Survey Methodology, June 1992
Vol. 18, No. 1, pp. 127–141
Statistics Canada

The Creation of a Residential Address Register for Coverage Improvement in the 1991 Canadian Census

L. SWAIN, J.D. DREW, B. LAFRANCE and K. LANCE[1]

ABSTRACT

The Address Register is a frame of residential addresses for medium and large urban centres covered by Geography Division's Area Master File (AMF) at Statistics Canada. For British Columbia, the Address Register was extended to include smaller urban population centres as well as some rural areas. The paper provides an historical overview of the project, its objective as a means of reducing undercoverage in the 1991 Census of Canada, its sources and product, the methodology required for its initial production, the proposed post-censal evaluation and prospects for the future.

KEY WORDS: Address Register; Census undercoverage; Geographical Information Systems (GIS).

1. OBJECTIVE

The concept of an Address Register at Statistics Canada dates back to the 1960s. Fellegi and Krótki (1967) first considered building one for the 1971 Census using administrative source files as the base. Their approach was mostly manual and yielded a very complete set of addresses with minimal undercoverage and overcoverage. In the mid-1970s (Booth 1976), the idea resurfaced in planning for the 1981 Census. This time the approach started with data capture of addresses from the previous Census and was augmented with information from Canada Post. In both cases, the generated address lists were being considered as a frame for a mail-out Census. However, costs of creation were high and would have needed offsetting reductions in other Census operations to be effective. In addition, the risks associated with changing the traditional enumeration method were considered too great. As a result, the construction of an Address Register was suspended in each case.

A renewed interest in the concept of an Address Register emerged from the International 1991 Census Planning Conference (Royce 1986, 1987) in October 1985. This interest derived from the potential for automation of Fellegi and Krótki's approach due to technological developments, such as the availability of machine readable administrative files with addresses and postal codes and the development of in-house software to parse addresses into standard components, to assign postal codes and to link postal codes to Census geography. It followed as well from the development of a statistical theory for record linkage (Fellegi and Sunter 1969) and computer systems based on this theory (Hill and Pring-Mill 1985).

As a result, a project was initiated in 1986 with the first research (Gamache-O'Leary *et al.* 1987) investigating the use of an Address Register for a mail-out Census rather than the traditional drop-off approach. It concluded that the new Census data collection approach would be less expensive only if the quality of the Address Register required minimal field updating prior to the Census. Two small pilot registers created in early 1987 put Address Register coverage at 90-95%, which was unacceptable without field updating (Drew *et al.* 1987), ruling out the use of an Address Register for a mail-out Census.

[1] L. Swain and B. Lafrance, Social Survey Methods Division; J.D. Drew, Household Surveys Division; K. Lance, Geography Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

However, the two pilot registers revealed the potential for an Address Register to aid in coverage improvement when used in conjunction with the traditional drop-off methodology. This fitted well with the emergence of coverage improvement as one of the top priorities for the 1991 Census. The results of the Reverse Record Check for the 1986 Census had indicated a dramatic rise in the undercoverage rate compared to previous Censuses (from 2.01% in 1981 to 3.21% in 1986 for the national total population; from 2.08% in 1981 to 3.28% in 1986 for the national urban population) (Statistics Canada 1990). It was therefore decided that *the research project should concentrate on the development of the Address Register to use in coverage improvement of the 1991 Census.*

The next section describes the two major tests conducted to develop and refine the procedures used to create the Address Register for the 1991 Census. As well, the second section outlines the joint agreement with the Province of British Columbia to extend the Address Register. The third section presents the administrative and geographic sources used in the production process and the structure and content of the Address Register booklets, the end product used by Census Representatives in the field. The fourth section describes the methodology used to exploit the sources in order to produce the Address Register booklets. In the fifth section, the proposed post-censal evaluation is discussed while the last section presents future prospects for the Address Register. A separate future report will detail an evaluation of the methodology.

2. BACKGROUND

2.1 The November 1987 Test of Coverage Improvement Methods

A substantial test of the use of the Address Register (AR) as a coverage improvement tool was conducted in November 1987 in five large Regional Office cities. It was designed to estimate both undercoverage and overcoverage of dwelling units for the traditional Census method of listing and for two experimental methods of using an AR for Census coverage improvement: Post-list and Pre-list. The Post-list approach had the enumerator compile the dwelling list in the usual Census manner (creating a Visitation Record) then reconcile it with a dwelling list for the Enumeration Area (EA) derived from the AR. Field follow-ups were done where necessary on any address discrepancies between lists. In the Pre-list method, the enumerator was given the AR in advance and updated it during a canvass of the EA to create the final dwelling list.

