinary individuals recorded in


SOUTHERN ONTARIO
COUNTIES AND GEOGRAPHICAL TOWNSHIPS
Georgian Bay
LAKE ONTARIO
ERIE
LEGEND
CENTRAL ONTARIO
MAIN STUDY REGION
1861-1871 LINKED DATA
0 60 120 KM
0 40 80 MI

administrative documents. A study of the population of even a relatively large city, like Toronto after mid-century, or of a major region or province clearly will generate an immense number of individual records. In the 1970s and 1980s such files quickly exhaust efficient data management capacities of even very large computers.

Moreover, our problem was not merely to design an efficient sampling strategy, but at the same time to preserve the capacity to conduct systematic record linkage. If one draws conventional samples from two or more historical listings, taking every nth person or otherwise randomly drawing cases, then one virtually eliminates the possibility of systematically linking records for the same individuals from the different listings. The random element of the samples, which ensures representativeness, also ensures that there is only the very slightest chance that any one individual will appear in both samples. One needs both to sample and to ensure that the samples are effectively closed populations, so that the same surviving individuals will appear in each.

Our solution was to devise a form of sampling we call letter sampling. It can be briefly described, although the actual procedure is rather more tedious (See other reports in this documentation). First, we needed to demonstrate that a random sample of surnames from a population of all surnames was, in fact, an adequate representative sample of the population itself. A sample of surnames, of course, preserves the possibility of record

linkage (at least for those who do not alter their surnames; unfortunately women tend to on marriage).

The procedure requires several stages. They can be briefly outlined. First, we wanted to examine in detail the character of surname samples for a nineteenth-century population, before proceeding to a random selection. Ideally one would compare the results of a large number of such samples with the characteristics of the population in question. Of course, a knowledge of those characteristics would obviate the need for sample estimates. Alternatively, there were the limited and flawed tabulations of the aggregate census reports. We chose a second alternative.

A large random sample of households could be drawn from the microfilmed copies of the nominal manuscript census of 1871, available through Public Archives of Canada and other libraries. The first or personal schedule of the manuscript census provides a variety of socio-demographic data for <u>all individuals</u> enumerated in their households of residence, including boarders, servants and visitors. The data collection itself was simple enough in principle, and relatively complicated in practice. The basic sample selections were proportional to township populations. Virtually all the reels of the microfilmed manuscript census of 1871 were scanned for the four provinces, township by township, and a prespecified number of sample households were identified in each (see sampling documentation). The information for all individuals in the sampled households was transcribed to paper and subsequently keypunched. The sample was a stratified, random sample with the

stratification ensuring that relatively small groups were adequately represented for purposes of analysis. Specifically, the stratified sample overrepresents the urban population in general, those of English-origin in Quebec and of French-Origin in Ontario and New Brunswick and the German-origin group in all areas in which they were at least 10 percent of the population (as determined from the aggregate census).

There were two primary objectives of this elaborate procedure. As noted above, it provided a surrogate national population from which letter samples could be drawn and to which they could be compared for a variety of characteristics and relationships. Second, it was clear that we had an unusual opportunity to supplement our methodological concerns with substantive ones: a relatively large stratified, random sample of a nineteenth-century national population provides for unique and very rich socio-historical analyses.

Analysis of the national sample has been reported to date in published articles on the complexity of nineteenth-century ethnic divisions of labour (Darroch and Ornstein, 1980), on the relationships between regional economies and household organization (Darroch and Ornstein, 1984) and on the complexity of households themselves (Darroch and Ornstein, 1983).

As for the original methodological objective, the results were consistently encouraging: the design effects of letter samples, which are technically cluster samples, were modest and the letter

samples adequately represented characteristics of the population from which they were drawn (see sample documentation).

The second stage of the study adopted a refined version of the letter sampling strategy. In this case the national sample was used to divide all surnames appearing in the census of 1871 into a set of 100 mutually exclusive clusters defined by Soundex phonetic codes, using the first letter of the surname and a phonetic classification of the next portion of the name. From these clusters a random sample of surname clusters or pockets was drawn, stratified by the size of the surname pockets.

This second phase of the study aimed to link individual records between two census years, 1861 and 1871. The magnitude of that task and the limits of funding and research time restricted the study to the large central region of Ontario, centered on Toronto. The area is represented in Figure 1.

By the time we were prepared to collect data again from the microfilm copies of the manuscripts, we were also able to substitute programmable data entry terminals for our paper and keypunch technology. In effect, however, the procedure was much the same. Coders were trained to search systematically the manuscript collection by city, town and township for the whole of the contiguous area of Ontario. All individuals with surnames that fell into the cluster sample were considered primary sampling units

and their complete census records, as well as the records of every other member of the same household, were transcribed exactly to computer file. The sampling and recording were repeated separately for 1871 and 1861 for each township or town in the region. A further selection had to be taken from the 1861 agricultural census, which had been taken as an independent enumeration.

The data collection for Ontario differed from the national sample in that all the information from the several schedules of the censuses was recorded for every member of the households (in 1861, the personal schedule is supplemented by information on manufacturing and industries and by the separate agricultural schedule; in 1871 there were nine full schedules, including agricultural, industrial and real estate censuses).

The last of the major steps in this research design was the linking of the individual records between 1861 and 1871 for the Ontario region. Record linkage has become a central feature of historical analyses using nominal data, from the early manual linkage of parish records in family reconstitution to quite elaborate computer algorithms for automatic linkage of large numbers of records.

After reviewing well-known computerized procedures we chose to develop a combination of computer and manual linkage that is particularly suited to these historical census records. In capsule form the procedure was as follows. The computer was enlisted to sort the records initially and to accomplish the merges of the individual data files after linkage. The sorting was no mean task:

there were over 34,000 individual records selected in the letter sampling for central Ontario in 1861 and over 40,000 in 1871. Using alphabetically sorted surname lists, the linkage proper combined a complex set of decision rules regarding records that would be considered to refer to the same historical person, with the pattern recognition capabilities of research assistants. The rules emerged out of a reading of the linkage literature and from trials undertaken by the principal investigators. They were quite complex, with separate decision algorithms applying to cases where information was limited to individuals and to cases where the family and household context added information. The algorithms were conditional ones in which the requirement of a precise or close match on one item of information, for example, on name spellings, age, or birthplaces depended on the precision of the match on others, say, on a wife's or parent's first name, age and ethnicity or on the names and ages of children. Uncertainty was systematically reduced by the accumulation of information across several items.

The research assistants learned their "trade" largely by trial and error under supervision. Results of initial trials showed quite high rates of replication for different individuals.

In all some 16,000 records were considered true links. In every case, the links were coded to include a subjective estimate of the level of certainty. First estimates for the entire region put the rate of linkage at about 55 percent of those at risk in 1861, taking account of mortality and, for women, marriage and name

change. The large residual represents an unknown combination of outmigration from the region, census underenumeration and record linkage failure. Both emigration to the U.S. and short-distance migration are known to have be very high during the period; they are probably the major component of the residual, but other evidence indicates that the limits of the nineteenth-century censuses and of the method are substantial. Previous studies suggest we might set the limit of census underenumeration at 10 to 12 percent in any census year, with the highest rate for 1861 (Knights, 1971; Stephenson, et al. 1978). Considering the likelihood of significant overlap in those subject to underenumeration, a combined rate for both years might be 15 to 18 percent. Estimating the errors and omissions of linkage adds another approximately 10 percent (see the differences between rates for different methods of tracing migrants in Katz, Doucet and Stern, 1982:ch. 3). For both years, then, the combined rate of linkage failure could be as high as 25 to 28 percent of the total, leaving 17 to 20 percent of the loss to migration.

REFERENCES

Darroch, G. and M. Ornstein. "Family Coresidence in Canada in 1871: Family Life-cycles, Occupation and Networks of Mutual Aid". Canadian Historical Association, Historical Papers, 1983:30-55.

_____. "Family & Household in Nineteenth Century Canada: Regional Patterns and Regional Economies". Journal of Family History, (Summer, 1984):158-177.

_____. "Ethnicity and Occupational Structure in Canada in 1871: The Vertical Mosaic in Historical Perpsective." Canadian Historical Review, (September, 1980): 305-333.

Hochstadt, S. "Urban Migration in Imperial Germany". Paper presented to the Canadian Historical Association Winnipeg, June 9, 1986.

Katz, M.; M. Doucet and M. Stern. The Social Organization of Early Industrial Capitalism. Boston: 1985.

Kertzer, D. and D. Hogan. "On the Move: Migration in an Italian Community, 1865-1921". Social Science History, (Winter, 1985):1-24.

Knights, P.R. The Plain People of Boston 1830-1860: A Study of City Growth. New York: 1971.

Social Science History, Special Issue, Spring, 1986.

Stephenson, C. et al. Social Predictors of American Mobility: A Census Capture-Recapture Study of New York & Wisconsin, 1875-1905. Newberry Library, Chicago: 1978.

Thernstrom, S. Poverty and Progress: Social Mobility in Nineteenth Century City. New York: 1969.

Coding and Data Processing for the Feasibility Study: Canadian Historical Mobility Project.

see Archive file for whole project description

There is ARPE only

1999

by: Gordc
Depar
York
anc
Micha
Insti... for Behavioural
Research and
Department of Sociology
York University
August, 1977.

Revised, 1980, 1984, 1994, 1999.

Appendix E: Part 1

Feasibility Study: Coding and Data Processing

Introduction

Our project was faced from the start with the need to create very large machine readable files of data transcribed from the microfilms of the Canadian censuses of 1861, and 1871. An examination of published work on nineteenth-century census-type data provides some, but not a great deal of guidance as to how to proceed. Only a very few projects, notably the Philadelphia Social History Project, have had experience with data files of the magnitude of those we propose to collect or, for that matter, were involved in the feasibility project. Acquiring experience with large historical data files was one of the reasons we designed the feasibility study to entail the collection and management of data which was much more extensive and diverse than that required only to test the proposed 'letter sampling" strategy.

Previous research did make it clear, however, that two particularly serious analytic difficulties had to be avoided. The first arises when early decisions about coding procedures result in some variables of interest simply being omitted. The second arises when a variable is created with a smaller number of categories than turn out, in later analysis, as necessary to capture fully the historically significant variation in the variable.

Both problems initially may seem obvious ones to avoid. Yet those who have coded large amounts of data and especially historical data will

recognize the considerable temptation to simplify coding procedures by ignoring some seemingly unimportant information, say the size of dwellings as recorded on censuses, or by collapsing categories of a variable, with- a great many legitimate categories, such as religion or occupation. There were enormous numbers of distinctly named protestant churches and sects and of distinct occupations in nineteenth-century North America. For a coding task of even moderate size, the additional cost of returning to the original source of the data to rectify omissions or errors is usually prohibitive. The coded data thus come to impose unnecessary constraints on the analysis itself.

In the light of these considerations we adopted the following principle in all phases of the data processing on this project: in the original coding of manuscript data (microfilm images) all the variables describing an individual are coded. Each variable is also to be transcribed exactly as it appears on the original document or coded in such a manner that exact original values are recoverable. Finally, the subsequent data processing of the records must always assign a unique value to each unique category of every variable.

It should be noted that this detailed and complete method of coding facilitates a full exploratory analysis of the data using all possible variables. In addition, it maximizes the value of the data to other researchers who may employ it for any secondary analysis which the original documents themselves permitted. Even if we had begun with a focused analytic purpose which, for example, did not require any information regarding religious affiliation or detailed occupations, the decision not to code these variables fully would obviously place severe limits on any future secondary

analysis of the data file while entailing only a relatively minor initial cost saving.

A second major consideration in the collection and processing of nominal manuscript data involves the units of analysis. The data ahould be in a form which makes it possible to use each of the following as units of analysis:

a. the individuals listed on the manuscript source, for example, to permit examining relationships such as that between a person's religion and his or her occupation;

b. the complete households, for example, to permit examining relationships such as that between the religion of the head of the household and the size of the household; and

c. the individuals listed, but in "contextual analyses" in which the context is given by the characteristics of the households as a whole or as given by the characteristics of other individuals in the household. For example, in the first case, it should be possible to examine school attendance of children as a function of the size of household in which they live; in the second case, to examine school attendance of children in relation to the occupation of their fathers.

In order to make this possible, households must be coded in their entirety. The important but perhaps not immediately obvious implication is that any sample from manuscript censuses must be a sample of households, not only of individuals. The formation of household composite variables, of course, again poses the requirement that <u>all the data on each individual</u> in every sampled household be coded.

The full range of contextual variables which could be of interest in

analyzing these data will not be apparent until the analysis is underway. For example, an examination of school attendance might lead one to relate this variable to the occupation of an individual's eldest brother--but it is hard to anticipate this beforehand. Other variables, like "household size" have been frequently employed in the published research and it makes sense to create them at the start. In this study, a set of variables describing each household is attached to each individual in the household. In addition, a set of "pointers" is created which will allow new contextual variables to be created easily when they are needed. For example, one of these pointers is equal to the person number (i.e., position in the sequence of individuals in the household) of each individual's father. Part 2 contains a list of all these variables describing the household structure.

Coding Procedures

Considerations of cost led us to employ a coding procedure for the nominal census data which involves two steps common to coding in projects such as this one. First, the census microfilm data are transcribed onto coding forms designed for the purpose and, secondly, the data is keypunched from the coding forms. It is possible to combine the two steps into one, by coding the census directly onto a computer terminal, and using remote data entry or by keypunching directly from manuscript records to computer cards. Remote data entry would allow for immediate error correction, a considerable advantage. Direct keypunching, of course, saves the time, cost and possible error involved in a two-step procedure.

At the time our coding was carried out, the facilities available through the York University Computing Centre all but precluded our using

direct data entry. Direct keypunching would require coders with keypunching skills. However it is likely that we would have chosen the two-step procedure in any case at this pilot study stage. The procedure is suited to gaining firsthand experience with the difficulties faced in reading and accurately transcribing microfilm records. The Canadian census data, available in 1975 at least, presented the problems of near illegibility of some of the original manuscript records and of relatively poor quality of the microfilming itself.

Of course, the two-step procedure requires that the keypunched data be checked and verified for illegal codes, inconsistencies and the like. In our case we opted to both verify the keypunching and to scrutinize the data for coding errors by using a computer programme in batch mode. This means that once errors are detected or suspected, we were required to return to the microfilm to locate the incorrectly coded record. The correction of the files has proved to be a time-consuming and tedious process which the principal investigators undertook almost entirely themselves. Two advantages to this verifying-checking procedure became evident. First, in a project the size of the feasibility study, much less the size of the proposed study, principal investigators simply cannot undertake much of the original coding, though close supervision is essential. We have found, however, that familiarity with all the problems of coding, and their inevitable implic4tions for data analysis, has been assured by primarily undertaking the verification and error checking ourselves. As we shall shortly describe the computer programme written for this purpose requires that every conceivable error and ambiguity which can be detected in the punched card file is examined and corrected. In effect the procedure has required us to review in

detail the entire machine readable file. Of course only some kinds of coding errors can be detected in this way. Hence, we also undertook as part of -he data collection the special "error detection" sample--amounting to a complete replication of the coding from the microfilms of a ten percent sample of the originally sampled households.

The second advantage of the procedure adopted has simply been our assurance that we have a very clean and consistent historical data file. In fact, as described below, the computer programme employed in checking the data has given us a unique record of the original errors and ambiguities in the file and the kinds of corrections made.

The Coding Forms & Instructions

Our coding forms have been designed to closely resemble original printed forms used for the nineteenth-century censuses (see figures 1, 2, and 3). The two coding forms included below in the text were the only forms employed for the collection of personal and household information. Note: Figures 2 and 3 refer to a pilot study for two Ontario counties, Essex and Kent. These data have not been made available as an electronic file. Only card files were created.

