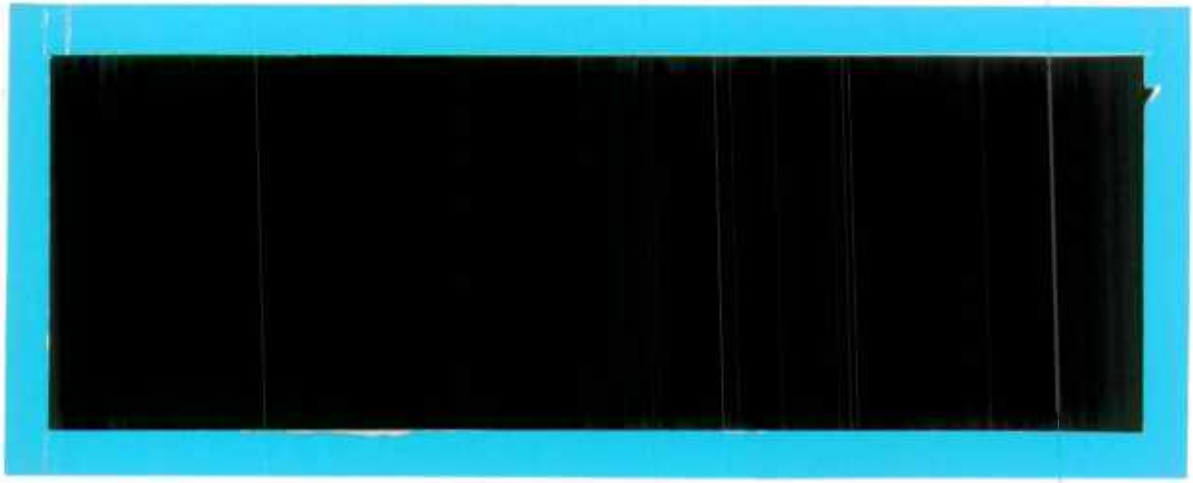

Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER NO. SSMD-88-008 E

METHODOLOGY BRANCH

METHODOLOGY FOR CONSTRUCTION OF ADDRESS REGISTERS USING SEVERAL ADMINISTRATIVE SOURCES

SSMD-88-008 E

J. Douglas Drew, John Armstrong, Alex van Baaren, Yves Deguire

Abstract

As part of the research program for the 1991 Census of Population, a study of the feasibility of constructing a dwelling address register for Canadian urban areas is underway at Statistics Canada. The initial pilot test indicated that use of an address register constructed using data from several administrative records systems could improve census coverage. In this paper the methodology used to construct address registers for additional pilot tests scheduled for the fall of 1987 is described. Topics examined include the quality of information available on the various administrative files, procedures used to parse free format address information and record linkage techniques used to unduplicate address lists. The benefits of using information not directly related to address in the linkage process are also considered.

KEY WORDS: Record linkage; Address standardization; Coverage Census.

Résumé

Dans le cadre du programme de recherche pour le recensement de la population et du logement de 1991, une étude de la faisabilité de la construction d'un registre d'adresses de logements pour les régions urbaines du Canada est en cours à Statistique Canada. Le test pilote initial montre que l'utilisation d'un registre d'adresses construit à partir des données provenant de plusieurs systèmes de dossiers administratifs pourrait améliorer le champ d'observation du recensement. Le document décrit la méthodologie utilisée pour construire les registres d'adresses pour d'autres essais pilotes prévus pour l'automne 1987. Les sujets examinés comprennent la qualité des renseignements fournis par les différents fichiers administratifs, les procédures utilisées pour la standardisation des adresses et les techniques de couplage des enregistrements utilisées pour le non-dédoublement des listes d'adresses. Le document examine également les avantages de l'utilisation des renseignements ne se rattachant pas directement aux adresses dans le processus de couplage.

MOTS CLÉS: Jumelage des enregistrements; standardisation des adresses; couverture; recensement.

1. INTRODUCTION

The notion of a machine readable household register that could be used in the conduct of population censuses is not new. Indeed, yesterday we heard from Redfern (1987) how not only household registers, but also population registers exist in Sweden, Denmark and some other European countries, and that the existence and use of these registers is in fact reshaping the role of Censuses in these countries. Also, the United States Bureau of the Census uses a list frame of addresses in the conduct of its decennial Census. Private vendor lists form the basis for their list, which is further improved by means of field checks (Whitford 1987).