The results (van Baaren 1988) concluded that the Post-list method was the more effective in improving coverage. This approach as a simple add-on to the standard Census enumeration process was fail-safe. If for some reason we failed to produce the AR (either in whole or in part) on time for the 1991 Census, the AR reconciliation step could simply be dropped without affecting the traditional enumeration process. The test data also provided estimates of the degree of coverage improvement and costs (Royce and Drew 1988). It was estimated that 34,000 occupied dwellings and 68,000 persons would be added by the AR to the medium and large urban centres for which it would be constructed (these urban centres representing those areas for which an Area Master File exists, *i.e.*, covering about 65% of the Canadian population). This would represent an improvement in coverage of 0.26 percentage points (the national undercoverage rate in 1986 being estimated as 3.21 percent). Relative to the two previous attempts at AR construction, costs were demonstrated to be low to the Census due to the highly automated approach and the proven benefit. As well, the risk was minimized since the traditional data collection method would still be used. Based on this cost, benefit and risk assessment, approval was given for creation of an AR for the 1991 Census.

From the November 1987 test, two concerns presented themselves. First, the ordering of the addresses in the AR booklets produced for each Enumeration Area (EA) didn't correspond to the order in the Visitation Records which made reconciliation a tedious and time-consuming task. Second, the overall overcoverage at 17% still seemed too high and more effort was required to eliminate erroneously placed or duplicate records. Both these problems were addressed by improving the methods for matching the AR to Census geography. Instead of linking addresses merely to EAs as had been done for the November test, procedures were developed to match the AR to the Area Master File (AMF) (Statistics Canada 1988) blockfaces. An algorithm was developed to sort addresses by block and within block in the same order they would be encountered by the enumerator in walking around the EA.

2.2 The September 1989 Test to Refine Procedures

Another substantial test was conducted in September 1989 involving four cities of various sizes: Moncton, Laval, Brampton and Calgary. Each was chosen because of unique difficulties that could arise based on the November 1987 test. The results (Dick 1990) showed a significant decrease in coverage from 84% in the 1987 test to 73%, a discouraging outcome. On the other hand, this test revealed a considerable reduction in overcoverage down from 17% to 8%. Importantly, despite the reduced coverage of the AR, its performance as a coverage improvement tool for the Census was still viable. On analysis, the new geocoding operation was found to be problematic, both in terms of its high costs, since it involved a great deal of clerical intervention, and in terms of its quality. The geocoding steps were therefore revamped for production, a key aspect of which was the adoption of CANLINK record linkage software (Statistics Canada 1989b) to improve quality and reduce costs of the AR/AMF linkage.

2.3 Joint Agreement with the Province of British Columbia

The Ministry of Finance and Corporate Relations in British Columbia was concerned about the high rate of undercoverage in their province in the 1986 Census (4.49% in 1986, up from 3.16% in 1981, for the provincial total population) (Statistics Canada 1990). Statistics Canada entered into a joint agreement with the Planning and Statistics Division (the provincial statistical agency) of the Ministry to help reduce undercoverage in British Columbia in the 1991 Census. Within this contract, the Address Register was expanded to include smaller urban areas in British Columbia, thereby increasing the population covered from 62% to 88%.

3. SOURCES AND PRODUCT

Production started in April 1990 and ended with the final Address Register (AR) booklet stapled in mid-May 1991, when 22,756 booklets had been compiled containing 6.6 million addresses for use in the Census data collection process.

3.1 Administrative Sources

In the September 1989 test, it was concluded that wherever possible the following four administrative sources ought to be used as sources of addresses to create the AR: telephone company billing files, municipal assessment rolls, hydro company billing files and the T1 Personal Income Tax file. However, the use of all four sources was possible only in Nova Scotia, New Brunswick, and eight major urban centres in Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton and Windsor). Because of the multiplicity of files,

the cost of files and refusals, only three sources were used for Newfoundland, Québec, Manitoba, Alberta (telephone, hydro and tax files) and for Regina and the rest of Ontario (telephone, assessment and tax files). For Saskatoon, only telephone and tax files were available. The primary source files used by the British Columbia government were those of telephone and hydro, though motor vehicles, cable and Elections files were also used.

3.2 Geography Sources

In building the AR, extensive use was made of a Geography Division system and files.

i. The Area Master File (AMF) (Statistics Canada 1988) is a digitized feature network (covering streets, railroads, rivers, *etc.*) for medium and large urban areas, generally with populations of 50,000 or more. Of interest for the AR were the street features which contained street name and civic number ranges which could be used to locate individual addresses onto a blockface, the primary linkage.

ii. The Computer Assisted Mapping System (CAM) orders blockfaces into blocks and blocks into a Census Enumeration Area (EA). CAM was used for the sequencing of addresses in the AR booklets. The EA maps produced by CAM were used by the Census Representatives for the 1991 Census. For the AR, the maps for all AMF areas were used in the second clerical operation.

iii. The 1990 Postal Code Conversion File (PCCF) (Statistics Canada 1991) is a national file of all postal codes, each of which is linked to a 1986 Census EA or a series of 1986 EAs. This input was used for secondary linkage of addresses at the EA level.

iv. The 1986/1991 EA Correspondence File relates the 1986 EA geography to the 1991 geography. This file was used for the secondary linkage at the EA level and the second clerical operation.