The variables are transcribed in the precise order in which they appear in the original and where it is necessary to code something less than the complete original entry, mnemonic codes with letters are used. All the data on an individual are coded in a total of eighty characters, for ease at the keypunching stage. Our main objective, especially in using mnemonic codes, was to minimize error at the coding stage, because of the high cost of

Figure 1: Division 1

HOUSEHOLD INFORMATION-1871 CENSUS OF CANADA

District	Subdistrict	E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write
1 3	4	5	6 9	10 12	13	14 15	16 31	32 47	48	49 51	52 53	54 55	56 57	58 73	74	75	76	77

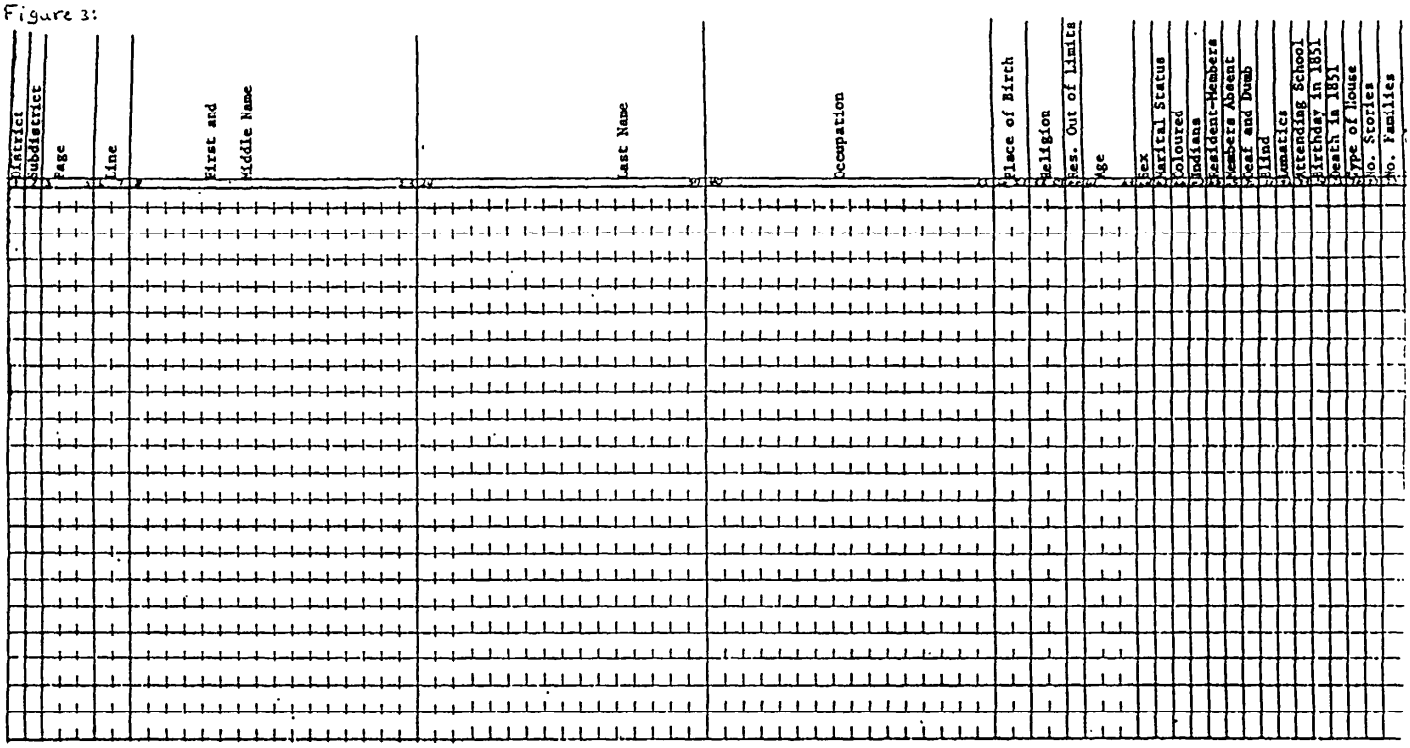
CODER'S INITIALS _____ DATE _ _/_ _/_ _ PAGE NUMBER _ _

Figure 2:

MOBILITY STUDY 1861 ESSEX-KENT BASIC HOUSEHOLD INFORMATION

Location	Page 1	Person	First Name	Last Name	Profession-Trade-Occupation	Birth Place	Married 1860	Religion	Age	Sex	Married 1st	Coloured	Pau. Present	Deaf-Dumb	Blind	Lunatic-Idiot	School	Cant Rend/Wr	Birth 1860	Death 1860	House Type	Stories	# Families	Being Built
1 - 3	4 - 6	7	8	24	25	41	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73

Figure 3:



PERSONAL CENSUS FOR 1851

CANADIAN HISTORICAL MOBILITY PROJECT: ESSEX-KENT STUDY

correcting errors. The pursuit of this objective is not without its costs, for the data can then only be analyzed after considerable transformation. Primarily we must substitute a unique numeric (rather than alphabetic) code to represent each possible value of each variable for the purposes of analysis. The transformation of alphabetic into numeric codes is accomplished by a computer programme also designed specifically for this project. The alternative to our procedure is the conventional one of recording the original data as numeric codes when it is first read from the microfilm. But the conventional procedure is both slower and more prone to error than our mnemonic coding method. Consider, for example, that if two digits are used to represent a religion variable, approximately sixty of the one hundred possible two digit combinations would be required. The numbers 43, 67, and 10 might represent Wesleyan Methodists, Adventists, and Roman Catholics respectively--instead of our codes WM, AD, and RC.

The mnemonic codes are superior in two respects: they are much easier for coders to recall and so they should increase efficiency and result in fewer coding errors; and, if an error is made, the resulting error is more likely to be detected in the data checking procedure. It may be noted that there are 26 x 26 or 676 valid two character mnemonic codes, of which sixty are required. Thus in most cases a coding error will take the form of an invalid and hence correctable code. If by mistake an Adventist is coded AT, rather than AD, the error is simply more likely to be spotted than an error in numeric codes.

Two sets of two character mnemonic codes were developed, one for religions and one for place of birth. The place of birth codes were also used to code the 'nation of origin' variable in the 1871 census, so for example, the code for England (a place of birth) was also used for English (reported as a nation of origin). The major difficulty which arose from the

adoption of this procedure was that certain religions and places of birth occurred so infrequently that it made no sense to create codes for them, yet we were committed to preserving the exact content of the original manuscripts. The solution was to use a special code, a blank followed by an asterisk, when-ever a mention of a religion, place of birth or nation of origin occurred which had not previously been assigned a mnemonic code by us. At the same time an additional coding form, called the "long form," was filled out which was key-punched and its contents merged by computer with the individual record file (see figures 1a, 1b and 1c).

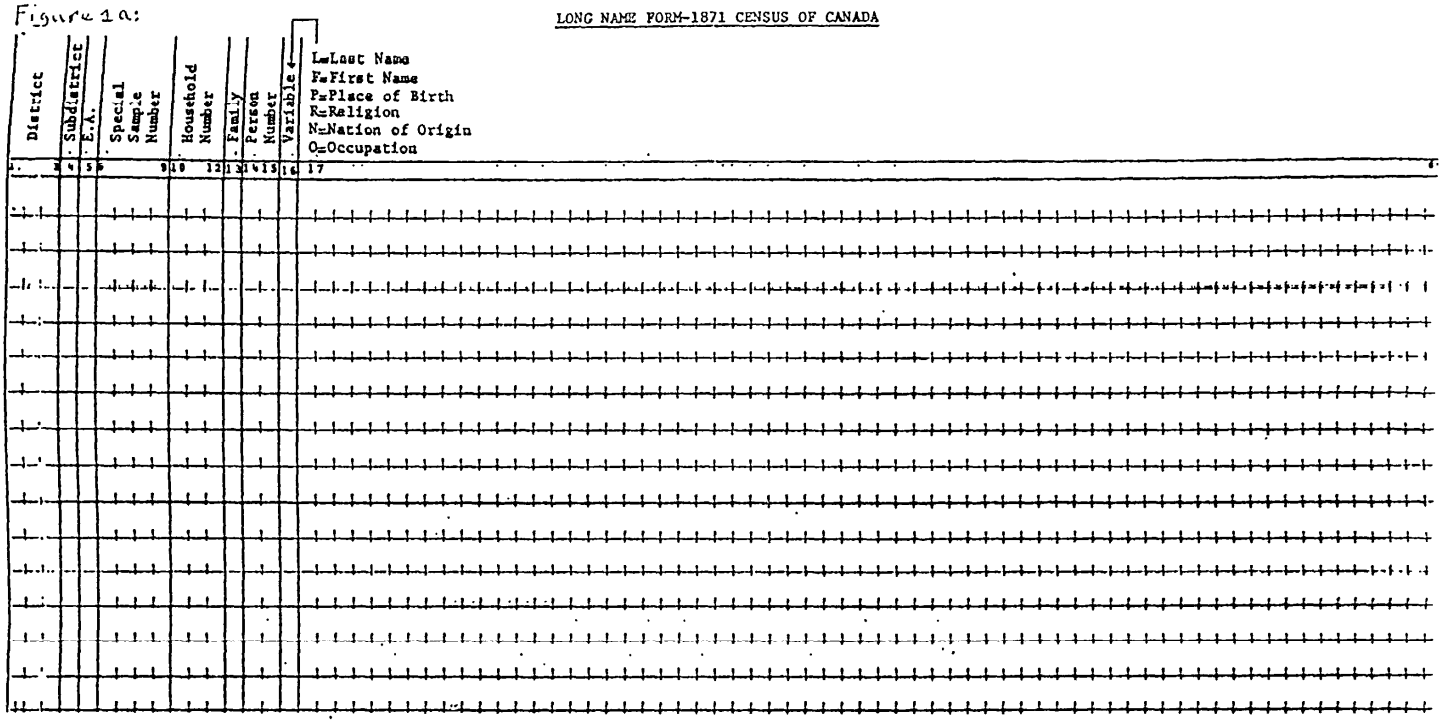
The 1871 long form has three fields: a location code which identifies the individual to whom the data referred, a code specifying the variable in question ('R' for religion, 'B' for place of birth, 'N' for nation of origin), and the exact mention or name for the variable as it is written on the census manuscript. For example, "West Indies" was a very uncommon place of birth requiring a "blank, *" code and a long name form.

The first and last name of each individual and his or her occupation were transcribed directly onto the coding form in the form in which they appeared on the manuscript, for no predetermined coding scheme could preserve the original content in full. Because of the fixed field coding scheme, a procedure was required to deal with cases where the name or occupation exceeded the length of the field (sixteen characters for each name and for the occupation) on the coding form. Here again, the "long form" was employed--an asterisk was placed in the last space allocated to the variable in question on the individual coding form and a long form was filled out containing only the uncoded part of the person's name or occupation.

The 1861 long form used the same three fields in altered format but using the same variable codes as in 1871. The form was also used for a secondary purpose. In 1861, as in 1851, the personal census schedule

Figure 1a:

LONG NAME FORM-1871 CENSUS OF CANADA



CODER'S INITIALS _____ DATE __/__/__ PAGE NUMBER __

Figure 2b: MOBILITY STUDY 1861: ESSEX - KENT LONG NAME FORM AND HOUSEHOLD CRAFT AND PRODUCTION INFORMATION

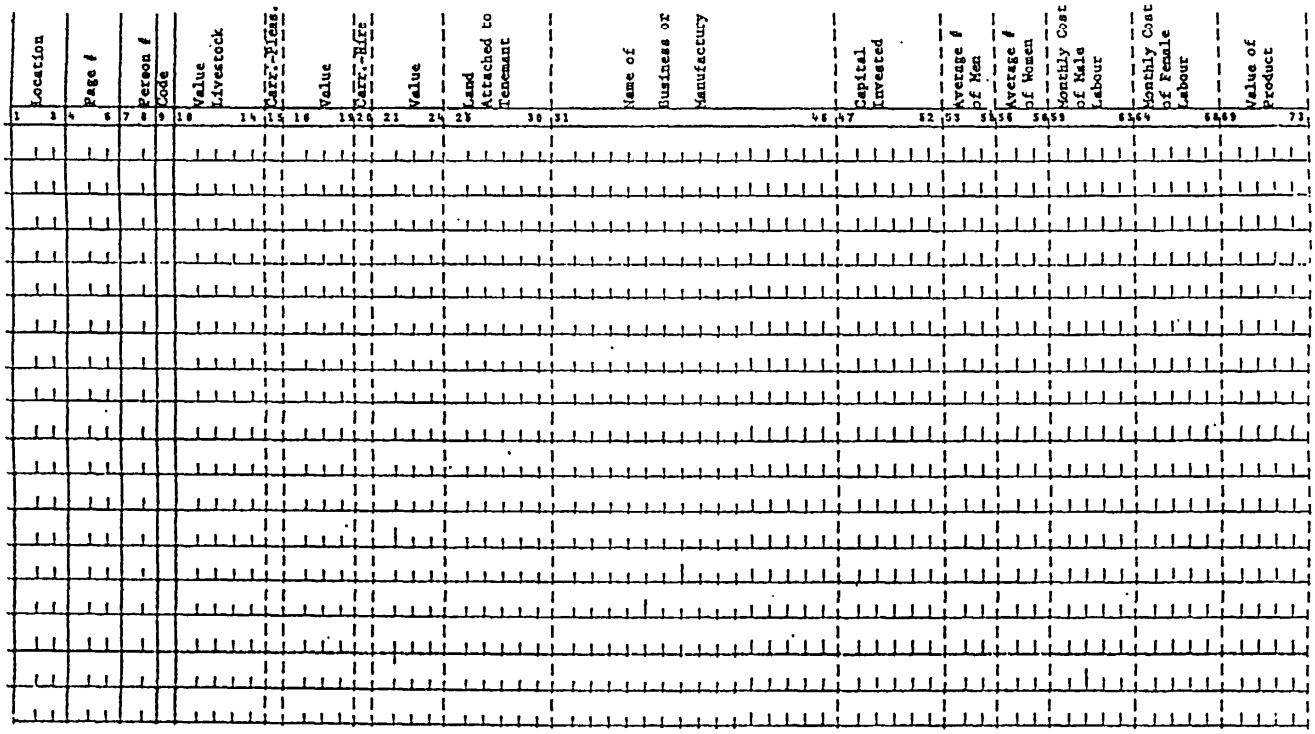
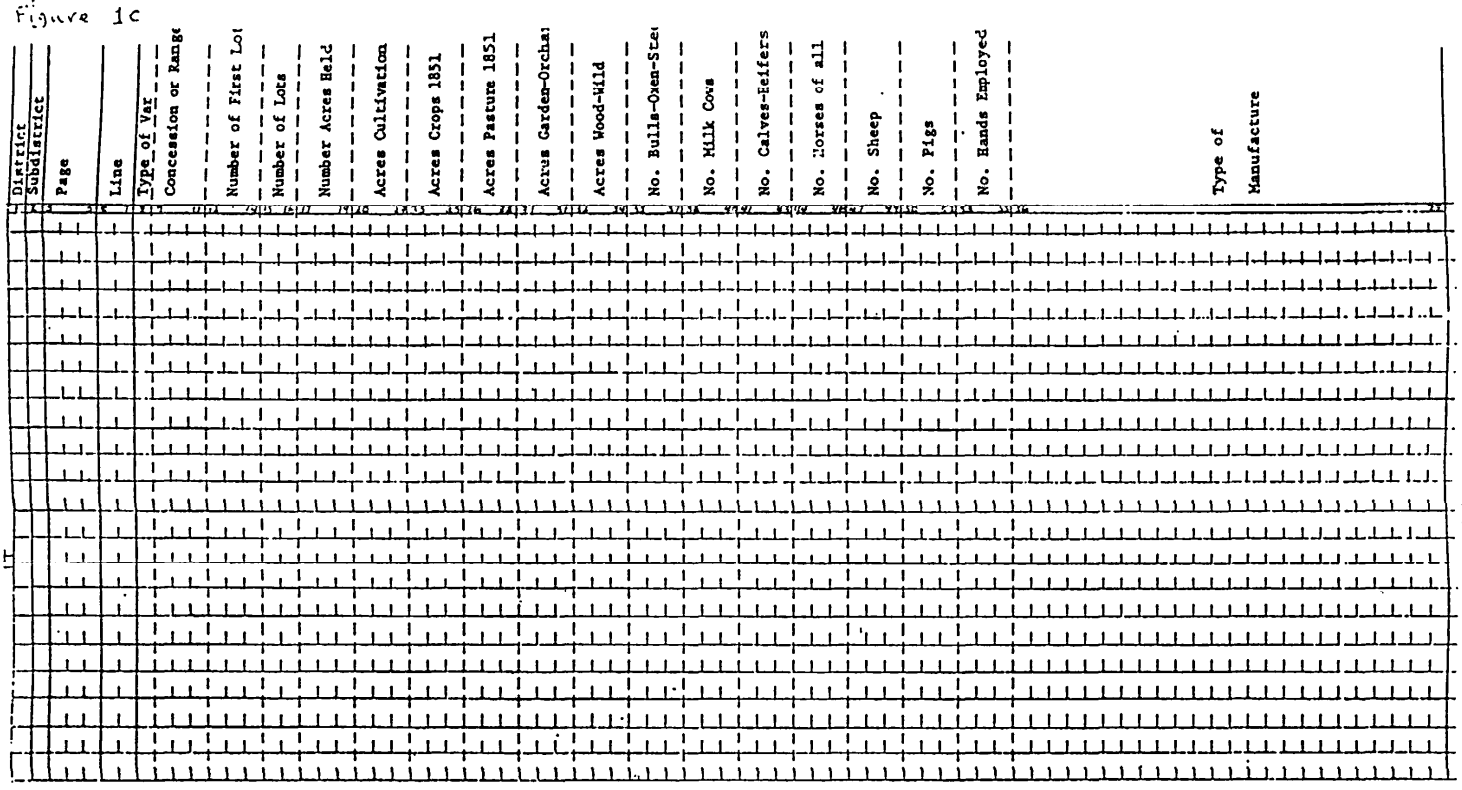


Figure 1c



AGRICULTURAL AND MANUFACTURING CENSUS 1851

CANADIAN HISTORICAL MOBILITY PROJECT: ESSEX-KENT STUDY

included information on household production and business and manufacturing information for those members of households who were owners and operators. Since this was occasional information which could be taken from the microfilm reels we provided for its direct coding. At the time of coding, comparable information for 1871 had not been microfilmed and provided on the same reels as the personal and household data.

Four of the variables which were coded, the surname, religion, place of birth, and nation of origin, were quite often identical for each person in a sequence of individuals listed within a household. So that the coders would not be required to tediously copy out these variables for each person in a sequence, blank fields in the keypunched data were automatically given the same values as the corresponding field for the previously coded person. Of course, the first person in each household must have all his or her variables coded.

Coding

The coding of the data for the sample of households from the 1871 census manuscripts was done through the facilities of the Institute of Behavioural Research. Four Bell and Howell microfilm machines were rented for a period of six months. Of several machines examined, these provided the best reproductions of the microfilms.

Under contract, but in continuous contact with the principal investigators, coders were selected and trained. In the end twelve coders were employed, working four at a time for approximately four hour periods. They were supervised by a full-time, experienced member of the staff of the Institute. The coding and keypunching of the entire file of households consisting of individual records required eight months and cost $12,426.

Thus, the average cost per household was about $1.24 or 20 cents per individual record.