In Canada, high quality vendor lists do not exist, and so we at Statistics Canada have considered at different times the creation of such a list ourselves. At this point I should note that currently in the Canadian Census, manual address lists are created by some 40,000 Census Representatives, each responsible for an area containing 200-300 dwellings. These lists are created coincident with the drop-off of Census questionnaires, and the address lists are not data captured.

The first study into the feasibility of creating a household or address register was carried out by Fellegi and Krotki (1967). They considered an approach of merging and unduplicating address information from multiple sources — which in their case consisted of the previous Census, municipal assessment roles, and electric utility billing lists. Pilot address registers were constructed and evaluated for two medium sized cities — Waterloo and London. They found the address registers covered 97% of dwellings, which was encouraging. However due to technological limitations of the day, construction of the address registers was largely a manual process, which did not favour implementation at that time.

During the 1970's a series of studies was undertaken which are summarized by Booth (1976). The approach considered was one of data capturing addresses from a previous Census, and using information from Canada Post to update the register and keep it current. The coverage under this approach was found to be comparable to that under traditional census methods. However the high initial data capture costs, despite anticipated savings in the longer term, were viewed as problematic, and the address register was not implemented.

Royce (1986) presented several potential uses and benefits of an address register to Statistics Canada programs, and also enumerated several factors that are now more conducive to construction of an address register than had been the case in earlier decades. These include the increased availability of machine readable administrative record systems with address information, the almost universal use of postal codes on these files, cheaper and more powerful computers, and improved record linkage methods and software. With all of these things in its favour, research into construction and use of an address register in the 1991 Census was started up a little over a year ago, considering an approach to construction like that investigated by Fellegi and Krotki, but with automation of virtually all of the steps. Due to lack of good address information in rural areas, attention is being restricted to urban areas.

Table 1 presents results from a small scale pilot register constructed for an area comprising 5000 dwellings in Ottawa (Drew, Armstrong, and Dibbs 1987). The address register coverage of valid dwellings was found to be about 1% below that of the 1986 Census for the test area. However, it was found that when the Census list and the address register were combined, the resultant list had 2.3% better coverage of dwellings than the Census. It should be noted that the areas chosen for this test were areas of suspected high undercoverage. Census dwelling undercoverage for the test areas was estimated at 3.7% after field verification, so that the 2.3% improvement in dwelling coverage obtained by combining the Census list and the address register represents about 60% of estimated

Census dwelling undercoverage. The dwelling overcoverage estimates for Census and address register lists needed to obtain the net dwelling figures involved in Table 1 were obtained using field verification.

Table 1
Ottawa Test: Net Dwellings as % of Census Net Dwellings*

	Dwelling Type		
	Single	Multiple	Total
Address Register	97.9	99.8	99.2
Census + Address Register	102.3	102.2	102.3

* Net Dwellings = Total Dwellings - Dwelling Overcoverage

Based on these encouraging results, a second test was scheduled; this test is currently underway in the field. In this test, two methods for use of an address register in Census data collection are being tested. Both methods are premised on the current drop-off methodology for delivery of questionnaires. The alternative of a mail-out Census based on an address register for 1991 was ruled out early in our research, when a study failed to show it would lead to any cost savings over the traditional methodology, under the assumption that a field check would be required to improve coverage prior to its use (Gamache-O'Leary, Nieman, Dibbs 1986).

Under the first method, which we call the pre-list method, address registers are pre-printed for each Census Enumeration Area. The task of the Census Representative under this method would be to update this list by making deletions and additions as necessary. The updating would be done coincident with the drop-off of Census questionnaires to all valid dwellings.

Under the second method, which we call the post-list method, Census Representatives (CR's) would create address lists from scratch coincident with drop-off of questionnaires, as under the current methodology. After drop-off, the CR would be issued a copy of the address register for his/her Enumeration Area, with instructions to match their manual list against the address register. Any additional dwellings found on the address register would be verified in the field and, if valid, would be added to the CR's list and a Census questionnaire dropped off.