3.3 Address Register Booklets

The end product consisted of a set of booklets of residential addresses, one for each Enumeration Area, covering urban areas of Canada for which an Area Master File existed. Figure 1 contains a fictitious example of a page from an AR booklet (reduced in size).

Each booklet was divided into two sections: a structured portion and an unstructured portion. The structured portion contained all the addresses tied to a blockface with all the blockfaces being sequenced into blocks within the EA. The sequencing mirrored that found on the map that the Census Representative (CR) used for listing the EA in his/her Visitation Record (VR). The unstructured portion contained the addresses that could be tied only to the EA rather than a blockface. These were sorted by odd/even civic numbers within street name. The volume of addresses split 90%-10% between structured and unstructured.

Besides the address data, each page in an AR booklet contained a series of columns to be used in the reconciliation operation between the AR and VR. In the reconciliation, the Census Representative manually compared the Visitation Record with the AR to identify matches and non-matches. If the address was only on the VR, it was added to the AR (undercoverage in the AR). If the address was only on the AR, field resolution was usually required by the CR, with the result that the address was designated either as a new address to be enumerated for the Census by the CR (undercoverage in the Census) or as an invalid address classified by type of error (overcoverage in the AR). Addresses were denoted as invalid if they were duplicates, if they lay outside the EA, or for any other reason. All valid addresses had the Census Household Number coded in the booklet by the CR. A telephone number for the address, if available,

ADDRESS REGISTER Protected PROVINCE 35 EA 261 Page 21 of 22
FED 038 VN 0

Block No.	Address			Hhld No.	Not Listed at Drop-off	Field Follow-up Required	Invalid			AR Ref No.	Telephone Number
	Civic No.	Street	Apt. No.				Duplicate	Outside EA	Other		
1	2	3	4	5	6	7	8	9	10	11	12
4	23	MAIN ST								1044566	5551111
4	19	MAIN ST								1044564	5561234
4	15	MAIN ST								1044562	5552321
4	11	MAIN ST								1044559	
4	9	MAIN ST								1044583	7475739
4	7	MAIN ST								1044581	5552222
5	30	CENTRE RD								1019615	5561029
5	34	CENTRE RD								1019617	
5	34	CENTRE RD	BT							1019618	5564261
5	60	CENTRE RD								1019627	
5	64	CENTRE RD								1019629	7478765
5	68	CENTRE RD								1019634	5556942
5	72	CENTRE RD								1019636	
5	76	CENTRE RD								1019640	
5	80	CENTRE RD								1019642	7476789
5	84	CENTRE RD								1019644	5568765
5	88	CENTRE RD								1019646	5559999
5	92	CENTRE RD								1019579	7473456
5	96	CENTRE RD								1019581	7450987
5	100	CENTRE RD								1019648	
5	108	CENTRE RD								1019579	5557171
5	112	CENTRE RD								1019581	5558888
5	116	CENTRE RD								1019583	7462009
5	120	CENTRE RD								1019586	7450235
5	124	CENTRE RD								1019588	5569630

Figure 1. Example of a Page from an AR Booklet (reduced in size).

was pre-printed in the last column of the booklet to assist the CR in any required Census follow-up operation.

4. METHODOLOGY

In this section, the creation of the Address Register (AR) is described. Figure 2 provides an overview of the steps involved.

4.1 Overview of the Methodology

The free-format addresses contained on the source files were first standardized into ordered component parts (steps 1 and 2) in preparation for the use of subsequent software. Then, postal codes were confirmed or corrected (step 3) so that those areas or worksites for which the AR was to be created could be selected from among all the addresses and locations contained on the source files (step 4). Because the same addresses could be contained on more than one file or more than once on the same file, unduplication of addresses based on both exact and probabilistic matching took place (steps 5 and 6).

Next, automated linkages were made of addresses to the blockface level using the Area Master File (step 7) or, where this was not possible, to Enumeration Area (EA) using the Postal Code Conversion File (step 8). After loading the addresses into a database management system (step 9), manual linkages were made of addresses to blockface (steps 10 and 11) or to EA

(step 12). Addresses within each EA were then sequenced by and within blocks (step 13) before being printed and collated in booklets by EA (step 14) for use in the Census.

4.2 Address Standardization (Steps 1, 2 and 3)

The Postal Address Analysis System (PAAS – step 2 of Figure 2) (Statistics Canada 1989c) performed two tasks: it broke up the free-format addresses from the source files into their component parts (street name, civic number, street designator, street direction, apartment number, municipality, province, postal code) and composed the address search key (ASK). ASK is an ordered concatenation of all the components of an address and is used during unduplication.

Although PAAS was an excellent product, analysis from the 1989 prototype had revealed certain shortcomings that we felt could be resolved by grooming or filtering the administrative file contents prior to using the generalized software. This FILTER step (step 1) concentrated on the following tasks: eliminating special characters with which PAAS refused to deal, repackaging address components in a manner compatible with PAAS, translating street designator short forms to acceptable ones, introducing commas between the street and municipality components of the free-format address to improve PAAS's comprehension, eliminating leading zeroes from civic numbers and numeric street names, and adding municipality and province names.