In addition to the main file, we collected data for a trial "letter sample" of households for Essex and Kent counties, Ontario for 1851, 1861 and 1871. This coding was subsidized for the most part by York University in providing funds for us to employ, part-time, six graduate student research assistants. Mainly they were employed in coding the data from microfilm. Two assisted in correcting the files. Two students are currently using the data procedures and computer programmes of the project for their own research. Pages 17 to 34, which follow, are copies of the coding instructions for the national sample from the 1871 census and for the coding of the special letter sample of the 1861 census for Essex and Kent counties. For coding the 1871 letter sample for Essex and Kent counties, we employed the coding forms and slightly revised instructions used for the national sample for that year. Coding of the 1851 census for the letter sample for Essex and Kent was not initially planned, but we have been able to complete part of this work to date. The instructions and coding forms used in 1851 are only slightly altered versions of those employed for 1861.

CANADIAN HISTORICAL MOBILITY PROJECT
CODING INSTRUCTIONS: PILOT PROJECT: NATIOML STUDY
NOMINAL CENSUS RETURNS ON MICROFILM, 1871

The objective of this coding is to transcribe selected cases of households from the 1871 census returns, which are on microfilm, to coding forms for key-punching. You will be coding information on all individuals in all families in selected households.

The only information to be coded is that *for* the households given to you on a "list of sample households." This list provides you with the information to locate the households in the 1871 microfilm.

The appropriate microfilm reels for the District (County> you are coding, will be given to you. The "list of sample households" gives district and subdistrict numbers, (a no. and a letter, 1A, 1B, 2A, 2B . . . 200A, 200B...) This refers to districts and subdistricts listed on the top and right-hand side of each microfilm frames. You should locate the appropriate microfilm frame by this number and check that it is correct by looking at the district names.

The "list of selected households" provides the following additional information for locating the households to be coded.

E.A. Numbers: These refer to the divisions within subdistricts, they are numbered on the microfilm at the top right-hand side (and called Divisions" there). Sometimes the selected households are all in the same divisions (E.A.), sometimes, in different ones.

NOTE: It is essential that you locate the household in the appropriate division (E.A.). This number will be coded, see coding instructions for column 5 below.

Special sample numbers (spec. samp. #), to be coded, see columns 6-9 below.

-Household numbers (H.Hold. =). These are the selected cases within the Division (E.A.'s) for which all information given on the microfilm for all individuals in the household is to be transcribed on the coding sheets. The household number will be coded, see columns 10-12 below.

- The "list of sample households" provides an additional instruction as to which households should be coded. Immediately after the household number several instructions are listed (-1) "code any case" means simply to code the household indicated whatever ethnicity - nationality (not birthplace) it may be.* (2) "French only" means you only code the case if the "Nation of origin" (called ethnic origin on microfilm) of the head of household (first person listed) is French. (3) "German only," and (4> "Non-French only" are the other instructions which will appear - in which case you only code the household to which the selected household number refers if the head of household is German ethnic origin.

The "Non-French, German, French only" instructions refer to ethnic origin or nation of origin (these refer to the same column on the microfilm), they do not refer to "Birthplace" which appears first, i.e., two columns to the left of "origins."

GENERAL INSTRUCTIONS

1. Always code with a sharpened pencil, use an eraser.

2. Exercise caution in coding: accuracy counts more than speed.

PLACEMENT OF INDIVIDUALS ON CODING SHEETS:

- Code one person per line, skipping no lines.

- Try to avoid splitting households on more than one sheet by seeing that the number of lines left is greater than or equal to the number of individuals in the new household to be coded.

- For large households which must continue on more than one sheet, or where you have miscalculated and not left enough space, code the family, household, and the person numbers at the top of the next sheet.

- Leave lines blank which will not contain a full household on a sheet, and go to the next sheet.

- You will probably average two households per code sheet.

3. If parts of a name (last, first, initials or occupation) are illegible, and quite indecipherable, then

A. code exactly those letters which can be made out -

B. place dashes (--) in appropriate columns for the illegible letters - one dash for each letter - if the whole name is illegible, then leave the whole space blank on the code form.

C. also place a question mark (?) in the last column of the entry - e.g. for last names, in col. 31; first names, col. 47; birthplace, col. 53; religion, col. 55; nation of origin, col. 57; occupation, col. 73.

NOTES: If name or occupation is partly illegible and also too long for the columns on the code sheet put a (+) sign in the last col. of the entry instead of the question mark - then complete the name on the "long name code sheets" (also see coding instructions below for the specific items).

It is very important that the first letter of the last name is correctly coded. If this letter is illegible then an entirely new household is to be substituted and coded. The substitute household is given on the list' of sample households directly below the normal listing within each section. You code the first household listed if you have to

substitute. There are a variable number of substitute cases depending on the number of cases required.

Substitution will only be occasional; it probably means you must erase the original codes for location, household and special sample no. and substitute the new codes - then complete the substituted case as it is found on microfilm.

If for the substitute case the first letter of the last name is again illegible, another substitution will be necessary. The next case listed in order in the substitute list must be taken.

If all substitute cases have an illegible first letter, then check with your supervisor.

When you have substituted, put beside the one taken from the list of substitute households.

More than one substitution is made only if the original case was being coded under the instruction "code any case." If the original case was "French only," "Non-French only," or "German only," then the substitute case must also be one of these. If the original, illegible case did qualify under these 3 conditions but the substitute does not, then do not code a household, go to the next section of the "list of selected households" and proceed as usual.

You may have to search the microfilm to locate a substitute, but it is always located near by the location of the original - perhaps in another division (E.A.), but always in the same general area.

4. All information coded should be PRINTED IN BLOCK CAPITALS. The letter "I" should have this form "I"; the number one should be written "I." The letter "L" is coded "L," not "1."

5. It is easy to miss the codes for married or widowed, school (attendance), able to read, write, etc., (i.e. col. 74-80), since this information is located at the extreme right-hand side of the microfilm frame.

6. Make certain that you are copying the precise spelling of the household nos. and names and occupations. They are given on the microfilm. Read them carefully first, and copy them letter by letter - the original enumerators made errors, copy the errors; spellings in the nineteenth century differed in many respects from current spellings; copy the nineteenth-century version.

7. Occasionally first names are listed in the last name column and vica versa. Read names before copying. If it is obvious that the error has been made code the names in the correct cols. If you are unsure, copy microfilm directly.

8. All information must be coded precisely in the columns marked on the code sheet. Last names, first names, and occupations are coded by beginning with the first column from the left, (col. 16, col. 32, col. 58). Other codes which have more than one col. (location codes, household no., person no., age) are right justified and leading zeros are to be filled

in, e.g. age, 1-9 is coded 001-009, 10-99 is coded 010-099 and 100- is coded 100-XXX is col. 49-51.

Birthplace, religion and nation of origin are always two (2) letter codes.

9. REPETITION OF CODES

a. For the first member of the household, all variables must be filled in regardless if they are the same as the values for the last member of the previous household.

b. For individuals within a household some codes may be left blank - they will be automatically duplicated in punching. The only variables for which blanks can be left are, last names (cols. 16-31); birthplace (col.~32-&3); religion (col. 54-55); nation of origin (ethnic origin), (col. 56-57).

The code must be given for the first individual listed on microfilm with a particular last name, birthplace, religion or ethnic origin. For individuals immediately following in the list who have the same code, the code should be left blank.

e.g. If the head of the household is born in Scotland, the wife in Ireland and all the children in Ontario, then for birthplace we must code the first three birthplaces - the second and later children will automatically be given the birthplace Ontario, if it is not coded.

10. Try to make out letters that initially appear illegible. In trying to make out the writing on the microfilms, try to familiarize yourself with the handwriting of the enumerator by looking around the microfilm frame on which a selected case is located. You may be able to recognize the differences between, say, M's and W's, S's and T's, etc. Each new numerator, of course, had a different script. But, if you are not certain, follow the question mark (?) codes for illegible cases as indicated on these instructions.

SPECIFIC CODING INSTRUCTIONS

Columns 1-12: are coded directly from the "list of selected households.'

Columns 1-3: District number, starting with district No. 1 and numbered consecutively. Code is right-justified in cols.1-3. This number is given as the numerical digits on the "list of sample households." Labelled, SUBDIS = 1A, 1B, 2A, 2B, etc.

Column 4: Subdistrict code within Districts. Given as the alphabetic code on the "list of sample households." Labelled, SUBDIS = 1A, 1B, 2A, 2B, etc.

Column 5: E.A. or DIVISION number within subdistricts. These numbers are also given on "list of sample households" as E.A. = X for each household selected. They range from

1 to 9 only.

Columns 6-9: Special Sample Number, given on "list of sample households as SPEC SAMP # = , one for each household listed. Code is right-justified in cols. 6-9.

Columns 10-12: Household number, given on "list of selected households," as HHOLD = . Also right-justified in cols.10-12.

The following information is coded from the microfilm:

Column 13: Family number: some households have more than one resident family. On the 1871 microfilms separate nos. are given for families (column 6 of the microfilm). Do NOT code these. Code the families consecutively within each household, beginning with 1, 2, 3 . . . Leave blank if only one family is listed in a household.

Columns 14-15: Person number: The coder assigns each individual within each family a separate number, consecutively 01, 02, 03 . .

NOTE: some households will in fact be boarding houses or hotels. One or more families may be listed as well as any number of unattached individuals.

Code a family no. (col. 13) for each one given on microfilm listing.

Columns 16-31: Last or family name; from microfilm - (manuscripts col. 7), begin at left hand, col. 16. Fill in coding form until next to last column - if name will exceed space (16 spaces max.), then fill last column (cal. 31) with an "asterisk" (A) and go to the long name coding form (#2). Provide the complete name there. Blanks are left for all persons with same last name after the first is coded.

NOTE: Make certain that all the different last names are filled, e.g. the last name is coded for all non-family members of the household (usually listed last) See no. 3 above for instructions regarding illegible letters.

Columns 32-47: First names; initials: Begin at left hand, col. 32, code as given (col. 7 on microfilm)

Leave one blank column between names and initials.

As above, if names and initials exceed space (17 spaces) fill in last column (col. 47) with asterisk (*) and code the complete name on long name form (#2).

See no. 3, above, for instructions regarding illegible letters.

Column 48: Sex: code M or F from microfilm.

If sex is missing put a question mark (?).

Columns 49-51: Age as given on microfilm. Code is right-justified.

NOTES: All single names use the first two letters of the birthplace as a code: Those consisting of two or three words use the first letters of the first two names. There are a few exceptions - they have asterisks beside them in the following code listing.

For a few birthplaces listed on the microfilm, there will not be a code here. Then place an asterisk in col. 53 and code them on the long name code form.

Where a birthplace is not legible on microfilm, put a question mark (?) in column 53 and leave column 52 blank.

Many birthplaces are abbreviated on the microfilm - some abbreviations are given beside the name in the following list.

PLACE OF BIRTH CODES INCLUDING ALL 1861 & 1871 MENTIONS

AFRICA	AF
AT SEA	SE
AUSTRALIA	AS
AUSTRIA	AU
BAVARIA	BA
BELGIUM	BE
BRITISH COLUMBIA	BC
CANADA	CA
CANADA EAST	CE
CANADA WEST	CW
CANADA WEST	WC
COLOURED AFRICAN	CF
DENMARK	DE
EAST INDIA	EI
ENGLAND	EN
FRANCE	FR
GERMANY	GE
GREEK	GR
HOLLAND	DU
ILLEGIBLE	?
ILLEGIBLE	?
ILLEGIBLE	IL
IRELAND	IR
ITALY	IT
LOWER CANADA	CE
LOWER CANADA	LC
MANITOBA	MA
NATIVE (IN NB)	NA
NATIVE FRENCH (IN NB)	NF
NEW BRUNSWICK	EN
NEW BRUNSWICK	NB
NEWFOUNDLAND	NE
NORTHWEST	NW
NORWAY	NO
NOT GIVEN	NG
NOVA SCOTIA	NS
ONTARIO	CW
ONTARIO	ON
ONTARIO	UC
ONTARIO	WC
POLAND	PO
PORTUGAL	PT
PRINCE EDWARD ISLAND	PE
PRUSSIA	PR
QUEBEC	CE
QUEBEC	LC
QUEBEC	QU
RUSSIA	RU
SCANDINAVIA	SV

23

SPAIN .....SP
SWEDEN .....Sw
SWITZERLAND .....SI
TERRE NEUVE .....NE
UNITED STATES .....US
UPPER CANADA.....Cw
UPPER CANADA.....UC
UPPER CANADA.....wC
WALES .....WA
WEST INDIES .....WI

Columns 54-55: Religion

NOTE I: The religion codes are formed in the same way as birthplace: exceptions have an asterisk.

NOTE II: As for birthplaces - no code is given for a religion, place * in col. 55 and code on long name form.

NOTE III: As for birthplace, if illegible, put question mark (?) col. 55 and leave col. 54 blank.

NOTE IV: Religions may also be abbreviated, some are listed below.

(See codes for religion next page)

RELIGIOUS CODES INCLUDING ALL 1861 & 1871 MENTIONS

ADVENTIST .....AD
AFRICAN ASSOCIATION BAPT.....AA
AMERICAN PRESBYTERIAN .....AP
ATHEIST .....AT
B(R) METH (E) .....BM
BAPTIST .....BA
BIBLE BELIEVER .....BB
BIBLE CHRISTIAN METHODIST ....BC
BME .....BM
BRITISH EPISCOPAL METHODIS ....BE
BRITISH EPISCOPAL METHODIS ....BM
C BRITIEN .....CE
C PRESB .....CP
CALVINISTIC METHODIST .....CM
CANADA PRES .....CP
CANADIAN PRESBYTERIAN .....CP
CHRISTIAN BAP CC
CHRISTIAN BAP .....CD
CHRISTIAN BAP .....CF
CHRISTIAN BRETHERN .....CH
CHRISTIAN CONF CC
CHRISTIAN CONF .....CD
CHRISTIAN CONF .....CF
CHRISTIAN CONFERENCE BAPT CC
CHRISTIAN CONFERENCE BAPT ....CD
CHRISTIAN CONFERENCE BAPT ....CF
CHURCH OF ENGLAND .....CE
CHURCH OF SCOTLAND .....CS
CONGREGATIONALIST .....CO
DEIST .....DE
DISCIPLE.....DI
DISCIPLE OF CHRIST .....DI
E .....CE
E CHURCH.....CE
E METH .....EM
E METHODIST .....EM
EPISCOPAL (OF ENGLAND) .....EP
EPISCOPAL METHODIST .....FM
EPISCOPAL OF ENGLAND.....BE
EPISCOPAL OF ENGLAND.....BM
EPISCOPALIAN.....EP
EVANGELICAL ASSOC .....EA
EVANGELIST .....EV
F (C) OF S .....FC
F (C) OF S .....FP
F BAPTIST .....FB
F C (P) .....FC
F C (P) .....FP
F C BAPTIST .....FB
FREE CHRISTIAN .....FR
FREE WILL .....FW
FREE WILL BAPTIST .....FW

26

ILLEGIBLE	IL
IND	IN
INDEPENDENT	IN
IRVINGITE	IR
JEW	JU
K PRESBYTERIAN	SP
KIRK (OF SCOTLAND)	SP
LATTER DAY SAINTS	LD
LUTHERAN	LU
MENNONITE	MN
MESSIAH	MS
METH E	EM
METHODIST	ME
MORMON	MO
MR	MM
NEW CONNEXION	NC
NEW JERUSALEM(C,CH,CHURCH)	SW
NO DENOMINATION	NR
NO RELIGION	NR
NO SECT	NR
NOT GIVEN	NG
PAGAN	PA
PLYMOUTH BRETHERN	PB
PRESB C OF SCOTLAND	SP
PRESB C S	SP
PRESBYTERIAN	PS
PRESBYTERIAN C. OF L.P.	PL
PRESBYTERIAN KIRK	SP
PRIMITIVE METHODIST	PM
PROTESTANT	PR
QUAKER	QU
R BAPT(IST)	RB
R PRESB	RP
RE BAPTIST	RB
REFORM BAPTIST	RB
REFORMED BAPTIST	RB
RF BAPTIST	RB
ROMAN CATHOLIC	RC
S KIRK	SP
S PRES(B)	SP
SCO PRESBY(TERIAN)	SP
SCOTCH PRESBY(TERIAN)	SP
SEVEN DAY ADVENTIST	SD
SWEDENBORGIAN	SW
TUNKER	TU
U KIRK PRESB	UP
U P	UP
U P PRESB	UP
U PRESB(YTERIAN)	UP
U S PRESB	AP
UN PRESB	UP
UNION BAPTIST	UN
UNIT PRESBYT	UP
UNITARIAN	UT
UNITE PRESB(YTERIAN)	UP
UNITED BRETHERN	UB
UNIVERSALIST	UV
W	WM
W C METH	WM
WESLEYAN	WM

Column 56-57: Nation of Origin - Ethnic Origin.

NOTE I: These codes are very similar to British codes, but a few differ: note them: them are underlined in the list which follows.

NOTE II: As For birth place, religion - if no code is given for an ethnic or naiton of origin, place * in col. 57, and code name on long name form.

NOTE III: As above, if illegible put question mark (?) col. 57, leave col. 56 blank.

NOTE IV: Ethnic-Nation of Origin listings are abbreviated as for birthplace.