The November 1987 test was restricted to a comparison of dwelling lists under the two methods relative to the traditional Census methodology, and for that reason it did not include any drop-off of questionnaires. For the test, persons with no previous interviewing experience were hired and assigned to a team doing only the pre-list method, or to a team doing only the post-list method. Under the test design, the same areas were listed according to both methods. Persons hired were not aware that two teams existed covering the same areas. As part of the evaluation, we will be data capturing the final address lists obtained under each method and carrying out a computerized match with resolution of any discrepancies through further field work.

This test is being conducted in the Census Metropolitan Areas (CMAs) of Vancouver, Edmonton, Toronto, Montreal and Halifax. For each CMA a stratified sample of 64 Enumeration Areas were chosen, with the stratification being on the basis of predominant dwelling type in the 1986 Census. This sample corresponded to areas of approximately 20,000 dwellings per CMA.

Table 2 presents the administrative files used as sources in constructing address registers for each CMA. Three national files already in Statistics Canada's possession were used for all cities — namely the Revenue Canada personal taxation file (TAX), and Health and Welfare Canada files of Family Allowance (FAM) and Old Age Security (OAS) recipients. In addition, for each site, two lists were purchased from among municipal assessment rolls (MUN), telephone billing lists (TEL) and electric utility billing lists (ELE). Edmonton was an exception in that due to a delay in obtaining one of the extra files, the address register was constructed using only four files.

Table 2
November 1987 Test: Source Files by CMA

CMA	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	x	x	x	x		x
Edmonton	x	x	x	x		
Toronto	x	x	x	x	x	
Montreal	x	x	x		x	x
Halifax	x	x	x	x	x	

2. STEPS IN ADDRESS REGISTER CONSTRUCTION

As mentioned earlier, the approach to address register construction we are investigating consists of merging and unduplicating address information from multiple administrative data sources. The four principal steps involved are discussed below.

Address Standardization

Address information on administrative files is typically in free format, by which we mean there is no fixed position or even order of appearance for the components of the address, such as street name, street number, apartment number, and so forth. It is necessary to analyse the address information to identify the components, in order that the address can be rewritten in a standard form to facilitate matching. This task turns out to be more complex than one might initially think.

At the outset of the address register research, evaluation studies of existing Statistics Canada software for address standardization revealed sufficient deficiencies that complete redevelopment was felt necessary to support an address register. An expert systems approach has been adopted which incorporates over 100 syntax rules concerning what constitutes a valid address (Deguire 1987). The system breaks the free format address into tokens, which are strings of consecutive letters or numbers, separated by blanks or delimiters such as commas. Some tokens are recognized by the system as keywords. Examples of keywords include 'Street', 'Rue', 'Apt', 'App' and so forth. Based on the pattern of numeric and alphabetic tokens, and known keywords, we have found that it is possible to uniquely decode over 95% of addresses unambiguously into components. While as few as 8 patterns account for 52% of addresses, the number of variations is large and over 1600 patterns are needed to handle 95% of the addresses. Currently the remaining 5% are reviewed and, where possible, deciphered manually. We are concerned about this 5% of cases, and plan to study whether further improvements can be made in the software, and in address register construction what would be the impact of discarding as opposed to attempting manual resolution of such cases.

Merging and Unduplication

After merging the standardized addresses from all the source files, the next step is to eliminate duplicates — that is, records referring to the same address. This is broken into two parts — exact matching to get rid of exact duplicates, and record linkage to identify duplicates where there is disagreement or only partial agreement on one or more of the standardized components. Such discrepancies occur for numerous reasons, such as variations in spelling, use of non-standard abbreviations, and so forth. The record linkage is carried out using Statistics Canada's record linkage software GIRLS (Hill and Pring-Mill 1985), which is based on the Fellegi and Sunter (1969) methodology.

More will be said in the next section about matching and record linkage in relation to construction of pilot registers for the November 1987 test.

Geographic Coding

Since we want ultimately to produce lists of addresses by Census Enumeration Area from the address register, the linkage of the address register to standard census geographic coding at least to the level of Enumeration Area is crucial. This linkage will bear directly on the coverage of the address register at the Enumeration Area level.