The FILTER and PAAS steps were applied in an iterative fashion. The first step was to discover what anomalies needed filtering for each administrative source. If the PAAS error rate after filtering was greater than 5%, error records were reviewed to find recurring problems that could be successively eliminated by further filtering until an error rate of less than 5% was achieved. As any address record that failed address standardization was eliminated from further consideration, it was vital to have a PAAS success rate as high as possible.

The PCVERFY step (step 3) used the Automated Postal Coding System (PCODE) (Statistics Canada 1989a) package for confirmation and generation of postal codes. It was not quite as effective as the PAAS software at address analysis and could only confirm or add postal codes for 84% of the output from PAAS. It confirmed 78% of the postal codes and changed another 6%. Only .003% of the source administrative records had arrived with no existing postal code. It was crucial to have correct postal codes because these would be used for worksite selection in the subsequent step.

Two problems arose in the PCVERFY step during production. If an address was missing a municipality/province component, the software continued to attempt to find a postal code instead of suspending further processing. As a consequence, enormous amounts of processing time could be spent trying to find postal codes. This problem was solved by including in the FILTER a step to add municipality and province names. The second problem occurred when a street name was numeric, as the processing time per address increased fourfold. This problem was not resolved and will necessitate modifications to the PCODE software.

4.3 Worksite Selection (Step 4)

This step partitioned the country by postal code into manageable worksites for processing with the sizes of worksites being based on the efficiency of CANLINK software for linkage of multiple large files. A geographic partitioning into worksites was adopted so they had dwelling counts in the 100,000 to 150,000 range based on the 1986 Census. Worksites were formed from an individual AMF (for a medium sized city), collections of physically adjacent AMFs (for small towns/townships), or parts of an AMF (for a large city). Geography Division's