(See next page for codes for Ethnic or Nation of Origin)

NATION OF ORIGIN CODES INCLUDING ALL 1861 & 1871 MENTIONS

ACADIAN .....AC
AFRICAN .....AF
AMERICAN.....US
ANGLO-SAXON .....AX
AUSTRIAN.....AU
BAVARIAN.....BA
BELGIAN .....BE
CANADIAN.....CA
DANISH .....DE
DUTCH .....DU
EAST INDIAN .....EI
ENGLISH .....EN
FRENCH .....FR
GERMAN .....GE
GREECE .....GR
HALF BREED .....HB
HINDOO .....HI
ILLEGIBLE ..... ?
ILLEGIBLE .....?
ILLEGIBLE IL
IRISH .....IR
ITALIAN .....IT
JEWISH .....JU
NATIVE INDIAN .....IN
NATIVE INDIAN .....NI
NORWEGIAN .....NO
NOT GIVEN .....NG
POLISH .....PO
PORTUGESE .....PT
PRUSSIAN.....PR
RUSSIAN .....RU
SCANDINAVIAN.....SV
SCOTTISH.....SC
SPANISH .....SP
SWEDISH .....SW
SWISS .....ST
WELSH .....WA

Columns 58-73: Occupation - Profession: Complete title as on micro-film manuscript.

If name exceeds 16 columns fill in last column (col. 73) with asterisk (*) and complete title on the long name code form (#2).

If an occupation is scratched out on the original form - but you can decipher it clearly, then code it as usual.

Column 74: Married or widowed; code M or W. If blank, code blank.

NOTE: If next item on the microfilm "married in last 12 months" is filled in - code L for married and W for widowed.

Column 75: School (attendance)
Code 1 if marked (usually a 1), otherwise leave blank.

Column 76: Unable to read: - labeled "read" on code form - Code 1, if marked, otherwise, blank.

Column 77: Unable to write: - labeled "write" on code form - Code 1, if marked, otherwise blank.

Column 78: Deaf and dumb: - Code 1 if checked, otherwise, blank.

Column 79: Blind: - Code 1 if checked, otherwise, blank.

Column 80: Unsound mind: - Code 1 if checked, otherwise, blank.

Processing of the Coded Data

The computer programs noted above are designed to deal with this data to accomplish two tasks: they carry out a series of logical checks on the coded data to allow coding errors to be corrected; and they take the original data and transform it into files on which data analysis can proceed. The two tasks are intimately related. For example, if the "blank *" is found in the religion field for a given individual, it is necessary to make certain that a long form containing the religion for this individual has also been created. When the data are transformed into files for analysis, the contents of the religion field on the long form must be merged into a specific position on the record for the relevant individual. Besides checking for the existence of long-form data, flagged by asterisks on the individual records, a number of other checks of the data are carried out. The entries for variables with a fixed set of codes, including religion, place of birth, nation of origin, marital status, and school attendance, are checked to see that they are among the predetermined acceptable codes. In addition, some infrequently occurring combinations of codes are also flagged by the program so that they can be checked. These include individuals who are listed as attending school but are under four or over nineteen years of age, married persons without a spouse present in the household, and all individuals listed as deaf and dumb, blind, or of unsound mind. These cases were scrutinized for possible error and many were checked against the microfilm records.

The entire census of 1871 does not record the relationship of the individuals in a household to each other.

If some error is tolerated, however, it is possible to deduce relationships among individuals within a household, making use of surnames, marital status, age, and the ordering in which an individual within a household are recorded on the census manuscript. A decision was made to carry out such an analysis for each household and using this information to attach to each individual record a number of summary variables describing the household of which he or she was a member (see below). Clearly there were cases where the family relationships are ambiguous. In all the detectable cases of ambiguity, a message was printed out by the data-checking programme to apprise us of the difficulty. For example, a child could logically have more than one person in the household as its mother--logically here being taken to mean that there are two or more women in the household with the same surname as that child, who are married or have been married, and whose ages differ by at least fifteen and no more than fifty years from that of the child. In all such ambiguous cases an "error" message was printed and the household scrutinized by the principal investigators to attempt to resolve the ambiguity. In the great majority of cases, the determination of family relationships among members of a household was unambiguous.

Three other potentially ambiguous situations were flagged by the program: married individuals without their spouses present (most of these proved to have been correctly coded), individuals who appear to be the children of those who are listed later in the sequence of persons in the household mostly this seems to indicate a widowed parent or aged couple living in the household of their child), and individuals who are identified as children of parents in the household, but who are separated from their parents by one or

more persons of a different surname.

In each of these cases, a "warning" message led us to reexamine the household. A "special allocation" procedure was developed whereby any alteration of the application of the computer program's rules for establishing family relationships could be recorded and the "imposed" relationships changed in the final data file. See Figure 4 and the accompanying "Layout" description for the form and kinds of "special allocations" permitted. We discuss the nature of this intervention by means of "special allocations" and provide illustrative cases in the following text.

re 4:
ECIAL
LOCATIONS

NS = No Spouse
SP = Spouses
NP = No Parent
CX = Children (type x)
HS = Household Size

	DISTRICT	SUBDISTRICT DIVISION	HOUSEHOLD NUMBER (1871 ONLY)	TYPE OF ALLOCATION	HUSBAND OR FATHER	WIFE OR MOTHER	FIRST CHILD	LAST CHILD (can be blank)
	1–5		6–8	9–10	11–16	17–22	23–28	29–34
1.								
2.								
3.								
4.								
5.								
6.								
7.								
8.								
9.								
10.								
11.								
12.								
13.								
14.								
15.								
16.								
17.								
18.								
19.								
20.								

Special Allocation Cards: Layout

(NOTE: 1861 data for a local study of Essex and Kent counties are referred to below).

1-3 District of household
4 Subdistrict of household
5 Division (enumeration area) of household

6-8 for 1861: blank
for 1871: household number

9-10 type of special allocation:

NS-- no spouse, used to prevent the assignment of two married adjacent individuals of opposite sex and with the same last names as a couple
SP-- spouses, used to assign two individuals as spouses who are non-adjacent
NP-- no parents, used to prevent the assignment of an individual as the child of any other person in the household
Cx- where x is 1, 2, 3, 4, 5, or 6--used to assign a child or children to a parent or parents other than those which the programme would automatically choose. The child or children are of type x,

x = 1 means child of couple
x = 2 means child of married
x = 3 means child of widower
x = 4 means widow
x = 5 father and stepmother
x = 6 mother and stepfather

HS--household size, used to break a household into 2 or more units

11-16, contain the numbers of four persons in either of the following two
17-22, forms (one or more may be blank in any given case)
23-28, form I: person number of the individual, counted from the first, right-justified.
29-34, form II: xxyyyy, For 1861-xxx is the page number, yyy is the person number of the individual. For 1871--xxx is the family number and yyy is the original person number. N.B. should either of these be in error on the original (which means they will be corrected by the programme), use the original values.

The positions are used as follows (left blank if they do not apply)
11-16--husband or male parent
17-22--wife or female parent
23-28—(first) child
29-34—(second) child
N.B. if both the person numbers for children are filled, then the programme assumes that all persons in between are to be treated as children.

Some Illustrations

	Type	11-16	17-22	23-28	29-34
1. Persons 5 and 9 are married	SP	5	9		
2. Persons 9 and 10 are not married	NS	9	10		
3. Persons 4 through 7 have no parents in the household	NP	b	b	4	7
4. Person 4 and 7 have no parents	NP	b	b	4	
in the household (N.B. requires 2 cards)	NP	b	b	7	
5. Persons 4 through 7 have persons 2 and 3 as father and mother respectively (i.e. type 1 children)	C1	2	3	4	7
6. Person 4 and person 7 have persons 1	C6	2	1	4	
and 2 as mother and step father respectively (i.e. type 6 children, requires 2 cards)	C6	2	1	7	

To cause the household to be broken into more than one unit for analysis, the four person numbers must contain the person numbers (counting from the first individual only and not using the original family or person numbers) of the last persons in each subunit.

e.g. if a household with 20 persons is to be analyzed in 3 units, 1-8, 9-13, 14-20, person numbers used are 8, 13, 20, b.

If the split is into more than five groups, two cards must be employed--in this case the second card must follow the first in deck placement and should be based on an entirely new count of the household. Say we wish to break a fifty person household at persons 3, 18, 25, 30, 36, 42, 50.

Then the two cards must read	HS	3	18	25	30
	HS	6	12	20	b

Should it be necessary to include special allocation data on parents and children simultaneously with households size information, then the parent and child data should be keyed to the original family and person number data.

We think that automatic creation of relationships, aided by our intervention in all ambiguous case, is an acceptable substitute for an original manuscript variable describing these relationships. Such a variable is a critical one in analysis.

Part 2 of this appendix fully describes the algorithm used to analyze the relationships with each household.

Two kinds of variables are generated which correspond to those described above as coded directly from microfilm. The first is a set of

summary variables describing the entire household which are attached to the record of each individual in the household: Among these variables are the number of people in the household, the number of married couples in the household, the number of children of widowers, and so on. The complete set of variables is also given in Part 2 of this appendix. The second type of variables describes each individual uniquely; they have different values for each person in the household. Examples of these variables are a child's number of older and of younger brothers and sisters resident in the household and the person number (i.e., position within the household) of each person's mother and father (some variables are chiefly of interest when it is necessary to create new summary variables for the household, for they allow the household to be read into storage and reanalyzed without going through the process of redefining the basic relationships).

Built into the family relationships algorithm is a set of decision rules to be followed at ambiguous points--for example, if two individuals could "logically" be taken as the mother of a given child, the algorithm assigns the child to the "mother" who is closest in the household listing. As noted above, a warning message is printed when this happens. What if our reexamination of a specific household leads us to conclude that the wrong person has been selected automatically as the mother? In such a case a "special allocation" form is filled out, from which a card is keypunched. The card contains an instruction to the programme to reallocate the parent-child relationship. In this case, when the raw data are reprocessed a message indicating this change is printed when the household is encountered, and a variable on the record records the fact that this "special allocation" has taken place.

Further processing of the basic data was required. Four variables require substantial recoding from alphabetic to numerical codes to be usable

in the data analysis. They are occupation, religion, place of birth, and nation of origin. The last three are two character codes, but a code for the complete set of occupations for such a large file has as many as forty characters! The four recording tasks are carried out at a single step: the programme first reads in a "dictionary" which specifies how the alphabetic codes are to be transformed, it then goes through the individual records and "looks up" the words in the dictionary and adds to the record the corresponding numeric codes from the dictionary.

The three two character variables can be handled fairly easily since there are no more than about eighty valid codes for each one of them; it is not difficult to assign numeric codes to the majority of the valid codes before the data are even coded. But there is nothing approximating a complete list of the occupations which can be found in the census abstracts. The aggregate census lists only about one hundred and twenty occupations, though over one thousand are found in our data. So, before the dictionary can be made up, the individual records must be analyzed by a program which identified all the unique occupational mentions. The program developed for this project punches a card for each occupation mentioned and these cards can then be used to make up the dictionary. Each occupation is assigned an eight digit code. Of course, each unique spelling of an occupation must be treated separately. The dictionaries are each described in detail in Part 3 of this appendix.

There is a final problem which is caused by the long forms. For example, the record for some individuals does not contain a valid two character religion code, but rather a "blank *" code and a 24 character string with the unique religion written out. All the infrequently occurring religions are coded in this way. Our solution is to place a three-digit numeric code for those religions coded on "long forms" in the last three positions of the 24

character string, i.e., the numeric code is simply punched onto the long form card itself after the data were collected. When the computer programme encounters a "blank *1" in the religion field of the original record of an individual$_1$ it scans the 22nd, 23rd, and 24th positions of the long form card to find the code. The long form values for place of birth, nation of origin, and occupation are handled in exactly the same way.

Finally certain cross-references among individuals within a household are likely to be of quite common use, though they certainly do not exhaust the possibilities. In particular, it will often be of interest to examine the characteristics of an individual in relation to five specific individuals in the household: his or her spouse, mother, father, family head, and household head (of course, the same person could fill more than one of these relationships). On our file each individual's record contains six additional variables describing these five individuals in terms of their occupation, religion, place of birth, nation of origin, age and sex. Thus, for example, one could examine the relationship between school attendance and a child's father's occupation and his or her mother's place of birth using only the data already on the records.

Appendix E: Part 2

Use of the Automatic Household Analysis Program

Ordinarily all data sets are analyzed twice. In the first run, no final records are created. The list of warnings, signalling ambiguities in the identification of parent-child relationships, married persons without spouses, and so on, are carefully examined and a special allocations form filled out for each case where the computer algorithm produces an incorrect result (see above). These special allocation forms are then used in the second run of the data, when final records are created.

Variables Created by the Automatic Household Analysis

As indicated above, two types of variables are created by the analysis: a set of summary variables which are attached to every person in the household which describes that household's general characteristics, including its size, the number of married couples in the household, the number of children with parents in the household, the number of servants in the household, etc.; and a set of variables which are unique to each person in the household, they include the number of older brothers an individual in the household has, the person number of his or her father, etc. These variables are listed below.

Steps in Automatic Household Analysis

1. The file of special allocation forms is searched to find any forms which refer to the household in question. The instructions on these forts override all automated allocations produced by the algorithm.

These special allocations permit the following programme overrides:

a. two individuals whose records do not meet the required conditions to be identified as a man and wife can be so designated,

b. two individuals who are automatically identified as man and wife can be separated from each other,

C. a child can be identified as having a specific parent or parents when they vould not so be identified automatically,

d. a child who is automatically identified as the son or daughter of a specific individual or couple can be separated from them,

e. child who is automatically identified as having two biological parents in the household can be identified as having one stepparent and one biological parent; also a person automatically identified as a stepparent can be identified as a biological parent,

f. the household can be broken into two or more groups of individuals which are analyzed as separate families or households and not together.

2. The marital status and presence of a spouse of the head of the household (the head is taken simply as the first person listed in the household) are ascertained.

3. All servants and visitors in the household are identified, using the occupation variable. Since a number of possible designations of a servant are possible in the manuscript data, e.g., servant, maid servant, general servant, servente, sevt, etc., the occupation of each individual (except for the household head who cannot be a servant) is compared to a list of occupations previously identified as including all the possible occupational

titles referring to servants and also all the misspellings of those titles which are found in the data. This procedure, of course, requires a preliminary examination of the occupational titles which occur in the entire file. Visitors to the household are identified only by the exact title 'VISITOR' in the occupation field.

4. All married couples in the household are identified. A couple must have identical last names, must both be listed as married, and must be listed on adjacent lines in the order of the household. If, as occasionally occurs, the two spouses are not listed sequentially or the maiden name of the wife is given as her last name (for example, in rare cases this appears in Quebec in the 1861 census) a special allocation form must be used to instruct the programme that the two individuals are married All such cases are identified and reexamined.

5. For all married persons in the household for whom a spouse cannot be identified, a warning message is issued, the processing then proceeds normally. In most cases a subsequent recheck of the microfilm found there was, in fact, no spouse present. occasionally, this warning led to our finding some coding error.

6. Widows and widowers are identified.

7. All probable parent-child relationships in the household are identified. A mother and child must have identical surnames and the age difference between them must be between 15 and 50 years. A father and child must have identical surnames and the age difference between them must be at least 17 years. For any other cases a special allocation form must be used

to create a parent-child relationship.

In those cases where, according to these criteria a child could have more than one person as his or her mother or father, a warning message is printed. The household processing is then carried on under the assumption that the actual parent is the one closest to the child in the listing of the household.

A warning message is printed whenever a child is found to be listed before his or her probable parent in the household. This is usually the case when an aged parent resides with the family of one of his or her children. A warning message is printed when a person identified as a child of a probable parent in the household is separated from that parent in the listing of individuals by one or more people with a different surname, for example, in the case of younger stepchildren. It is important to note that where a child-parent relationship exists but their surnames are not identical, due to a name change at marriage or to remarriage, the automatic routines will fail to identify the relationship. If the relationship were known from some other data source or investigators were willing to deduce it from an inspection of the household, a special allocation form could be used to instruct the program to identify the relationship as part of the file. If a child is identified as related to only one member of a married couple--as occurs where the child's surname is identical to that of the couple, but the age comparisons indicate that only one of the couple is likely to be a parent--a warning message is issued. The child is identified as having a stepparent and data processing proceeds normally. In cases where a visual examination of the household suggests that there are both children and stepchildren (this is usually revealed by the presence of two distinct age-ordered groups of siblings), a special allocation form can be used to treat one group as stepchildren, even if the age requirements for biological parenthood are

satisfied for all.

8. Daughters-in-law are identified, they are the wives of men for which at least one of the parents is found in the household. If the family lives with the parents of the wife, the in-law relationship cannot be identified without using outside sources, because of the name change at marriage.

9. For each set of siblings identified above, a set of variables to measure the number of older and of younger brothers and sisters (separately) is computed.