A number of possibilities exist for establishing this link and work needs to be done to evaluate them. One means would be through a Postal Code to Enumeration Area link. Such a link was established by data capturing Postal Codes for the one-fifth sample of dwellings in the 1986 Census, and plans exist for updating and maintaining that link. Plans also exist for evaluating the accuracy of this link, keeping in mind that to use it in linking an address register to Census Enumeration Areas would impose requirements for accuracy and updatedness well beyond what has been needed to support current uses.

Edit and Imputation

The final step in address register construction consists of fine tuning. For instance, logical gaps in apartment numbers can be imputed. Some clearly erroneous addresses which escaped detection at earlier steps in address register construction may be spotted clerically and deleted.

3. PRELIMINARY FINDINGS FROM CONSTRUCTION OF PILOT REGISTERS

In this section, we present some preliminary analysis of the address register construction process, based on the pilot registers for the November 1987 test. More critical and complete analysis will be possible when results of the current field work become available.

Table 3 presents the gross coverage of the pilot registers at various stages of construction as a percentage of 1986 Census dwellings for the test areas. Column (2) indicates the initial number of addresses with Postal Codes corresponding to those in the selected Enumeration Areas in each city according to the most recent version of the Postal Code to Enumeration Area conversion file, whose vintage was February 1987. That is, it represents the number of addresses after merging of standardized addresses from all source files and before elimination of duplicates. The four source files used in Edmonton contained, in total, twice as many addresses as the 1986 Census, while the five files used in other cities contained on average three times as many addresses as the Census.

After elimination of exact duplicates, the gross coverage (compared to the 1986 Census) was brought down from an average of 273% (column 1) to 122% (column 3); this demonstrates the success and importance of the address standardization step.

Table 3
Gross Coverage as % of 1986 Census Dwellings Pilot Address Register at Steps During Construction

CMA	After Merge	After Elimination of Exact Duplicates	After Postal code Verification	After Record Linkage	Final
(1)	(2)	(3)	(4)	(5)	(6)
Vancouver	283	117	109	103	104
Edmonton	194	110	103	99	101
Toronto	283	113	103	102	102
Montreal	312	136	125	111	108
Halifax	297	134	126	109	110
Average	273	122	113	105	105

Column (4) represents a step that was unique to construction of the pilot registers. Postal Codes of addresses were verified using Statistics Canada software designed for this purpose, and cases where Postal Codes were in error and the corrected Postal Codes fell outside the sample Enumeration Areas were dropped. Note that, in constructing a full scale address register, such cases rather than being dropped would be shifted to the Enumeration Area where they belong. This Postal Code verification step resulted in 9% of the records being dropped with, of course, none being added; it represents a potential source of undercoverage that would be unique to the pilot registers. We plan to assess the extent of such undercoverage, which may range from minimal to being fairly significant depending on the degree of independence of Postal Coding errors from file to file.

The record linkage step reduced the gross coverage by a further 8% (column 5), resulting in average gross coverage of 105%. Column (6) represents the gross coverage after edit and imputation. On average, gross coverage was unaffected, but for individual CMAs it increased or decreased by 1-2%, which is quite a large amount relative to the anticipated net undercoverage of the registers. If results are similar to those for the earlier pilot register for Ottawa, net undercoverage relative to the Census may be close to 1%, which, given the 5% gross overcoverage of the registers, would imply 6% net overcoverage. The overcoverage stems from duplicate records which were undetected in the record linkage process or from appearance on the register of dwellings which are no longer valid.

Results from the field test will tell us the under- and overcoverage not only for the address register, but for its alternative uses in Census data collection. We also plan to do indepth studies of reasons why addresses were missed on the address register, and whether improvements in the methods and software could reduce the undercoverage.

Table 4 presents some results on the record linkage step in address register construction. It presents for pairs matched during record linkage, the percentage of times individual components of the address used in linking either agreed, partially agreed, or disagreed. It should be noted that street number was a blocking factor in record linkage, that is searches for links took place only amongst records which agreed on street number.

Another point worth noting was that during the merge and exact matching, a record was kept of the source files on which each address appeared, and during record linkage it was the version appearing on the most source files that was retained. Two levels of partial agreement were allowed as comparison outcomes for street and municipality names. (These are combined in Table 4). The first level consisted of cases of minor misspellings due to omission of a letter or transposition of two letters. Two names were declared to agree at the second level of partial agreement if their phonetic versions coded using the NYSIIS (New York State Identification and Intelligence System) scheme were identical. NYSIIS coding is intended to eliminate the effects of common spelling errors.