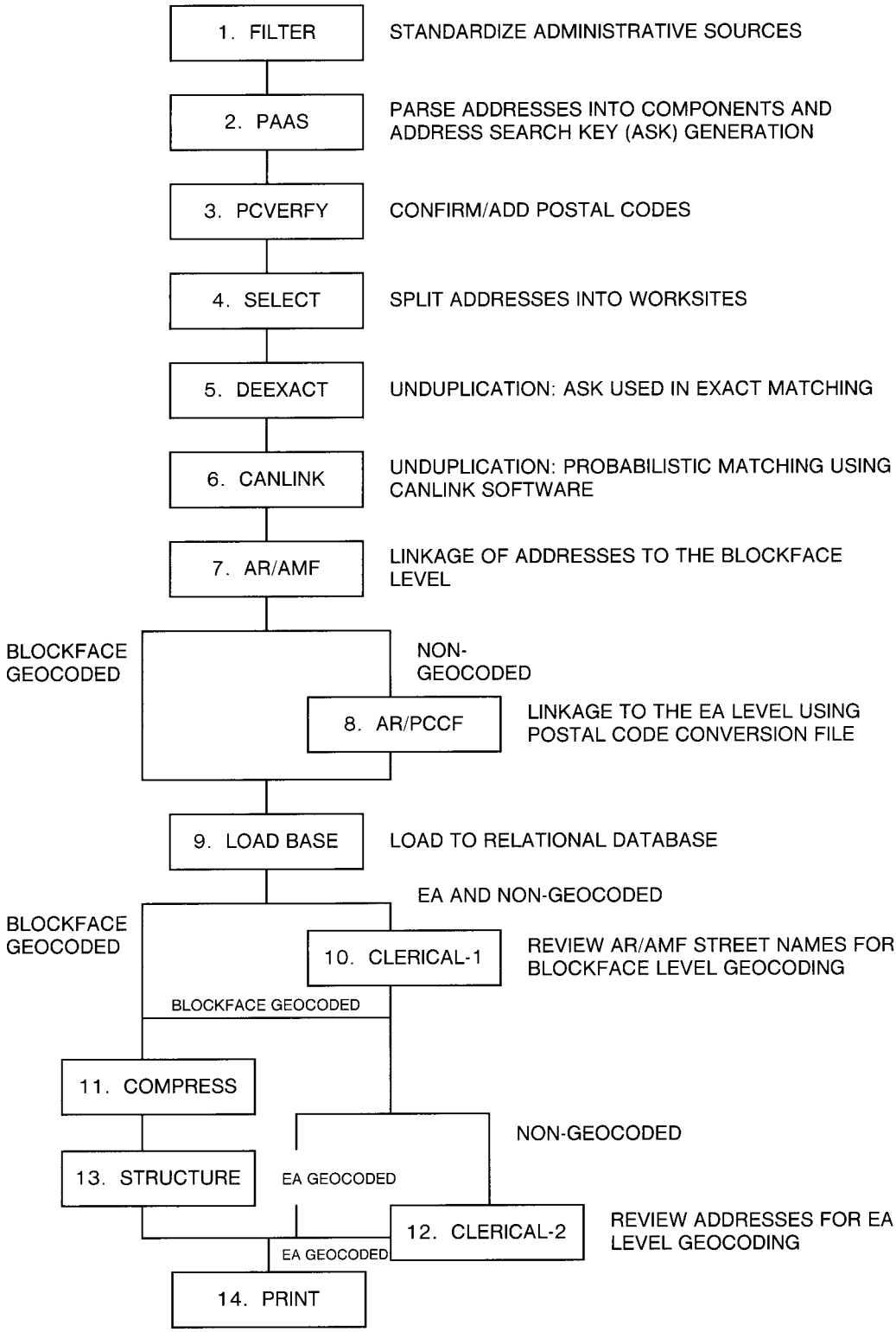

Figure 2. Overview of the Methodology.

Postal Code Conversion File (PCCF) which links postal codes and detailed Census geography was used to do this partitioning in the SELECT step (step 4). Once partitioning was completed, there were 105 distinct worksites and the original 43.4 million addresses had been reduced to 20.5 million addresses, with the dropped addresses having postal codes outside the AMF areas (*i.e.*, smaller cities and rural areas).

4.4 Unduplication (Steps 5 and 6)

In order to delete addresses included more than once on the source files, an unduplication process was conducted in two stages: an exact match with DEEXACT (step 5) and a probabilistic match using CANLINK software (step 6).

The DEEXACT step utilized the address search key (ASK) produced by the PAAS software and all records with an identical ASK were collapsed into a single record. With DEEXACT, the 20.5 million records from the SELECT step were reduced down to 10.1 million records. This reduction shows the importance of performing the address standardization.

Step 6 utilized the CANLINK generalized record linkage software (Statistics Canada 1989b). It clusters close records into groups called "pockets" and only records within the same pocket are actually matched together. For this application, civic number was used as the pocket. The components of the address (street name, municipality name, postal code, *etc.*) were used for matching purposes and weights were assigned for agreement or disagreement of each component. The development of levels of partial agreement for street name, municipality name and the last three characters of the postal code allowed for spelling variations and letter transpositions within the fields. The CANLINK step accounted for a further reduction to 6.7 million records. More details on the use of CANLINK in address unduplication are given in Drew *et al.* (1988), where its application in the November 1987 test is described.

4.5 AR/AMF Linkage (Step 7)

The major concern from the 1989 test was the strategy used to link addresses to their respective blockface. Because of the 11% drop in coverage from 84% to 73% compared to the 1987 test, a thorough investigation was needed and possibly a new approach. The other concern was that automated matching accounted for only 80% of the records matched while the other 20% were picked up clerically. This would have represented a daunting manual workload in full production. In order to circumvent these two concerns, another CANLINK application was developed for the AR/AMF linkage (step 7).

The original 1989 test files for Brampton still existed, so this became the test site for developing this step. The revised approach yielded 10% more matches, which increased the coverage back up to 1987 levels. As well, the automated matching was now responsible for 97% of the matches with 3% being picked up clerically, a significant improvement on the earlier 80%-20% split. Based on these results, the CANLINK approach was adopted for Census production.

In the construction of the new matching strategy, the first area of study involved a comparison of the contents of fields that would be used for matching purposes. This revealed certain anomalies that could be corrected prior to use to improve the number of linkages. The processing modifications to existing fields covered the following areas: removal of blanks between compound street names; alignment of street directions and civic numbers; conversion of numeric street names to numbers (on the AMF); removal of special characters in street names (on the AMF); correction of spelling variations in municipalities (on the AR); and a

recreation of certain PAAS translations for street names (on the AR). Several new fields were also generated: NYSIIS (New York State Identification and Intelligence System) and SOUNDEX versions of the street name, employing two phonetic encoding packages used to eliminate the effects of common spelling errors (Statistics Canada 1989d); a duplicate street name flag (on the AMF) to identify situations where a street name was not unique; a unidirectional street flag (on the AMF) to identify streets that had only a single street direction coded; and an official street name flag (on the AR) to indicate that the street name matched an official AMF street name. The AMF records contained only street data so we appended the Census Subdivision name and a province code and then attempted to assign postal codes to blockface civic numbers. When the postal codes differed between the "from" and "to" civic numbers, we generated subblockfaces for each unique postal code.

For this application, three distinct pockets were created for each record, effectively triplicating the files. The primary pocket was the most stringent in nature and was designed to find all the good match possibilities quickly in the first pass of the files. It was composed of street name/Forward Sortation Area (FSA)/odd or even civic number flag. The second pocket was postal code/odd or even civic number flag which allowed for poorly parsed addresses to be matched on postal code. The third was the NYSIIS version of the street name/odd or even civic number flag which allowed records with spelling variations in street name and missing postal codes to be considered as potential matches.

The function rules established for partial matches for street name, municipality name and the last three characters of the postal code were taken directly from our existing CANLINK application used for internal unduplication where they had already demonstrated their effectiveness.