Below we reproduce some selected examples of the output of The Automatic Household Analysis Program. Each case includes a copy of the program's print-out to which we have attached short captions noting the nature of the ambiguity or error indicated by the message. Each case also gives a copy of the original coding form.

```
**SEPCHILD IN ERROR FOR PERSON   1C 1  456 214   101  9              2     1     0     1
**REVCHILD IN ERROR FOR PERSON   1C 1  456 214   101  1              2     6     1     1

 1 C 1  456 214   101   1001    RENARD          ELI             M  28 CN  RC  FR   TAVERNKEEPER      M   1 1
 1 C 1  456 214   101   2002    RENARD          JOSEPHINE       F  46 CN  RC  FR                     M   1 1
 1 C 1  456 214   101   3003    BROOKER         WILLIAM         M  18 CN  RC  EN   FARMER
 1 C 1  456 214   101   4004    BROOKER         JAMES           M  16 ON  RC  EN   FARMER
 1 C 1  456 214   101   5005    BROOKER         ELIZABETH       F  14 CN  RC  EN
 1 C 1  456 214   101   6006    BROOKER         THOMAS          M  14 ON  RC  EN
 1 C 1  456 214   101   7007    BROOKER         HENRY           M  12 CN  RC  EN
 1 C 1  456 214   101   8008    BROOKER         ELLEN           F  10 ON  RC  EN
 1 C 1  456 214   101   9009    RENARD          ELLY            F   2 CN  RC  FR
```

The error messages are given above the household listing on this run. The messages SEPCHILD and REVCHILD refer, respectively, to person number 9, Elly Renard, and person number 1, Eli Renard. The SEPCHILD message identifies Elly, aged 2, as the possible daughter of Eli and Josephine Renard, given the surname similarity and age differences. But the relationship is ambiguous because Elly is separated from her possible parents by six Brooker children. Elly is probably the sister of Josephine's six surviving children by a previous marriage to a Mr. Brooker. The automatic household analysis program would have assumed this relationship, assigning Elly to the closest probable parents. The "error" message here serves as a warning to the investigator that the ambiguity warrants consideration and possibly the automatically established relationships should be altered by using the Special Allocation feature of the program. We did not alter the allocations in this case. Note that Elly is given as being of french origin, like her assumed father, Eli, while the other children are of english origin.

The REVCHILD message identifies Eli Renard, aged 28, as a possible son of Josephine Renard, aged 46, merely because of the name similarity and the age differential (18 years). The automatic program will also have recorded this relationship. On the basis of the other information (age sequence of the children, married couple indicated) this ambiguity is resolved by overriding the automatic allocation by means of a Special Allocation, making Eli the husband of Josephine, as assumed above.

	Subdistrict	E.D.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
1	2	3 5	6 9	10 12	13	14 15	16 31	32 47	48	49 51	52 55	56 58	59 62	63 73	74	75	76	77	78	79	80
21	C	1	0456	214	1	01	RENARD	ELI	M	028	ON	RC	FR	TAVERNKEEPER	M		1	1			
						02		JOSEPHINE	F	046					M		1	1			
						03	BROKER	WILLIAM	M	018			EN	FARMER							
						04		JAMES	M	016				FARMER							
						05		ELIZARETH	F	014											
						06		THOMAS	M	014											
						07		HENRY	M	012											
						08		ELLEN	F	010											
						09	RENARD	ELLI?	F	002			FR								
21	C	1	0443	231	1	01	MELOCHE	ANTONIO	M	034	ON	RC	FR	HUNTER	M		1	1			
						02		MARY	F	029					M		1	1			
						03		ELIZABETH	F	006											
						04		CECELIA	F	004											
						05		ROSIA	F	001											

CODER'S INITIALS DTA DATE 1 1 PAGE NUMBER 10

***AGE---SRW IN ERROR FOR PERSON 69C 2 702 31 101 4 -247447400 0 0 0 1

69	C	2	702	31	101	1001	HOLMS	GEORGE	M	61	EN	WM	EN	FARMER	M
69	C	2	702	31	101	2002	HOLMS	MARY	F	47	GN	WM	EN		M
69	C	2	702	31	101	3003	HOLMS	MARGRET	F	30	CN	WM	EN		
69	C	2	702	31	101	4004	HOLMS	MURCUS	M	(25)	CN	WM	EN	FARMER	
69	C	2	702	31	101	5005	HOLMS	ALBERT	M	13	ON	WM	EN		1
69	C	2	702	31	101	6006	HOLMS	HARVY	M	7	CN	WM	EN		1

The error message above the case, AGE---SRW, indicates for person number 4, Murcus Holms, aged 25, a farmer, is listed as attending school. In this case a subsequent review of the microfilm record indicated that no coding or punching error was made and the record is maintained.

Subdistrict	E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
1	2	3 4 5	6 7 8 9	10 11 12	13 14 15	16 31	32 47	48	49 51	52 53	54 55	56 57	58 73	74	75	76	77	78	79	80
9	A	2	0702	031	01	HOLMS	GEORGE	M	061	EN	WM	EN	FARMER	M						
					02		MARY	F	047	ON				M						
					03		MARGARET	F	030											
					04		MURRAYS	M	025				FARMER		1					
					05		ALBERT	M	013						1					
					06		HARVS	M	007						1					
9	A	2	0801	220	01	CAMPBELL	ARCHIBALD	M	043	SCFD	SC		LABOURER	M						
					02		ANN	F	041	IR				M						
2	C	3	0802		01	COOK	ABEL	M	038	ONEMEN			CARPENTER	M						
					02		DAVINA	F	007						1					
					03		ISABELLA	F	006											
					04		IDA	F	004											
					05		GEORGE	M	039		BA		LABURER	W						
					06		JAMES	M	022		FM		CARPENTER							

CODER'S INITIALS AC DATE __/__/__ PAGE NUMBER 06

```
106 B 0 1101 217 1  1     BRAS          ?  BENEDICT        M 031 GE  JE  GE   IMPORTER        M
106 B 0 1101 217 1  2                       SOPHIE         F 027                             M
106 B 0 1101 217 1  3                       ELLEN          F 004 QU
106 B 0 1101 217 1  4                       IVA            F 003
106 B 0 1101 217 1  5     LAURENCE          SARAH          F 020 CN  CE  SC   SERVANT
106 B 0 1101 217 1  6     WIX               CATHERINE      F 019 IR  RC  IR   SERVANT

***RELIGION IN ERROR FOR PERSON NUMBER C 1101 217  1  1          347     0     0    0
```

The error message below the case, RELIGION IN ERROR, indicates that for person number 1, JE is not a valid code (in columns 54-55 of the coding form). A check of the microfilm showed that the actual religion was Jew and the valid code would be JU. The punch card was altered and the file corrected. The corrected code is also recorded in the right margin of the coding form.

Subdistrict	E.D.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
BC	1402		101	1	01	MCCULLOUGH	ANN	F	068	QU	CE	EN		W						
					02	HUBBELL	MARGARET	M	DH	8			TEACHER	M						
					03	MCCULLOUGH	EMILY	F	036		CS									
					04		JOHN	M	027		CE		BANK CLERK							
					05	HUBBELL	HELEN E	F	018	ON										
					06		GEORGE H	M	016	QU			CLERK							
					07		WILLIAM	M	015											
					08		STUART	M	008											
					09	DILLON	JANE	F	054	IR		IR	SERVANT	W						
BC	1401		2.71		01	ERAS	? BENEDICT	M	031	GE	CE	GE	IMPORTER	M						
					02		SOPHIE	F	027					M						
					03		ELLEN	F	004	QU										
					04		IVA	F	003											
					05	LAURENCE	SARAH	F	020	ON	CE	SC	SERVANT							
					06	WILX	CATHERINE	F	019	IR	RC	IR	SERVANT							

CODER'S INITIALS DA DATE 1/1/ PAGE NUMBER 14

```
***MARNO SP IN ERROR FOR PERSON 1740 1 3202 226   1C1  9              C     C     O     C
```

174	D	1	3202 226	101	1001	STEVENS	RCBT	M	63	NB	BA	EN	BOARDING HOUSE M
174	D	1	3202 226	101	2002	STEVENS	MRS R	F	63	NB	BA	EN	M
174	D	1	3202 226	101	3003	STEVENS	BEVERLY	M	29	NB	BA	EN	
174	D	1	3202 226	101	4004	STEVENS	C E	M	19	NB	BA	EN	
174	D	1	3202 226	101	5005	SIME	P C	F	37	NB	BA	EN	W
174	D	1	3202 226	101	6006	SIME	FRANK	M	14	NB	BA	EN	
174	D	1	3202 226	101	7007	SIME	MARY	F	8	NB	BA	EN	1
174	D	1	3202 226	101	8008	SIME	R S	M	3	NB	BA	EN	
174	D	1	3202 226	101	9009	WEDDERSPOON	JUNE	F	21	EN	CE	EN	SERVANT M
174	D	1	3202 226	101	10010	GILFORD	MARY	F	20	WA	RC	IR	SERVANT
174	D	1	3202 226	101	11011	QUIRK	MARY	F	21	EN	CE	EN	SERVANT
174	D	1	3202 226	101	12012	STOCKLM	MARY	M	24	SC	CS	SC	CLERK
174	D	1	3202 226	101	13013	BELYEA	A	M	28	NB	BA	EN	EXPRESS DRIVER
174	D	1	3202 226	101	14014	CHANDLER	C H	M	30	NB	CE	EN	POLICE OFFICE CE
174	D	1	3202 226	101	15015	CHISHOLM	P	M	35	NS	PL	SC	GROCER W
174	D	1	3202 226	101	16016	CASE	JOHN	M	20	NB	CE	EN	
174	D	1	3202 226	101	17017	GOLDING	E	M	19	NB	CE	EN	CLERK
174	D	1	3202 226	101	18018	GILCHRIST	R	M	25	SC	CS	SC	CLERK
174	D	1	3202 226	101	19019	LAWSON	W S	M	18	NS	BA	EN	CLERK
174	D	1	3202 226	101	20020	FOSTER	F F	M	22	NS	BA	EN	CLERK
174	D	1	3202 226	101	21021	MCGRR	S	M	25	NB	PL	IR	GROCER
174	D	1	3202 226	101	22022	MCDONALD	M	M	24	NB	BA	SC	ATTORNEY AT LAW
174	D	1	3202 226	101	23023	RANKIN	M	M	25	NB	PL	IR	CLERK
174	D	1	3202 226	101	24024	WHITE	L A	M	19	NB	WM	EN	STUDENT AT LAW

The case serves to illustrate the nature of some boarding houses considered as households. The error message MARNO SP, indicates that person number 9, June Wedderspoon, apparantly a member of the household staff, is listed as married but without an adjacent spouse. A check of the microfilm confirmed the record.

HOUSEHOLD INFORMATION-1971 CENSUS OF CANADA

Subdistrict E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
101	32,01	22,01	01	01	SAVAGE	TOM	M	035	ENG	CEN	CEN	CLERK	M						
				02		ELLEN	F	025											
101	32,02	22,61	01	01	STEVENS	ROBT	M	063	NB	BAPT	EN	BOARDING HOUSE	W						
				02		MRS R	F	063											
				03		BEVERLY	M	029											
				04		G E	M	049											
				05	STONE	P C	F	037											
				06		FRANK	M	014											
				07		MARY	F	008											
				08		R S	M	003											
				09	WEDDERSPOON	JUNE	F	021	ENG	CEN	CEN	SERVANT	M						
				10	GILFORD	MARY	F	020	WAL	RC	IR	SERVANT							
				11	QUIRK	MARY	F	021	ENG	CE	CEN	SERVANT							
				12	STOCKUM	—	M	024	SC	CS	SC	CLERK							
				13	BELYEA	A	M	028	NB	BA	EN	EXPRESS DRIVER							

CODER'S INITIALS PAT JENSEN DATE MAY 4/72 PAGE NUMBER 55

Subdistrict	E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
01		3202	22	01	14	CHANDLER	CH	m	03	ONB	CE	EN		X						
					15	[illegible]	P	m	03	5NS	PL	S,C	GROCER	W						
					16	CASE	JOHN	m	02	ONB	CE	EN								
					17	GOLDING	F	m	01	9			CLERK							
					18	GILCRIST	R	m	02	5S	CS	S,C	CLERK							
					19	LAWSON	WS	m	01	8NS	BA	EN	CLERK							
					20	POSTER	FF	m	02	2			CLERK							
					21	MCGEARR	S	m	02	5NB	PL	ER	GROCER							
					22	MCDONALD	M	m	02	4	BA	SC	ATTORNEY AT LAW							
					23	RANKIN	M	m	02	5	PL	ER	CLERK							
					24	WHITE	L	m	01	9	NB	WM	EN	ST	UDENT AT LAW					

CODER'S INITIALS JS DATE __/__/__ PAGE NUMBER __

```
****REVCHILD IN ERROR FOR PERSON 174D 2 3502  48   101  1              4     4     1     0

   174 D 2 3502  48   101  1001   RALSTON          THOMAS G          M  24 QU  BA  IR    B & SHOE MANUFAT M
   174 D 2 3502  48   101  2002   RALSTON          REBECCA           F  21 QU  BA  SC                    M
   174 D 2 3502  48   101  3003   RALSTON          ANNIE             F   0 NB  BA  IR
   174 D 2 3502  48   101  4004   RALSTON          ANN               F  49 EN  BA  EN                    W
   174 D 2 3502  48   101  5005   MCLOY            ELIZABETH         F  16 IR  RC  IR    SERVANT
```

The error message given above the case, REVCHILD, indicates that person number 1, Thomas Ralston, aged 24, has been automatically allocated as a child of a parent listed after him in the household, ie., Ann Ralston, aged 49, a widow. The allocation seemed highly probable here and was not altered. In general, nineteenth century census enumerators in Canada seemed to reflect the household structure by listing first whoever had assumed the position of the head of the household followed by their spouse, if any, their children in order of their birth, other relatives, with boarders, servants and visitors last.

Note too that the coding form only gives an "*" in the last column for occupation while this version of the automatic program provides the complete occupational title taken from the long name forms.

Subdistrict	E.D.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
1	2 3	4 5	6 9	10 12	13	14 15	16 31	32 47	48	49 51	52 53	54 55	56 57	58 73	74	75	76	77	78	79	80
1	02	3501	0601	1	01	SWEDEN	ELIZA	F	064	NB	CE	EN		W						
					02		ROBERT	M	027				SHIP BROKER							
					03		ISABEL	F	024											
					04		MARY	F	022											
1	02	3502	0481	1	01	RALSTON	THOMAS G	M	024	QU	BA	ER		*m						
					02		REBECCA	F	021				S.C	m						
					03		ANNIE	F	000	NB		IR								
					04		ANN	F	049	EN		EN		W						
					05	MELOY	ELIZABETH	F	016	IR	RC	IR	SERVANT							
1	02	3601	1071	1	01	SMITH	DANIEL	M	057	NS	BA	IR	HOUSE CARPENTER	m						
					02		JANE	F	047	NB	WM			m						
					03		EDWARD	M	020				MARINER							
					04		WILLIAM	M	017											
					05		MARIA	F	014											
					06		DANIEL	M	008											

CODER'S INITIALS PAT JENKINS DATE MAY 4 1975 PAGE NUMBER 58.

```
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101  5          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101  6          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101  7          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101  8          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101  9          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101 11          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101 12          2    2    0    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101 11          4    0   10    0
***WHOCHIL? IN ERROR FOR PERSON 183A 1  201 117   101 12          4    0   10    0
***REVCHILD IN ERROR FOR PERSON 183A 1  201 117   101  1         10    4    3    0
***REVCHILD IN ERROR FOR PERSON 183A 1  201 117   101  3         10    4    4    0

183 A 1  201 117   101   1001   ROY          LEZZAR      M  43 NB  RC  FR  FARMER        M  1 1
183 A 1  201 117   101   2002   ROY          MARY        F  25 NB  RC  FR                 M
183 A 1  201 117   101   3003   ROY          TIMOTHY     M  30 NB  RC  FR  FARMER        M
183 A 1  201 117   101   4004   ROY          MARY        F  28 NB  RC  FR                 M
183 A 1  201 117   101   5005   ROY          RAPHEAL     M   8 NB  RC  FR
183 A 1  201 117   101   6006   ROY          JERCAM      M   6 NB  RC  FR
183 A 1  201 117   101   7007   ROY          MARY        F   4 NB  RC  FR
183 A 1  201 117   101   8008   ROY          ELLEN       F   2 NB  RC  FR
183 A 1  201 117   101   9009   ROY          ELIZEBETH   F   0 NB  RC  FR
183 A 1  201 117   101  10010   ROY          LUCIE       F  65 NB  RC  FR                 W
183 A 1  201 117   101  11011   ROY          TURECA      M  21 NB  RC  FR
183 A 1  201 117   101  12012   ROY          STEPHEN     M  20 NB  RC  FR  FARMER
183 A 1  201 117   101  13013   LAPLANTE     EDWARD      M  13 NB  RC  FR  FARMER
183 A 1  201 117   101  14014   LAPLANTE     JERCAM      M  12 NB  RC  FR
183 A 1  201 117   101  15015   LAPLANTE     JOSEPH      M  10 NB  RC  FR
183 A 1  201 117   101  16016   LAPLANTE     MARY        F   8 NB  RC  FR
183 A 1  201 117   101  17017   LAPLANTE     ELIZEBETH   F   6 NB  RC  FR
183 A 1  201 117   101  18018   LAPLANTE     JOHN        M   4 NB  RC  FR
183 A 1  201 117   101  19019   LAPLANTE     PHILEMON    F   2 NB  RC  FR
```

see description attached.

This case illustrates what we take to be a truly ambiguous case regarding the determination of the relationships among the household members. The ambiguities are indicated by the WHOCHIL? IN ERROR messages. They first indicate that the allocation of the five Roy children, numbers 5 to 9, could reasonably be to either of two sets of parents listed above them, to Lezzar and Mary Roy or to Timothy and Mary Roy. The names and age differences between the children and these couples do not resolve the ambiguity. The program automatically assigns the children to the potential parents listed most immediately above them, arbitrarily, but not unreasonably imposing a resolution. We did not alter this allocation.

There are four other WHOCHIL? IN ERROR messages, two each for persons numbered 11 and 12, Tureca (?) and Stephen Roy, aged 21 and 20. The first of the messages indicates that the previously mentioned couples could also be the parents of these two men - but the second message notes the fact that they are listed immediately after Lucie Roy, aged 65, a widow (column 74). The age difference between Lucie Roy and both these men (44 and 45 years) has the program conclude that she is their widowed mother. The program again assigns the two men to the closest previously listed potential parent or parents, in this case Lucie. The logic of the overall listings suggest to us the allocations are appropriate.