Table 4
Comparison Rule Outcomes for Address Pairs Matched by Record Linkage (Percentages)

Matching Category	Outcomes			
	Agreement	Partial Agreement	Dis-Agreement	Missing
Street Name	49	31	20	
Apt. Number	93		7	
Civic Number Suffix	95		5	
Postal Code: Dig. 1-3	100			
Postal Code: Dig. 4-6	95	4	1	
Municipality	87	2	11	
Family Name	35		18	47

Another field where partial agreement was allowed was in the last three characters of the Postal Code, where two out of the last three characters being the same constituted partial agreement.

It is interesting to note the low frequency with which the street name agreed for matched records, with full agreement only half of the time. This appears to be due to frequent misspellings and abbreviations. Another point worth noting regards the use of family name as a match variable. This variable was used only for record linkage purposes, and was deleted from the final register. Due to the different ages of the source files, failure to link on family name was not counted against linking a pair of addresses; however, agreement on name was considered quite important, that is it received a high positive weight. In order to assess the impact of using family name, for one city we repeated the record linkage without name, and found that 1% less records were linked.

The next two tables examine the contributions of the various files to the final address register. Table 5 presents coverage of the source files as a percentage of address register gross coverage — that is, what percentage of the address register records were traceable back to each of the source files. This table confirms as we had suspected that coverage of the tax, telephone and electric utility files is high. The electric utility files came out best, and it appears, at least in the two provinces we have looked at, that bulk metering of multi-unit structures, which previously had been a weakness of this source is no longer a significant factor. The low tax file coverage in Montreal and Toronto was due to frequent errors in the tax file address leading to its not being the retained version. The coverage of the municipal assessment file, except for Toronto, was quite low since they generally have only one record per owner for multiple unit structures.

Table 5
Gross Coverage of Sources Files
(% of Address Register Gross Coverage)

City	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	73	26	26	48		87
Edmonton	82	32	18	49		
Toronto	60	22	18	78	76	
Montreal	57	24	16		72	86
Halifax	78	30	19	47	72	

Table 6 gives the percentage of addresses uniquely contributed by each source. Once again, electricity files performed very strongly, and the telephone files were not far behind. The tax files performed well in the case of Halifax and in Edmonton. The Edmonton result is anomalous in the sense that of the four files used, the tax file was the only one with high coverage of addresses.

Table 6
Unique Contribution by Source File
(% of Address Register Gross Coverage)

City	Source File					
	TAX	FAM	OAS	MUN	TEL	ELE
Vancouver	5	1	1	1		13
Edmonton	28	5	4	4		
Toronto	2	0.5	0.5	7	12	
Montreal	3	0.5	1		9	17
Halifax	10	1	1	2	9	

It should also be noted that these results are for the contribution of individual files to gross coverage. It will be of interest, once the field results from the November 1987 test are available, to see the contribution of each file to net coverage. The usefulness of files such as Family Allowance and Old Age Security would be very questionable if a substantial proportion of the 0.5-1% unique addresses they contribute are in fact found to be in error.

Future Directions

Analysis of the results from the November 1987 test will be completed by the spring of 1988. Also estimation of the developmental requirements, and cost and timing implications of different scenarios for use of an address register in the 1991 Census will be completed by that time. A decision on the extent of use of an address register in the 1991 Census, based on these two inputs, is scheduled for the spring of 1988. If a decision to use an address register on a wide scale is taken, this will imply a high priority to developmental work leading up to 1991.

The work to date has identified areas for further research, some of which would have to proceed in parallel with development if the decision taken is in favour of implementation. The research should continue also if it were decided to use an address register on a test as opposed to a production basis in 1991. Areas where further research is needed are discussed below.

Updating Methodology

To date we have only considered the initial creation of an address register. The sources and approaches that are best for creation are not necessarily the best for updating. Consideration has to be given both to the frequency with which updating is needed, and the implications on systems design of the frequency and proportion of updates. One possible approach to updating would be to do an exact match on successive versions of source files to identify changes, which would then be linked to the existing address register. The handling of deletions of addresses that are no longer valid under such an approach needs special investigation. Another possibility would be the use of data sources such as construction and demolition permits or updates from Canada Post.