However, there were three AMFs to which we had difficulty matching in the course of production: Red Deer, St. Thomas and Charny. The problem with all three was missing civic number data on the AMF. Knowing that these would require heavy clerical intervention, a field operation was mounted in December 1990 to update the maps from the Computer Assisted Mapping System (CAM). CAM maps from Geography Division were sent to Regional Office staff who added the missing civic number ranges. These updated maps were subsequently forwarded to Geography Division for inclusion in the next round of updates to the AMF. For the creation of the AR, the civic number ranges for the three AMFs were used manually in the clerical operation.

Success in matching was quite similar across all provinces except for Québec. In Québec, the automatic matching to the blockface dipped by about 10-12% to 73% as it was not as effective at dealing with French addressing as it was with English addressing. Three situations were identified as causes for the drop in the automatic match rate: the use/non-use of articles within the street name (*e.g.*, Savane, de la Savane, la Savane), the use of complete personal names as street names with a high degree of spelling variability (*e.g.*, Jean-François Belanger, J.F. Belanger and Jean F. Belanger) and the lack of street designators. As a result, the clerical operations described below, especially the first one, were of increased importance for matching in Québec relative to the other provinces.

During the AR/AMF processing with the CANLINK software, the only problem that arose was in exceeding an internal pocket maximum on the number of records allowed. The solution was to identify the streets causing the problem from the pocket report (they were always major thoroughfares) and set up special pre-processing programs that would add the fifth digit of the postal code in calculating the pocket value for those streets to make it more discriminating. This had the effect of reducing the number of records within the pocket.

4.6 AR/PCCF Linkage (Step 8)

This step (step 8) attempted to obtain an automated link to the proper Enumeration Area (EA) for those addresses which could not be matched to the blockface using the AMF in step 7.

The principal inputs were the Postal Code Conversion File (PCCF), which gave the correspondence between postal codes and 1986 EAs, and the 1986 to 1991 EA Correspondence File. By matching the two together we could identify postal codes that were uniquely matched to a single 1991 EA, as well as postal codes matched to two or more possible 1991 EAs, requiring manual work to resolve later in step 12.

Again, Brampton became the test vehicle. The analysis of the postal code/EA matching revealed that 38% of the postal codes could be uniquely assigned to a 1991 EA. The linkage to these postal codes of the AR records unmatched to a blockface yielded a further 5% increase in total matches. Overall, the automated match rate increased to 89% (84% to the blockface and 5% to the EA), up from 64% in the September 1989 test, almost cutting in half the amount of manual intervention.

4.7 Loading the Base (Step 9)

To facilitate queries and in anticipation of future usage, ORACLE had been used in the 1989 test as the database management system and was used again for the 1991 production. The ORACLE load step (step 9) involved the transformation of the up-to-now sequential file into four separate component files, one for each of municipality, blockface, street and address.

4.8 Clerical Procedures (Steps 10, 11 and 12)

The clerical procedure for the 1989 test was a review of all unique combinations of street name/street designator/street direction from both AMF and AR records along with an AR record count for each street combination. The objective was to replace an unmatched AR street combination with the legimate AMF combination. By comparing similar street combinations and determining which ones should in fact have been identical, hitherto uncoded AR records could be matched manually to a particular blockface. This procedure had worked well in 1989 and had proved useful in two problem situations: those where there were large discrepancies in street name spelling and those where the AR street name field contained both the street name and a street designator short form that the PAAS software had not understood in parsing the address.

We expanded the capability of this clerical procedure (step 10) to compare AR street combinations with other similar AR street combinations to handle instances where a particular street might have a number of AR spelling variations with no AMF equivalent. This expansion permitted some additional manual coding of addresses to blockface.

To summarize, in this first clerical procedure (Clerical-1), all addresses not coded automatically to blockface in step 7 (that is, those coded automatically to EA in step 8 and those not yet coded) were examined for possible manual coding to blockface.

Following the Clerical-1 procedure, we added a Compress step (step 11), which was applied to all records coded to the blockface. For each unique value of street name/street designator/street direction within a worksite, all the corresponding address records were checked for uniqueness using the civic number/apartment number as the key. Where multiple records occurred, they were collapsed with all pertinent data blended into one single record, a further step of unduplication.

As a result, at the end of step 10, the database contained addresses coded automatically or manually to blockface, automatically to EA or uncoded as yet.

Step 12 now dealt with those residual addresses that could not be linked to a unique EA but could be matched to two or more possible EAs via step 8. A complete set of CAM-generated maps was produced for the AR project. The Clerical-2 step consisted of examining these maps for the candidate EAs to assign these residual addresses to the proper EA wherever possible.

Overall, the ratio of automated to manual matching was 91%-9%. The automated portion comprised 87% from the AR/AMF linkage to blockface, and 4% from the AR/PCCF linkage to EA. The manual portion was split 3% matched to the blockface from the Clerical-1 operation and 6% to the EA in Clerical-2.