Thus the outcome of the program allocations in this case is a household in which the first couple is considered childless, or without children residing in the household, the second couple is considered to have five children living in the household while Tureca (?) and Stephen Roy are taken to be younger sons of Lncie Roy.

Two additional, REVCHILD IN ERROR, messages are given for persons numbered 1 and 3, indicating that both of these men, Lezzar and Timothy are possibly also the sons of the widow, Lucie Roy. Thc namc similarity and the age differences again are the basis for the message. The program's automatic allocation was not altered. The household is considered to have four surviving sons of Lucie living in it.

HOUSEHOLD INFORMATION-1871 CENSUS OF CANADA

Subdistrict	E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
A	1	0201	117	1	01	ROY	LEZZAR	M	043	N.B	RC	FR	FARMER	M		1	1			
					02		MARY	F	025					M						
					03		TIMOTHY	M	038	N.B	RC	FR	FARMER	M						
					04		MARY	F	038					M						
					05		RALPHEAL	M	008											
					06		JEROAM	M	006											
					07		MARY	F	004											
					08		ELLEN	F	002											
					09		ELIZABETH	F	000											
					10		LUCIE	F	065					W						
					11		THERESA	M	021											
					12		STEPHEN	M	020				FARMER							
					13	LAPLANTE	EDOARD	M	013				FARMER							
					14		JEROAM	M	012											
					15		JOSEPH	M	010											

CODER'S INITIALS _____ DATE __/__/__ PAGE NUMBER 06 cont'd

Subdistrict E.D.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
				16		MARY	F	009	WB	PC	FR								
				17		ELIZABETH	F	006											
				18		JOHN	M	004											
				19		PHILEMON	F	002											
01	0243	120	1	01	ROY	LUI	M	032	NB	RC	FR	FARMER	W		1	1			
				02		MARY	F	027					M		1	1			
				03		ELLEN	F	006						1					
				04		DENNIS	M	003											
				05		LORANCE	M	001											
				06		PAUL	M	031				LABORER							
				07		JOHN	M	016				LABORER							
				08	CULTON	MARY	F	026											
			2	01	RICHARDS	AUGUSTIN	M	042	QU	RC	FR	CARPENTER	M		1	1			
				02		DELMAS	F	031					M		1	1			

CODER'S INITIALS PA DATE 05 JUNE PAGE NUMBER 07

196	A1	101	2	1	1	LOWE	WILLIAM P	M	040	EN	WM	SC	SHOPKEEPER	M	
196	A1	101	2	1	2		SARAH	F	033			EN		M	
196	A1	101	2	1	3		ELIZABETH	F	009	NS					1
196	A1	101	2	1	4		AMELIA	F	005						1
196	A1	101	2	1	5	MCFIE	LIZZIE	F	021	PL		SC	SERVANT		

***PL BIRTH IN ERROR FOR PERSON 196A 1 101 2 101 5 582 0 0 0

This case illustrates the results of a single coding error. The error message below the household listing, PL BIRTH IN ERROR, for person number 5, Lizzie McFie, indicates that the mnemonic code PL ir columns 52 and 53 is not a valid birthplace code. The punched card and the coding form were checked and the error found on the form, as the copy included shows. A subsequent check of the microfilm located the source of the error. The coder had placed the correct code for religion in the birthplace columns.

For this case we provide a reproduction of the microfilm of the original manuscript record. It indicates the likely source of the coding error. The poor quality of the microfilm in this case, as in many others, is obvious, though somewhat exaggereated by reproduction. More specifically, the entry for religion was altered by the enumerator and is difficult to interpret. The coder translated the entry, probably correctly, as Presbyterian, scratched out and replaced by PCLP - meaning Presbyterian Church of the Lower Province, a denominational subheading used in the census abstracts of 1871. We had provided a valid code of PL for this subheading.

The error is indicated on the coding form and the record corrected.

Page 1 — Census of 1871. Province of Nova Scotia — District No. 196, West Halifax — Sub-District 1, Ward One — Michael Walsh, Enumerator.

Schedule No. 1—Nominal Return of the Living.

3	4	5	6	7 Names	8	9 Age	10	11	12 Religion	13 Origin	14 Profession, Occupation or Trade	15
1	1	1	1	Tobin Thomas S.	M	65	—	N. S.	Catholic	Irish	Private Means	M
1	1			" Frances Agnes	F	63	—	Ireland	"	"		M
1	1			" Michl Stephen	M	36	—	N. S.	"	"	[illegible]	M
1	1			" Catherine	F	28	—	N. S.	"	Scotch	—	M
1	1			" Thos Finlayson	M	1½	—	N. S.	"	Irish	—	—
1	1			Coy Anne Elizabeth	F	20	—	Newfoundland	"	English	Servant	W
1	1			Stuart Mary	F	29	—	Ireland	"	Irish	Servant	—
1	1			MacNeill Margaret	F	37	—	N. S.	"	[illegible]	Servant	—
1	1			Dircen Mary	F	28	—	N. S.	"	Irish	Servant	—
1	1	2	2	Lowe William	M	40	—	England	W. Methodist	Scotch	Shopkeeper	M
1	1			" Sarah	F	33	—	"	"	English	—	M
1	1			" Elizabeth	F	9	—	N. S.	"			—
1	1			" Amelia	F	5	—	"	"			—
1	1			McKe Lizzie	F	21	—	N. S.	Presbyterian	Scotch	Servant	—
1	1	3	3	Nolan [illegible]	M	57	—	Ireland	Catholic	Irish	[illegible]	M
1	1			" Mary	F	[illegible]	—					M
1	1			" Catherine	F	23	—					—
1	1			" Mary	F	15	—					—
1	1	4	4	McLeod [illegible]	M	30	—	N. S.	C. of E. [illegible]	Scotch	[illegible]	M
				" [illegible]	F	23	—	Scotland				
				" [illegible]	M	21	—	[illegible]			[illegible]	
						11						
					9	75						

```
196 A1  102  12  1  1   SAWYER          JOHN JAMES      M 062 US  CE  EN   RETIR GENTLEMAN  W
196 A1  102  12  1  2                   MARY            F 028 NS
196 A1  102  12  1  3                   FRANCES         F 026
196 A1  102  12  1  4                   ALICE           F 025
196 A1  102  12  1  5                   EMILY           F 024
196 A1  102  12  1  6                   ARTHUR          M 022                BANK CLERK
196 A1  102  12  1  7   JONES           CATHERINE       F 030                SERVANT
196 A1  102  12  1  8   LCNNERGAN        ELLEN            F02 7I  RR  CI    RSERVANT
196 A1  102  12  1  9   PUBLICOVER      JULIA           F 035 NS             SERVANT         W

:***AGE      IN ERROR FOR PERSON 196A 1  102  12   101  8          3      0     0     0
:***PL BIRTH IN ERROR FOR PERSON 196A 1  102  12   101  8       1225      0     0     0        I
:***RELIGION IN ERROR FOR PERSON 196A 1  102  12   101  8        664      0     0     0        Q
:***NATION   IN ERROR FOR PERSON 196A 1  102  12   101  8         85      0     0     0
:***SEX      IN ERROR FOR PERSON 196A 1  102  12   101  8          0      0     0     0
```

This page of the printout also illustrates an error message arising from mispunching. The messages given are AGE IN ERROR, PL BIRTH IN ERROR, RELIGION IN ERROR, NATION IN ERROR, SEX IN ERROR, all referring to person number 8, Ellen Lonnergan. A comparison of the printout with the coding form shows that the entries for all these variables have been punched slightly to the right of the correct columns.

E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
1	0101	002	1	01	LONE	WILLIAM P	M	040	EN	WMS	EC	SHOPKEEPER	M						
				02		SARAH	F	033			EN		M						
				03		ELIZABETH	F	009	NS					1					
				04		AMELIA	F	005						1					
				05	MCFIE	LIZZIE	F	021	PL	RC	SC	SERVANT							
1	0102	012	1	01	SAWYER	JOHN JAMES	M	062	US	CE	EN	RETIRED GENTLEMAN	W						
				02		MARY	F	028	NS										
				03		FRANCES	F	026											
				04		ALICE	F	025											
				05		EMILY	F	024											
				06		ARTHUR	M	022				BANK CLERK							
				07	JONES	CATHERINE	F	030				SERVANT							
				08	LONNERGAN	ELLEN	F	027	IR	RC	IR	SERVANT							
				09	PUBLICOVER	JULIA	F	036	NB			SERVANT	W						

CODER'S INITIALS _____ DATE __/__/__ PAGE NUMBER 07

63

```
197 G1  1001 223   1  1     GRAHAM          JOSEPH            M 050 NS  CE  IR                    * M
197 G1  1001 223   1  2                     CATHERINE JANE    F 037         SC                      M
197 G1  1001 223   1  3                     ALEXANDRE R       M 019         IR                    *
197 G1  1001 223   1  4                     EMMA              F 012                                     1
197 G1  1001 223   1  5                     LOIOIA SUSANNA    F 008                                     1
197 G1  1001 223   1  5                     JOHN WILLIAM      M 005                                     1

****CANNOT SUBSTITUTE LONG FORM DATA  6 FOR PERSON 197G 1 1001 223   101   1001  NO SUCH CASE
****CANNOT SUBSTITUTE LONG FORM DATA  6 FOR PERSON 197G 1 1001 223   101   3003  NO SUCH CASE
****PERSON # IN ERROR FOR PERSON 197G 1 1001 223   101  6          5     6   101     1
```

Illustrates two error messages, listed below the case. The first indicates that the program CANNOT SUBSITUTE LONG FORM DATA for persons number 1 and 3, Joseph and Alexandre Graham. Both of these persons have a "*" in column 73 of the coding form and punch card, indicating that their occupations were entered on Long Name Forms. The program's initial search for a long name card, matching the identification of these two, was unsuccessful. The appropriate forms were located and cards punched.

The second message, PERSON # IN ERROR, for person number 6, John William Graham, points to a punching error - two persons are given as number 5.

Subdistrict	E.A.	Special Sample Number	Household Number	Family	Person Number	Last Name	First Name(s) and Initials	Sex	Age	Birth Place	Religion	Nation of Origin	Occupation	Married	School	Read	Write	Deaf-Dumb	Blind	Unsound Mind
1	2	3–6	10–12	13	14–15	16–21	22–37	38	39–41	42–43	44–45	46–47	48–73	74	75	76	77	78	79	80
G	1	1,0,0,1	223	1	01	GRAHAM	JOSEPH	M	050	NS	C	E	IR	*	M		1			
					02		CATHERINE JANE	F	037				SC		M					
					03		ALEXANDRE R	M	019				IR	*						
					04		EMMA	F	012						1					
					05		LOUDIN SUSANNA	F	008						1					
					06		JOHN WILLIAM	M	005						1					
AV		1073	008	1	01	BAYERS	JAMES	M	037	NS	P	L	GE FISHERMAN		M		1			
					02		JANE	F	037						1					
					03		JAMISON	M	012											
					04		BENGAMIN	M	007											
					05		CAROLINE	F	003											
					06		GIBSON	M	004											

CODER'S INITIALS V.M. DATE __/__/__ PAGE NUMBER 09

Appendix E: Part 3

Canadian Historical Mobility Project: Numeric and Mnemonic Codes for Place of Birth, Nation of Origin, Religion and Occupation, Census Data, 1861 and 1871

The following coding schemes were employed in coding the national sample from 1871 census manuscript data on microfilm (and the sample of census manuscript data for Essex and Kent Counties, Ontario in 1851, 1861 and 1871). The mnemonic codes for place of birth, nation of origin and religion were used in transcribing the data from microfilm to coding forms. Using mnemonic rather than numeric codes at this stage was intended to reduce coding error. Numeric codes are employed on the SPSS file.

The codes include all mentions of place of birth, nation of origin, religion and occupation for all members of all households in the 1871 national sample and the 1861 and 1871 samples for Essex and Kent Counties, Ontario.

For the place of birth, nation of origin and religion codes, initial lists of mnemonic codes were taken from the census abstracts and provided for coders in the coding instructions. Provision was made for coding all other mentions in the course of coding (see p. 11 above in reference to "long name" coding and coding instructions). Subsequently, the full codes were constructed. The occupational coding is clearly the most complex and conceptually difficult. The religion codes were also given a general conceptual ordering in terms of church-sect status (for details, refer to the specific descriptions given below).

The occupational code dictionary consists of all occupational mentions in the three samples (currently excluding 1851 data for Essex and Kent) and corresponding numerical codes. Given that occupation as reported in the

census is a critical variable in this study, as in most current quantitative historical analyses, we have created a quite complex multiple code. We have been informed by previous coding schemes, including Armstrong's work for British occupational-industry codes, based on Booth's early work, the work of the historians Hershberg, Katz, Blumin, Glasco and Griffin (The "5 Cities Study," Historical Methods Newsletter 7 [June 1973], and the Philadelphia Social History Project's very elaborate coding scheme. The full description of the latter given in the Historical Methods Newsletter 9 (nos. 2 and 3, March-June 1976) was not available to us at the time our coding scheme was created. The logic of the two schemes is similar.

I. Four Digit and Two Character Mnemonic Codes for Place of Birth and Nation of Origin in the Canadian Census of 1871.

General Coding Scheme. 1977. Revised and Expanded. See variable list.

First Digit	Group	Later Digits	Mnemonic Code	Description
0	Missing & Illegible	000	NG	Not given
		100	IL	Illegible
1	Upper Canada	000	UC	Province as a whole
	Ontario	000	ON	
		xyz		District level code from 1871, xyz is the sequential code from 001 to 090 from the 1871 census.
2	Lower Canada	000	LC	Province as a whole
	Quebec	000	QU	
				District level code from 1871, xyz is the sequential code from 091 to 173 from the 1871 census.
3	All other Canada including Nfld, and P.E.I.	000	NB	New Brunswick as a whole
		0yz		District level code from 1871, yz are the last two digits of the sequential code between 74 and 87, from the 1871 census.
		100	NS	Nova Scotia (Terre Neuve) as a whole
		110		Cape Breton
		1yz		District level from yz are last two digits of the sequential code between 88 and 06, from the 1871 census
		200	PE	P.E.I.
		300	NE	Newfoundland
		400	BC	British Columbia
		500		Canada West
		600		Manitoba
		610		Red River
		700		Northwest
		710		Rupert's Land
3900	Canada--Canadian		CA	

First Digit	Group	Later Digits	Mnemonic Code	Description
4	United States--American	000	US	
5	France--French	000	FR	
6	United Kingdom & Ireland	000		Britain--British
		100	EN	England--English
		110		Great Britain
		200	WA	Wales--Welsh
		300	SC	Scotland--Scottish or Scotch
		400	IR	Ireland--Irish
		500		Guernsey
		510		Jersey
		520		Isle of Man
		530		Orkney
7	Other European, including Australia	000	AU	Austria
		100	GE	Germany
		110	BA	Bavaria
		120	PR	Prussia
		130		Bohemia
		200	BE	Belgium
		300		Scandinavia
		310	DE	Denmark
		320	NO	Norway
		330	SW	Sweden
		340		Greenland
		400	DU	Holland--Dutch
		500	GR	Greece
		600	IT	Italy
		610		Sicily
		700	PO	Poland
		800		Portugal
		900	RU	Russia
8	Other European including Australia (continued)	000	SP	Spain
		100	~~SW~~ SZ	Switzerland
		200		Australia ✓
		900	JU	Jewish
9	All other, non-European	000	NI	Native Indian
		100	HB	Half-breed
		200	AF	Africa
		210		Cape of Good Hope
		300	EI	East India
		310		Ceylon
		320		Malta

First Digit	Group	Later Digits	Mnemonic Code	Description
		400	HI	Hindoo or Hindu
		500		West Indies
		510		Trinidad
		520		Jamaica
		530		Bermuda
		540		Mexico

II. Three Digit and Two Character Mnemonic Codes for Religion in the Canadian Census of 1871

In general, an attempt was made to code religious affiliations according to their position on a dimension varying from established church to minority church to sect. The distinction varies as much with time as with religious affiliation; by 1671 many sects and minority churches were moving toward established status. Our primary sources of information have been:

S.D. Clark, Church and Sect in Canada (Toronto: University of Toronto Press, 1948) and David Millett, "The Age of Organized Religion," (unpublished manuscript, no date). We thank David Millett for making his manuscript available to us and Ted Mann for his comments.

The first digit of the code divides the religions into major groups. The second and third digits provide a detailed code for all mentions in the sample of religions which are known to be affiliated with the major church of column 1. The first digit is 9 for all "other" mentions, including those for which no known major affiliation was given. Codes in columns 2 and 3 are also loosely ordered in terms of increasing sectarianism, where this was given for the late nineteenth century in Millett. The size of the recorded congregation was used as a surrogate for this information in some cases - smaller congregations were assumed to be more sectarian. Many religious mentions however were not clearly classifiable and they are listed after the known ones in alphabetic order.

All religious affiliations-mentions are coded separately, unless they are clearly only different spellings. All spellings are given exactly as they were transcribed from the census manuscripts. Mnemonic codes are indicated in brackets for those codes used in the coding of the 1871 national sample of 10,000 households (and in the 1871 Essex and Kent County letter sample). All mentions of religions found in our coding of these 10,000 households are included in the overall list.