Use of Address Register in Enumeration Area Delineation

For collection and dissemination purposes, Census Enumeration Areas should contain approximately the same number of dwellings, and they must respect higher level geo-statistical and geo-political boundaries. Since dwelling counts used in Enumeration Area delineation are currently primarily based on the previous Census, they are sometimes quite out of date. Dwelling counts from an intercensally updated address register should improve the delineation process, and reduce the expense and disruption of having to split Enumeration Areas due to discovery of substantial growth during field operations for the Census.

Use of the Address Register as a Frame for Household Surveys

Currently most household surveys at Statistics Canada are based on area samples, which require costly face to face interviewing, at least in the first month households are sampled. Telephone frames by themselves are not a viable alternative for large national surveys, due to the bias associated with undercoverage of the non-telephone universe (Drew and Jaworski 1986). The alternative of dual frame methodologies combining area frames and telephone frames is fairly inefficient in the sense that a relatively large area sample is needed to cover the small non-telephone population. An address register with telephone numbers for roughly 75% of urban households (see Table 5) has appeal as a frame which can afford the benefits of telephone interviews for a large portion of the urban population, while identifying and permitting the adoption of an efficient sample design for remaining urban and rural households.

Plans are to convert a portion of the Labour Force Survey sample to an address register based design in the areas where pilot registers are being maintained for use in Enumeration Area delineation. As part of the test, methods for dealing with address register undercoverage will be investigated.

Refinement of Address Register Methodology

Finally research is needed into the address register construction process itself. Issues such as the impact of more or different source files need study. Can additional sources with high coverage be found, and if so what would be the implications of their use in address register construction?

Also, we saw that the software for address standardization, and for validating Postal Codes does not successfully handle all cases. More needs to be known about the problem cases. Are they cases of address errors appearing on one file while valid versions of the same address appear on another file? If this were the case, ignoring problem cases on individual files might be the recommended course of action. If cases not handled by the software are due to systematic failure of the software to handle valid addresses, or if

particular addresses tend to be in error on all files, then ignoring these cases would lead to coverage problems. Detailed study of problem cases, including addresses missed on the pilot registers is needed to answer these questions.

In summary, the findings to date are encouraging, both in terms of the technical feasibility of producing at a reasonable cost an address register with high coverage of urban addresses, and in terms of the potential for such a register to reduce undercoverage in the Census. A number of avenues of further research into uses and improvements of the methodology for construction and updating are planned for the coming year, to be integrated with developmental work should the decision be taken to proceed with implementation of an address register for the 1991 Census.

REFERENCES

Booth, J.K. (1976). A Summary Report of All Address Register Studies to date. Internal Report, Statistics Canada.

Deguire, Y. (1987). Research into the Parsing and Standardization of Free Format Addresses at Statistics Canada. Internal Report, Statistics Canada.

Drew, J.D., Armstrong, J., and Dibbs, R. (1987). Research into a Register of Residential Addresses for Urban Areas of Canada. *Proceedings of American Statistical Association, Section on Survey Research Methods.*

Drew, J.D., and Jaworski, R. (1986). Telephone Survey Development on the Canadian Labour Force Survey, *Proceedings of the American Statistical Association, Section on Survey Research Methods.*

Fellegi, I.P. and Krotki, K.P. (1967). The testing programme for the 1971 Census in Canada. *Proceedings of the American Statistical Association, Social Statistics Section, 29-38.*

Fellegi, I.P., and Sunter, A.B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Gamache-O'Leary, V., Nieman, L., and Dibbs, R. (1987). Cost Implications of Mail-out of Census Questionnaires using an Address Register. Internal Report, Statistics Canada.

Hill, T., and Pring-Mill, F. (1985). Generalized Iterative Record Linkage System. *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.

Redfern, P. (1987). European Experience of Using Administrative Data for Censuses of Population: The Policy Issues that Must be Addressed. International Symposium on Statistical Uses of Administrative Data, Ottawa.

Royce, D. (1986). Address Register Research for the 1991 Census of Canada. Journal of Official Statistics, 2, 447-456.

Whitford, D. (1987). Research Program for the 1990 Decennial Census, *Proceedings of the American Statistical Association, Section on Survey Research Methods.*