Although ORACLE was an appropriate vehicle for the 1989 prototype, it proved to be costly and eventually a bottleneck once in full production with the AR as just one user on a Bureau-wide database. It allowed for only 8-10% of the worksites on-line at any one time, and had to export and import sites continuously to free up space and reload to carry on processing. A second ORACLE database was therefore set up for exclusive use of the AR team. In fairness to ORACLE, not all the processing being done was conducive to any database management system. The product was being built and as a consequence large portions of the tables were being examined to make sweeping field changes, to eliminate duplication and to select records for printing. ORACLE did offer tremendous flexibility to change software procedures quickly and generate new ones as production unfolded.

4.9 Use of the Computer Assisted Mapping System (Step 13)

The Computer Assisted Mapping System (CAM) was a new research initiative for the 1991 Census whose development ran concurrently with AR development. The system generated all the Enumeration Area maps within AMF coverage areas. This was a major departure from the manual map generation process of the past. CAM also provided a structure to EAs that located blockfaces within blocks and sequenced the blocks within the EA (step 13). An off-shoot to CAM for AR purposes was set up to sequence the dwellings on the blockface. This was necessary to organize the address lists in a manner corresponding more closely to the way the Census Representatives do their listing.

CAM was fully implemented by the time of AR production. In order to remain compatible with it, the same vintage of the AMF that CAM employed was used. However, a small portion of blockfaces had no structure data assigned to them. For any EA where this percentage was greater than 5%, either CAM was re-executed for that worksite if time permitted or an alternate system, Point-in-Polygon Assignments (PIPA), that locates blockfaces within their EA was executed. Although PIPA shifted addresses from the structured portion of the AR booklet (based on blockface coding) to the unstructured portion (EA coding), at least the affected addresses were not dropped during the print selection process, which was the case when sequencing data were missing.

4.10 Printing and Booklet Production (Step 14)

The last production step was the printing and gathering of booklets (step 14) for the almost 23,000 Enumeration Areas containing at this point 6.6 million addresses. Major concerns which were addressed included print speed and quality (a continuous-page printer was used), durability of booklets (the booklets had front and back covers and were stapled) and compilation costs (the booklets were gathered and attached in-house).

5. POST-CENSAL EVALUATION

The post-censal evaluation can be broadly categorized into four study areas: field operations, data capture of AR booklets, update of the AR and determination of the AR contribution to coverage improvements.

Evaluation of field operations will focus on the effectiveness of training, how complete the reconciliation work was, and causes of errors, with a view to improving the methodology for future Censuses.

The data capture operation will yield two separate outputs. First, addresses printed in the booklets will be deleted if invalid, and if valid their Census Household Number will be captured. Second, the new addresses added by the Census Representatives will be captured. It will then be possible to calculate the AR overcoverage and undercoverage rates and the AR contribution to Census coverage. Addresses placed in the wrong EA can be investigated and traced back to the source of error. Through the Census Household Number, the number of persons added and characteristics of dwellings and persons can be studied.

From a cost perspective, the unit cost per dwelling added by the AR will be calculated, in view of the cost of creating the AR and using it in the Census.

6. FUTURE DIRECTIONS

The Address Register (AR), although initially set up as one of the procedures for reducing Census undercoverage, is a developmental project with potential impact on other programs within Statistics Canada as well as other government agencies.

The more immediate objectives for the future development of the AR are as follows: to incorporate the addresses identified during Census enumeration; to evaluate the effectiveness of the AR in improving coverage of the 1991 Census; to document and evaluate the production activities; and to develop a longer-term plan for the AR addressing its cost-effectiveness as a household frame, the optimal updating strategy and its potential for use by external agencies.

Within these guidelines, a project plan was prepared and is presented below under six main topic areas.

6.1 Relationships between the Census and the Address Register

Besides the potential for coverage improvement, other ways in which the AR could contribute to the Census will be explored. Some preliminary thoughts in this regard include possibilities for the AR to be used as a processing control file, for telephone numbers to be used for follow-up purposes, for creation of control numbers of dwellings in an Enumeration Area, for certification of dwelling counts for processing, or for migration analysis. Consideration will be given to whether the AR should be used before or after Census Day, and to how the AR might be used for those addresses where only a higher level of geography than the EA can be ascertained.

6.2 Relationships between Geography and the Address Register

As is evident in the description of the methodology, the creation of the AR relied heavily on many of the products from Geography Division (*e.g.*, the Area Master File, the Postal Code Conversion File). Their contributions and limitations in building the AR will be reviewed. For any new products developed by Geography Division, their possible use in the AR will be investigated with a view to incorporating the AR needs directly into the new product. As well, the AR will be integrated into the Geography Division's Geographical Information System (GIS).

The AR may be able to provide update indicators to the Area Master File (AMF) or for the delineation of Enumeration Areas. The AR could be used to establish priorities especially in high-growth areas or in areas where there are poor civic number ranges in the AMF. The updating of the Postal Code Conversion File might be served by postal code/Enumeration Area or postal code/blockface combinations from the AR. After each Census, all Census households are encoded with blockface centroids. Since the bulk of AR records have already been geocoded prior to the Census, a link of the AR with the Census Household Number will reduce the amount of manual geocoding work after the Census. This last project is already in progress.