First Digit	Group	Later Digits	Mnemonic Codes	Description
0	Missing & Illegible	00	NG	Not given
		10	IL	Illegible
1	Catholic, Church of Rome	00	RC	
2	Church of England	00	CE	Also E. Church
		01		English
		02	EP	Episcopalian Episcopal "Church of --"
3	Church of Scotland	00	CS	
4	Lutheran	00	LU	
		10		Ev Lutheran Evangst Lutheran
5	Methodist	00	ME	
		01	WM	Wesleyan Methodist Wesleyan
		02	EM	Episcopal Methodist E Methodist E Meth Meth E
				From here, minority churches and sects in order of increasing sectarianism
		10	BE	British Episcopal Methodist
		11	NC	New Connexion
		12	PM	Primitive Methodist
		13	BC	Bible Christian Methodist
		14	BB	Bible Believer
		15	CM	Calvinistic Methodist
				From here, unclassifiable mentions
		30		Dutch Meth
		31		EGL Wesl
		32		Evangel Meth evangelical methodist evangelist M evangst meth
		33		I meth E
		34		I meth C
		35		J meth E
		36		Meth M E
		37		Methodist H
		38		Methodist N

NOTE: revised and expanded codes, see Variable list.

First Digit	Group	Later Digits	Mnemonic Codes	Description
6	Presbyterian	00	PS	Presbyterian
		From here, minority churches and sects in order of increasing sectarianism		
		01	CP	Canadian Presbyterian
				Canada Pres
				C Presb
		02		Free Kirk
				Free Church
				F C Presb
				F C Presbyterian
				F Churc Presb
				F Church
				FE Presb
				F Presb
				F Presbyterian
				Free Presb
				Free Presby
		03		K Presbyterian
				Kirk
				Kirk of Scotland
				Presbyterian Kirk
				Presb C S
				Presb C of Scotland
				S. Kirk
				S. Pres
				S. Presb
				Sco Presby
				Sco Presbyterian
				Scoth Presby
				Scotch Presbyter
				Scotch Presbyterian
		04		United Presbyterians
				United Presb
				Un presb
				U Presbyterian
				U Presb
				U P Presb
				U Kirk Presb
		05	AP	American Presbyterian
				U S Presb
		06		Reformed Presbyterian
				R Presb
		07		Evangelical Union

First Digit	Group	Later Digits	Mnemonic Codes	Description
6	Presbyterian (continued)	From here, unclassifiable mentions		
		30		ES Presb
				Est Pres
		31		Irish Pres
		32		N Presbyterian
		33		Old Presbyterian
				Old Presbyterian Kirk
		34		Presb N A
		35		W Presbyterian
7	Congregationalists	00	CO	
8	Baptist	00	BA	
		From here, minority churches and sects in order of increasing sectarianism		
		01	FW	Free Will
				F Baptist
				FWC Bapt
				FWC Baptist
				Free Christian
		02	RB	Reformed Bapt
				Reform Baptist
				RF Baptist
				R Bapt
				R Baptist
		03		Regular Baptist
				Regl Baptist
		04	UN	Union Baptist
		05	CC	Christian Conference Baptist
				Christian Conf
				Christian Bap
		06	AA	African Association Bapt
		From here, unclassifiable mentions		
		30		Baptist Christian
		31		C Baptist
				C Bapt
		32		CM Bapt
				CM Baptist
		33		Cal Bap
				Cal Bapt
				Cal-st Baptist
				Calvin Baptist
		34		Close Com Baptist
		35		First Baptist

First Digit	Group	Later Digits	Mnemonic Codes	Description
8	Baptist (continued)	36		Lu C Baptist
		37		N Bapt
		38		Open C Baptist
		39		Second Advent Baptist
9	Other	00		Amish Omish
		01	AD	Adventists
		02	SD	Seven Day Adventists
		03		Apostotic (sic)
		05		Bethern
		06	CB	Christian Brethern
		07	PB	Plymouth Brethern
		08	UB	United Brethern
		10		Christian
		11		Church of Christ C of Christ
		12		Christian Delp
		13		Church of God
		15		Dain Ward
		16	DI	Disciple Disciple of Christ
		17		Dunkers
		20	EV	Evangelical
		21		Evangelist
		22	EA	Evangelical Assoc
		24		German Episcopal
		25		Independent
		26	IR	Irvingites
		27	LD	Latter Day Saints
		28	MS	Messiah
		30	MN	Mennonites
		31	MO	Mormon
		32		Mnece Munice
		33		NSB Assoc
		34	SW	New Jerusalem C New Jerusalem New Jerusalem Ch New Jerusalem Church New Jirusalem Swedenborgians

First Digit	Group	Later Digits	Mnemonic Codes	Description
9	Other (continued)	35	QU	Quaker
				Friend
				Friends
		36		Prot Cong Zion
				Protest Congr
				Protest Congrega
				Congt Protest
		37	PR	Protestant
		38	TU	Tunkers
		39	UN	Uniterian
		40	UV	Universalists
		50		Greek
				Greek Orthodox
		60		Mahometan
		70	JU	Jews
		71		Hebrew
				Hebrew Church
		72		Reformed Jew
		80	NR	No religion
				No Denomination
				No Sect
		81	AT	Atheist
		82		Free Thinker
				Free Thinker of England
		83		Materialist
		84	PA	Pagan
		85		Infidele
		90	DE	Deist
		91		Spirtulist
				Spirituecist

Occupational Dictionary: Eight Digit Codes for Occupation and Industry in the Canadian Census of 1871 and in the Censuses of Essex and Kent Counties, Ontario in 1861

The occupational dictionary for the pilot project classifies every occupation mentioned either in the 1871 main file or the Essex-Kent 1861 and 1871 files. It provides an eight digit code constructed as follows:

Cols. 1 - 2:	A Detailed Industrial Classification
Col. 3:	Occupational Class Position
Cols. 4 - 6:	Detailed Occupational Codes
Col. 7:	Vertical Status Code
Col. 8:	Degree of Difficulty

Cols. 1 - 2: A DETAILED INDUSTRIAL CLASSIFICATION

The *detailed industrial classification* (variable label INDUS) is adapted from Armstrong's more detailed occupational allocation for all occupations separately distinguished in the English 1861 national occupational census abstract. Armstrong employed Booth's occupational list for the majority of the mentions classified. This is, clearly, a "functional" classification in Katz's sense (1972). The following shows the actual INDUS codes:

1-6. PRIMARY SECTOR

1. Farming
2. Other Agriculture
3. Logging
4. Fishing
5. Hunting

8. MINING SECTOR (inc. quarrying and well drilling)

10. BUILDING SECTOR

20-39. MANUFACTURE SECTOR

20. Machinery and tools (makers)
21. Shipbuilding
22. Metal workers
23. Watches, instruments and toys
25. Earthenware, inc. brickmakers
26. Coals, gas, chemicals
30. Furs, leather, glue, tallow, etc.
31. *Wood workers, inc. furniture and paper*
32. Carriages and harness
33. Printing and bookbinding
35. Textiles
36. Dress and textile products

37. Food, inc. drink and tobacco
39. Unspecified

40-49. TRANSPORT SECTOR

40. Navigation
41. Warehouses and docks
42. Railways
43. Roads

50-59. DEALING SECTOR

50. Raw materials, inc. fuels
51. Textiles, inc. textile products
52. Food, tobacco and all spirits
53. Furniture, utensils, and stationary
54. Hotels and lodging and restaurants
55. Other dealers
59. Unspecified

60-69. BUSINESS, GOVERNMENT AND PROFESSIONAL SERVICE SECTOR

60. Banking, insurance, accountancy
61. Public administration, communication, army, navy, police and prisons
63. Law and Medicine
64. Art and amusement
65. Literature and science, inc. newspapers
66. Education
67. Religion
69. Unspecified

70. DOMESTIC AND PERSONAL SERVICE SECTOR

80. INDUSTRY NOT KNOWN (inc. labourers)

90-99. RESIDUAL

91. Property owning and independent (inc. gentlemen)
92. Students
93. Other

Col. 3: OCCUPATIONAL CLASS POSITION

Occupational class position is an attempt to provide a relatively detailed classification in terms of the most likely implications which occupational titles have in terms of four main criteria of social class as distinguished from status or prestige. The criteria used for determining occupational class position are:

(1) Is the "occupational title" in or out of the labour force? e.g., lawyer and law student; seamstress and mother.

(2) Titles which clearly imply property ownership are coded as Merchants, Manufacturers, Agents & Dealers (code=1---). We include farmers as a *separate, second category of property owner.* Clearly there is some unavoidable error in such a classification. Examples of property owners are, manufacturer, miller, hotel keeper, etc. Note that we specifically exclude from this code occupations which could be either small manufacturers-owners or skilled, artisanal "makers" employed by others. This is a large group of nineteenth-century occupational titles, many of which were likely used interchangeably when the position of a person slipped from one category of artisan to the other. We do distinguish all occupational mentions most likely subject to this petit bourgeois/artisan slippage in terms of a specific range of detailed codes (see below cols. 4-6). There are reasons to be able to examine this group separately in analysis given the ambiguity of the class position which specifically characterizes them.

Within the non-propertied occupation titles, we attempt to employ the following three criteria, (a) probable skill level of the occupation, (b) the nature of the work process implied, and (c) the level of authority, i.e., directing other's work as a primary aspect of the occupation. These lead to six separate occupational class codes. The actual codes and criteria are as follows:

- Professionals, Managerial and Supervisory Occupations (code=2---) includes occupations which imply specialized formal training of any kind (doctors, lawyers, career soldiers, for example), and which suggest self-employment. It also includes those who are not self-employed but who direct others' work as a primary aspect of the job, managers, foremen, superintendents, bailiffs, inspectors. The two types are distinguished by different detailed codes, cols. 4-6, see below.

- White Collar (code=3---) occupations entailing administrative, clerical and technical work as an employee, i.e., not implying direct control of others' work as a major aspect of the job.

- Artisanal (code=4---) occupational titles imply a high level of specialized skill and likely degree of autonomy in the work-process. Many of these occupations are further classified as ambiguously petit bourgeois/artisanal in the detailed codes, cols. 4-6 as mentioned. (Equivalent to "5 Cities Study," Historical Methods Newsletter, June 1973; Category III).

- Semi-Skilled and Unskilled (code=5---), with the exception of labourer as a specific occupational title. These "blue-collar" occupations imply little to moderate skill levels, tedious and physically demanding work-processes and minimum authority or autonomy in the work. E.g. of semi-skilled are barbers, drivers, lumbermen and shantymen; unskilled are messengers, operatives, miners. (Equivalent to "5 Cities Study," Category IV).

- Labourer (code=6---) is retained as a separate code on the grounds of the particular insecurity and hardship characterizing common labour in the nineteenth century.

- Servants (code=7---) also retained as a separate code due to the special interest in the class implications of servant employment.

- Farmers (code=8---) are also retained as a separate code due to their predominance in the nineteenth-century.

- Outside Labour Force (code=9---) denotes those occupation titles that could not be placed in the above schema (e.g. gentlemen, students).

NOTE: A small number of occupations (>00.1%) were Miscoded and could be treated as "illegible" or "missing."

Cols. 4 - 6: DETAILED OCCUPATIONAL CODES

Detailed occupational codes are individual codes for each separate mention of a different occupation. Col. 6 is reserved for apparently synonymous occupational titles in French or English, but misspellings (Farner for Farmer, _armer for Farmer, etc.) are given the same code as the correct spelling.

These codes incorporate several important, additional distinctions.

I - even numbers were used for English titles, odd for French, with one exception; for dealers (as coded in terms of industry) we applied a special convention to facilitate analysis of this category. The convention is, for all occupational titles with the word "dealer" itself in them the 3 digit detailed code ends with a 5, those with the title "merchant" end with a 3, those with "marchand" end with a 4. (Note this is the only inconsistency in the general rule even = English, odd = French).

The object is to be able readily to separate the specifically labelled dealers and merchants from other occupations. We assume this is the majority of dealing occupations, although obviously others are also dealers (e.g.,

miller).

II - to distinguish between professionals and managerial/supervisory occupations within category 3 of the occupational class code above; all professional occupations are given 3 digit, detailed codes 000 - 599; supervisory/managerial 600 - 799.

III - to distinguish those occupations which implied small property owners or ambiguous and variable petit bourgeois/artisanal occupations. The detailed 3 digit codes 800-999 were used in all cases where it was thought the occupational title was even possibly in this category. This coding should be considered only with other sources, such as city directories.

IV - for the general coding procedures see Detailed Coding procedures (below) which include several additional conventions to ensure logically ordered and readily recodable detailed codes for every distinct occupational mention.

Coding Procedures for Detailed Occupational Codes

a. Within each industry, distinct occupations are numbered using the round numbers, i.e., 000 for the first, 010 for the second, 020 for the third, etc. The first three corresponding French occupations are then coded 001, 011, 021, etc.

b. When there are very similar titles, e.g., shingle cutter, shingle maker and shingle weaver, they receive consecutive (even, because they are English) numbers, in this case, 320, 322, 324. Since there are no corresponding French occupations, the numbers 321, 323 and 325 are not used.

c. The last digits 8 and 9 are used only for apprentices in English and French. The listings for carpenters are as follows:

Carpenter 280
Charpentier 281
Menuisier 283
House carpenter 284
Shop carpenter 286
App carpenter 288
App menuisier 289

d. When there is a difficulty in translation, as above, there being both "charpentiers" and menuisiers for carpenters, the titles are grouped together. Occasionally more than ten numbers are required to include a whole group of similar occupations, in which case the numbers over a sequence of 20 are used.

e. In general, within industrial code (col. 1-2) numerical gaps are left between dissimilar occupations and a few more detailed occupational groupings are provided for obvious differences, e.g., in Industry 63, Law and Medicine codes 000 to 099 are used for legal occupations, 100 to 199 for medical occupations and "the" phrenologist has a number 200; in Industry 40, the Navigation industry, codes 000 to 099 are for sailors, and other ship workers, 100 to 199 are for sea captains, pilots and other "managerial"

occupations.

Col. 7: VERTICAL STATUS CODE

This code (variable label STATUS) is the more conventional *"vertical" status code*. We have adopted Michael Katz's code (for Hamilton, Ontario 1851-1861) directly which includes status groups ranked 1 to 5, with 6 reserved for "unclassifiable" (Katz classified all female occupations 6). We have used Katz's detailed occupational listing and follow it exactly. All occupational mentions in our data, not actually given in Katz's listing, are coded 9, unclassifiable mentions. *Note that this Katzian code places Farmers within the ranks of "White Collar" occupations.*

Col. 8: DEGREE OF DIFFICULTY

This column is reserved for a subjective assessment made by the coders of the *degree of difficulty* (variable label OCPROB) in making a specific allocation. It refers only to problems of legibility or deciphering the actual mention as given on the original microfilm manuscript data. We used direct transcriptions of these data, but where a letter could not be made out reasonably, a blank space was coded, e.g., _oale .

The codes are 0 = no problem
1 = some difficulty
2 = considerable difficulty

For those mentions for which reasonable guesses could be made, a classification was given, e.g., Coaler is coded 61420190
oale is coded 6142019**2**

commarchand (merchant's clerk) is coded 59340120
Canie marchand is coded 5934012**2**.

Note: There remains, of course, a truly ambiguous category of either undecipherable or uninterpretable occupational mentions.

The general convention was to minimize the ambiguous category by reasonable guesses. Of a total of over 1,000 occupational mentions in the three files fewer than 50 were counted as truly ambiguous.

COMMENTS:

All the occupational coding was carried out by the two principal investigators and by Bruce Bellingham, a research assistant, working on aspects of the social history of Essex and Kent counties.

The procedure first punched all occupational mentions on separate computer cards with an appropriate identification code. Each mention was then given the full occupational code. Then for each different occupational code a definition card was punched as:

Cols. 1-4 = CLAS
25-32 = The eight character occupational code
55-80 = occupational title

The final occupational dictionary consists of the definition cards with the (possibly several) distinct original mentions filed after them. A computer program written for the purpose searches the file for duplicated codes and misclassified cards.

OCCUPATIONAL CLASSIFICATION: 19th Century

1. Merchants, Manufacturers, Agents and Dealers (property-owners and dealers), OCCUP codes 1000 to 1999

2. Professionals, Managerial and Supervisory Occupations, OCCUP codes 2000 to 2999

3. White Collar, OCCUP codes 3000 to 3999

4. Artisanal (Five cities study, category III), OCCUP codes 4000 to 4999

5. Semi-Skilled and Unskilled (Five cities study, category IV), OCCUP codes 5000 to 5999

6. Labourer, OCCUP codes 6000 to 6999

7. Servants, OCCUP codes 7000 to 7999

8. Farmers (also property-owning), OCCUP codes 8000 to 8999

9. Outside Labour Force (e.g. gentlemen, students), OCCUP codes 9000 to 9999

0. Occupation Blank, OCCUP code 0

   Miscoded, OCCUP code 1 to 999

The Design of the Sample from the 1871 Canadian Census

The 1871 sample combines two separate samples: a stratified sample of all households in all four provinces of Canada; and a "two-stage" sample of households that included at least one person of a particular national origin in a particular province. The two are referred to, respectively, as the "main" and "special" samples and are described in turn.

The Main Sample

This a stratified sample of all **households** from the entire census. The population is stratified by province and, within provinces, between urban (defined as communities of 3,000 or more) and non-urban areas. So there are eight strata (four provinces by urban/non-urban). Table 1 gives the population and number of households in each stratum, obtained from Volume 1 of the published Census. Table 4 (photocopied from the original documentation) gives the places identified as "urban" in the sample design. In Table 1, note that the estimated population is slightly different from the population recorded in the 1871 Census volumes, for example the estimated population for non-urban Ontario (in the first row) is 240,483, versus the published figure of 243,568. This discrepancy arises from our having sampled and the sample weights therefore involve post-stratification to remove these small errors. This assumes, of course, that the published Census is an exactly correct count of the census returns.

In order to increase the precision of comparisons between the two Atlantic provinces and Ontario and Quebec, the former were sampled with higher probability. Also urban areas (which included a minority of the population in 1871) were sampled with higher probability. Table 3, which is appended, is from the original sample design report and describes the urban sample in detail.

Because of the unequal probabilities of selection, weights are required to obtain unbiased estimates of population characteristics. Different weights are required to make optimal use of the data for estimates of characteristics of the entire population, for each province separately, for urban and rural areas, and for the combination of urban and rural areas. Table 1 also shows the value of the variable PROVURB in the dataset, which serves to identify the eight strata

Reflecting the paired selection of households, a second factor figures into the weights. The actual sample was drawn by dividing the population, within each of the 206 districts, into "cells" consisting of sequences of consecutive households in the census microfilms, that is the districts were put in order, then subdistricts, divisions and households. From each cell, two selections were made for the main sample. There were a small number of errors in the selection of observations in cells--about 40 errors in 10,000 households; in which one or three cases (not two) is in the sample. These errors are corrected by changes in the weights (in one-case cells, the weight was doubled; in three-case cells, weights were multiplied by a factor of two-thirds).

There is a complication in considering the data as a sample of individuals, rather than a sample of households. The dataset includes every person in every selected household; in other words, for persons in selected households the probability of selection is one. So, the weights for the household sample can simply be applied to each individuals in a household. The attractiveness of the stratified sample of *households* is that errors derived from it are no larger than would be obtained from a simple random sample; in more technical terms, the "design effect" is not larger than one. The same is not true, however, of the resulting sample of persons in households, which is a cluster sample, so that the given weights may result in standard errors (as given by SAS, SPSS and other programmes that assume simple random sampling) that **underestimate** the true values.

Table 1

Province	Urban-ization	Value of the Variable Provurb	Population, in the Published Census	Number of Selections*	Mean Number of Cases in the "Cell"	Population, estimated from the sample	Post-Strat-ification Correction
Ontario	Non-urban	10	243,568	1750	274.84	240,483	139.182
	Urban	11	43,450	1170	74.75	43,728	37.137
Quebec	Non-urban	20	151,395	1110	269.82	149,749	136.392
	Urban	21	29,220	792	74.52	29,512	36.893
New Brunswick	Non-urban	30	37,348	958	71.12	34,021	38.985
	Urban	31	6,231	286	41.90	5,992	21.787
Nova Scotia	Non-urban	40	53,415	1086	97.94	53,185	49.185
	Urban	41	9,086	360	50.02	9,013	25.239
Total			573,713	7512			

* corrected for a small number of cases where one or three, rather than 2 selections were made in a cell

Appropriate Weights, for Different Analytical Goals

In order to obtain national estimates which yield population counts (i.e. one obtains estimates of the actual numbers in the population), use the weight TOTWT. (popwgt)

In order to obtain national estimates which yields number of observations close to those in the sample (i.e. crosstabulations and other tables reflect the sample size, approximately), use the weight NATWT. (sampwgt)

popwgt sampwgt
TOTWT and NATWT can be used for all kinds of analysis and, because the sample is large are generally sufficient. They do not, however, maximize one's ability to detect differences between provinces and between urban and non-urban areas, ~~this~~ because they do not take account of the higher sampling ratios in urban areas and in the two Maritime provinces. The next three weights are designed to make use of this property of the sample.

In order to make comparisons between provinces (providing numbers of observations approximately equal to the actual sample sizes), use the weight PROVWT. (provwgt)

In order to make comparisons between urban and non-urban areas (providing numbers of observations approximately equal to the actual sample sizes) use the weight URBWT. (urbwgt)

In order to make comparisons between urban and non-urban areas within provinces (providing numbers of observations approximately equal to the actual sample sizes) use the weight PRURBWT. (prurbwgt)

The Special Samples

There are three components, separately derived, in the special sample:

- *a sample of German households in all four provinces*, from all districts in the four provinces with at least 15 percent German origin population;
- *a sample of French households in Ontario and New Brunswick*, selected from districts in Ontario and New Brunswick in which at least 15 percent of the population were of French ethnic origin; and
- *a sample of non-French households in Quebec*, from districts in Quebec with at least 15 percent British (combining English, Irish, Scottish and Channel Islanders.

These samples are designed to allow particular, theoretically interesting comparisons, for example between the "charter" ethnic groups and the Germans (which in 1871 constituted the only non-French, non-British group of any size). Tables 4, 5 and 6, which are appended, are from the original sample design report and describe the special samples in more detail.

In order to cut down the cost, the special samples were restricted to districts which, the published census volumes showed, included enough of the group for the effort to yield a sizeable number of cases; in practice districts with at least 15 percent of the desired group were included in special samples. Because they have large non-British populations (which would be coded in the main sample) and were in the urban-sample (with a higher sampling fraction, see above), Montreal and Quebec City were excluded from the sample of the non-French in Quebec. From the additional random samples of households in these districts, we coded households *with at least one person in the target group*. For example, a household with one person of German original and five of French origin would qualify for inclusion in the German special sample if it was selected.

(ETHWGT)

In order to use the special samples, one should employ the weight ETHWT *and always analyze the sectors of the population divided into categories of the variable* ***ETHCOMP***. The special samples, it should be (ETHSEL) emphasized, are not representative of the *entire* groups from which they are drawn. The German households, for example, are from areas in which Germans are relatively numerous; these may or may not be the same as German households in more ethnically-isolated circumstances.

Table 2

Group	Number of Households		
	Population	First-stage sample	Second-stage Sample
Districts in all four provinces at least 15 percent German origin	70462	6225	1483
Districts in Ontario with at least 15 percent French origin	12794	1308	417
Districts in New Brunswick with at least 15 percent French origin	12911	536	245
Districts of Quebec (excluding Montreal and Quebec City) at least Non-French origin	57868	1620	769

Some General Advice on Using Weights

Because the 1871 sample is fairly large, *unless small subsamples* are the object of analysis, statistical significance will rarely be at issue. That is, effects that are just large enough to be significant will generally correspond to small and uninteresting substantive differences. For this reason, the weights are calculated conservatively--so minimizing Type I error.

SPSS, SAS and most other statistical packages treat data as if they were derived from a simple random sample. While the samples are not actually simple random samples, *at t he household level* the stratified sample is at least as efficient as a simple random sample. The same is not true for the cluster sample of persons.

It is possible (but not simple) to make exact estimates of the standard errors of population (or sub-population) statistics, taking account of stratification and clustering. To do so you need to use the following variables: CELLNO, which is the number of the stratum for each household (these numbers begin with 1 are incremented, but may be restarted in each district and sometimes within districts) and NSEL, the number of main-sample household selections in the stratum. It would be helpful to begin by printing out selected variables for a few hundred observations so see the pattern of sample allocation.

To pursue these issues, a reasonable a knowledge of sampling theory is essential, and you should probably consult Michael Ornstein as well.

Table 3: Cities and Towns with a Population of 3000 or More in 1871, compiled from the Census of Canada 1870-71, Vol I, Table 1 (pp 2-83)

Province	Population over 5,000 Census District Reference	Name	Number of Households	Population 3,000-5,000 Census District Reference	Name	Number of Households
Ontario	2 g	Chatham	1137	1 k	Windsor	857
	10	London	2804	7 f	Stratroy	558
	15 d	Brantford	1513	13 f	Ingersoll	750
	21 b	St.Catherines	1437	14 e	Woodstock	759
	24	Hamilton	4830	23 c	Dundas	607
	33 c	Guelph	1223	25 g	Goderich	713
	46, 47	Toronto	9798	29 b	St. Mary's	578
	51 b	Port Hope	943	30 c	Stratford	777
	60 c	Belleville	1326	31 d	Galt	714
	66	Kingston	2229	37 g	Owen Sound	625
	68	Brockville	1856	42 e	Barrie	599
	77	Ottawa	3729	48 e	Oshawa	616
				50 b	Bowmanville	587
		Total	32385	52 c	Lindsay	736
				54 b	Colburg	775
				56 c	Peterborough	814
					Total	11065
Quebec	104-106	Montreal	16134	102 c	Joliette	396
	120 f	Sorel	857	117 e	St. Jean	506
	131	Trois Rivieres	1049	121 h	St. Hyacinthe	571
	145-147	Quebec	7944	140 a	Sherbrooke	710
	153 a-c	Levis	1053			
					Total	2183
		Total	27037			
New Brunswick	174	St. John	3369	180 a	Woodstock	682
	179 a-e	Fredericton	917	183 b	Bathurst	663
				184 b	New Castle	600
		Total	4286			
					Total	1945
Nova Scotia	196 a-g	Halifax	3989	192 c	Yarmouth	913
				195 f	Lunenburg	470
		Total	3989	197 a	Dartmouth	690
				198 a	Amherst	614
				199 k	Truro	690
				200 f	Picton	507
				201 c	Antigonish	522
				205 b	Sydney Mines	691
					Total	5097
Province Totals		Ontario	32385			11065
		Quebec	27037			2183
		New Brunswick	4286			1945
		Nova Scotia	3989			5097
		Total	67697			20290
		Grand Total	87987			

Table 4: Districts with 15% or More German, as compiled from the Census of Canada 1870-71, Vol I, Table I (pp. 2-83), and Table III Nation of Origin (pp. 252-333)

	Census District No.	District Name	Number of Germans	Total Population	Proportion of Germans	Total Households	Sampling Fraction	No. Hhlds Selected in First Stage	No. Hhlds Selected in Second Stage
Ontario	5	Elgin W	1138	12796	.089	2296	.1695	392	36
	6	Elgin E	3512	20870	.168	4024	.1695	686	93
	11	Norfolk S	2843	15370	.185	2860	.1695	480	68
	12	Norfolk N	2541	15390	.165	2837	.1695	480	81
	17	Haldimand	3357	20091	.167	3510	.1695	595	107
	18	Monck	5628	15130	.372	2903	.1020	296	98
	19	Welland	5916	20572	.288	3856	.1058	408	134
	21	Lincoln	4844	20672	.234	3795	.1267	481	127
	22	Wentworth S	3957	14638	.270	2629	.1695	450	114
	27	Bruce	5525	31332	.176	5270	.0453	247	39
	30	Perth N	5543	25377	.218	4355	.0453	202	43
	31	Waterloo S	8892	20995	.424	3391	.0453	170	71
	32	Waterloo N	13158	19256	.684	3222	.0453	144	96
	59	Prince Edward	4866	20336	.239	3780	.0453	182	49
	60	Hastings W.	2764	14365	.192	2616	.0453	132	25
	63	Lennox	4649	16396	.284	2983	.0453	143	39
	64	Addington	5453	21312	.256	3681	.0453	169	44
	71	Dundas	5563	18777	.296	3139	.0453	143	42
	72	Stormont	2220	11873	.187	1936	.0453	91	21
	83	Nipissing	266	943	.282	225	.0453	26	8
		Total	92635	356491	.2599	63508	.09317	5917	1335
Nova Scotia	195	Lunenburg	16612	23834	.697	3681	.0453	180	120
	197	Halifax	3425	19955	.172	3273	.0453	158	28
		Total	20037	43789	..4576	6954		338	148
		Grand Total	112672	400280	.28148	70462		6255	1483

*In the District of Monck, Subdistricts 18 a-d; the District of Welland, Subdistricts 19 a-f; the District of Lincoln, Subdistricts 21a and 21 c-f the sampling fraction was .1695. In the remaining subdistricts, 18 c-g, 19 g-l, and 21 b, the sampling fraction was .0453.

Table 5: Districts in Ontario and New Brunswick with 15% or more French Nation of Origin, as compiled from the Census of Canada 1870-71, Vol I, Table I (pp. 2-83) and Table II (pp. 252-333)

	Census District No.	District Name	Number of French	Total Population	Proportion of French	Total Households	Sampling Fraction	No. of Hhlds Selected in First Stage	No. of Hhlds Selected in Second Stage
Ontario	1	Essex	10539	32697	.322	6036	.10627	606	166
	75	Prescott	9623	17647	.545	2779	.10627	300	135
	76	Russell	5600	18344	.305	2964	.10627	319	98
	83	Nipissing, S	151	943	.160	225	.10627	16	0
	84	Nipissing, N	207	848	.183	153	.10627	13	5
	88	Algoma, E	255	977	.261	219	.10627	10	3
	89	Algoma, C	536	2177	.246	418	.10627	44	10
		Total	26911	73633	.36547	12794	.10224	1308	417
New Brunswick	181	Victoria	7184	11641	.617	1788	.036232	75	40
	182	Restigouche	1143	5575	.205	876	.036232	36	9
	183	Gloucester	12680	18810	.674	2564	.036232	113	73
	185	Kent	9356	19101	.560	2917	.036232	120	68
	186	Westmoreland	1071	29335	.460	4766	.036232	192	55
		Total	41064	84462	.42819	12911	.041515	536	245

Table 6: Districts of Quebec with 15% or More non-French Nation of Origin as compiled from the Census of Canada, Vol I, Table I (pp. 2-83) and Table II (pp. 252-333)

Census District No.	District Name	Number of French	Total Population	Proportion of French	Total Households	Sampling Fraction	No. of Hhlds Selected in First Stage	No. of Hhlds Selected in Second Stage
91	Pontiac S	3195	14,591	.219	2319	.04762	112	87
92	Pontiac N	260	1,219	.213	207	.04762	9	5
93	Ottawa W	11531	23,794	.485	3895	.04762	182	92
94	Ottawa C	2929	5,282	.555	825	.01988	18	9
95	Ottawa E	7054	9,553	.738	1499	.01988	30	6
96	Argenteuil	3902	12,806	.305	2109	.01988	46	32
101	Montcalm	10794	12,742	.847	2073	.01988	40	4
107	Hochelaga	20224	25,640	.789	3680	.01988	78	15
112	Chateauguay	11288	16,166	.698	2602	.01988	54	21
113	Huntingdon E	2383	8,834	.300	1493	.01988	30	22
114	Huntingdon W	2541	7,470	.340	1167	.01988	24	20
117	St. Jean	9415	12,122	.777	1948	.01988	44	15
125	Missisquoi	7114	16,922	.420	3022	.01988	60	36
126	Brome	3471	13,757	.252	2448	.01988	54	39
127	Shefford	12683	19,077	.665	3363	.01988	72	34
136	Drummond	10487	14,281	.734	2339	.01988	45	14
138	Richmond	3716	11,213	.332	1850	.01988	35	23
140	Sherbrooke	3544	8,516	.416	1388	.04762	40	43
141	Stanstead	3212	13,138	.245	2555	.04762	126	92
142	Compton	3785	13,665	.277	2376	.01988	45	33
144	Comté de Québec	14681	19,607	.749	3091	.01988	66	13
155	Lotbiniere	17340	20,606	.841	3129	.04762	154	26
156	Megantic	12074	18,879	.640	2827	.04762	140	55
159	Dorchester	7872	9,564	.823	1446	.01988	30	4
169	Bonaventure	9545	15,923	.599	2369	.01988	48	17
171	Gaspé C	2396	5,278	.454	843	.01988	18	6
172	Gaspé S	4897	7,296	.671	1005	.01988	20	6
	Total	202335	357,931	.56529	57868		1620	769

1871 Data File, Added Documentation – weighted files (April, 2001).

The 1871 file has several variables for weighting cases, since the sample was designed with several forms of over sampling in order provide more optimal estimates of specific comparisons – between provinces and between urban areas, or a combination of these two. The over sampling and the appropriate weight variables are described in the sample section toward the end of the basic documentation, but the weight variables are **relabelled** in the current data file. The original and current weighted variable names are:

TOTWT = POPWGT - for population estimates.
NATWGT = SAMPWGT – for sample estimates.
PROVWT = PROVWGT – for comparisons of provinces.
URBWT = URBWGT- for comparisons of urban areas (see urban variable).
PRURBWT = PRURBWGT – for comparisons of the 8 rural/urban by province (2 X 4) sectors of the sample.
ETHWT = ETHWGT. – for comparisons of **selected** ethnic populations – see below.

In addition, the variable ETHWGT must only be used to analyze the population when comparisons are made among the very specific ethnic selections or categories defined by the variable ETHSEL. A frequency tabulation for ETHSEL **without any weight on**, will reveal the specific groups and their frequencies in the special samples (as described in "The design of the Sample" section in the original documentation).

There are several other considerations in using weighted samples in the SPSS version of the files. For unknown reasons, we have found that if a weight variable is applied for an analysis, even if the weight is taken off before one exits the file (and the file not saved as weighted), when the file is reopened, the total N may appear as something over 40,000 cases. This is IN ERROR. When the file is retrieved for analysis, the weight procedure in SPSS "Data/weight" must be reset to no weight (it may appear as such, but should be reset). We recommend that before tabulations are undertaken, one establish that the N for the sample will be either 62,281, when no weight is applied (the ful number fo records), or 24,741, the size for appropriate sample estimates.

Also note that SPSS will provide unbiased parameter estimates, but **NOT** correct **estimates of error** for samples weighted such as this one. SPSS makes error estimates assuming simple random samples of a size equal to the sum of the weighted cases, rather than taking into account of probabilities of selection of cases. Normally, in a sample the size of this national sample the question of statistical significance is not a key one, unless one is examining fairly small subgroups. Nevertheless, some statistical packages, notably STATA, do compute correct estimates of error for stratified, weighted and "clustered" samples. (As the original documentation indicates, this is an equivalent to a simple random sample of households, but it is a cluster sample of individuals within them).