6.3 Documentation, Evaluation and Improvement of Procedures

A user guide documenting procedures and a technical guide to document programs, sample problems and solutions and quality assurance are being prepared for the work done to date.

As with any new project, much is learned during the creative process and procedures are developed as required and as time and budget permit. After the fact, there are usually efficiencies to be gained by reviewing these procedures.

For the automated procedures, projects already underway include a more efficient use of ORACLE or choice of another system, the use of desk-top computers rather than the Statistics Canada mainframe computer, standardization of the filter, enhancements to PAAS, amalgamation of sites into provincial databases, the dropping of some fields earlier in the process, consideration of other postal coding software, improvement of address place name matching and an improvement of the Area Master File linkage with French addresses.

For the manual procedures, improved handling of adjacent Enumeration Areas across boundaries of Federal Electoral Districts and of the lack of civic numbers on CAM maps are to be pursued. The editing system to correct addresses will be reviewed for possible improvement as well.

Telephone numbers were added at a later stage within the AR production. A thorough evaluation of their coverage and accuracy will be undertaken especially in view of the potential uses of telephone numbers in the Census and other Statistics Canada surveys. For the latter, initial emphasis will be placed on testing within the context of the upcoming redesign of the Labour Force Survey.

Computer systems developed for the initial production have already been cleaned up to a large extent for better efficiency of mainframe expenditures, for programs and disk and tape storage, for file manipulation, for output, libraries and file access. Better system controls will be prepared.

This AR was produced only for urban areas. Future methodological development will examine the potential for extension to rural areas.

6.4 Updating Methodology

The AR was created from among four sets of administrative files: telephone files, municipal assessment files, hydro files and the T1 tax file from Revenue Canada. As well, the AR is currently being updated to be consistent with the 1991 Census so that the Census is also a source. The relative contributions of these source files, both in volume and quality, will be investigated so that a decision on acquisition of files for updating can be made.

An integral part of the updating strategy is the development of a methodology for updating. The definition of an update will be needed along with an update system. The cost effectiveness of ongoing updating, dependent on the various needs which result from projects identified throughout these future directions, will be considered as well. Is ongoing updating cost effective

when compared to updating only in time for the Census? What requirements will there be from other possible uses? Answers to these questions will lead to an updating strategy.

6.5 Other Uses of the Address Register in Statistics Canada

Besides the Census and geographical relationships presented earlier, a number of other uses are suggested within Statistics Canada. The potential use of the AR in the Labour Force Survey (LFS) will be investigated as part of the LFS Redesign Project. The possibility of using the AR in urban areas either to improve sampling under the existing area frame or as a list frame to reduce the number of stages in the sample design are two major areas highlighted for research. With telephone numbers on the AR, more telephone interviewing would be possible.

The use of the AR as a survey frame for other Statistics Canada surveys will be examined. In addition, since the AR currently uses telephone files as a primary source of information, it has these files on hand for further exploitation. The Special Surveys Program, the General Social Survey and the existing Labour Force Survey are areas which use or require telephone files.

Another potential application within Statistics Canada is as a housing database if the AR were enriched with housing data from the 1991 Census and data obtained from municipal assessment files, for example. The existence of such a database might reduce the amount of information on housing that would have to be collected in future Censuses. Data needs and availability have to be explored.

6.6 Uses of the Address Register External to Statistics Canada

If the AR is to be used outside Statistics Canada, issues of confidentiality of the source files and releasability of the AR must be addressed and meet the requirements of the Statistics Act. Some source files were provided to Statistics Canada in confidence, either contractually (*e.g.*, some files from Alberta) or legally (the T1 file from Revenue Canada).

6.7 Conclusion

The breadth and diversity of the ideas contained above in future directions demonstrate the potential of the Address Register as a geographical product with applications in many areas of Statistics Canada and elsewhere.

ACKNOWLEDGEMENTS

The authors would like to thank the many persons from the following areas for their dedication and perseverance in the creation of the Address Register: Phillip Reed and the AR Production Unit, Geography Division, the Labour Force Survey Sample Control Unit, Census Methodology, Survey Operations Division, the Main Computer Centre and Household Surveys Division. The authors would also like to thank the referee, Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed and Carol Sol for their helpful suggestions for this paper.

REFERENCES

BOOTH, J.K. (1976). A summary report of all address register studies completed to date. Report E-414E, Statistics Canada.

DICK, P. (1990). Address register - September 1989 test. Draft internal report, Statistics Canada.

DREW, J.D., ARMSTRONG, J.B., and DIBBS, R. (1987). Research into a register of residential addresses for urban areas of Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 300-305.

DREW, J.D., ARMSTRONG, J., VAN BAAREN, A., and DEGUIRE, Y. (1988). Methodology for construction of address registers using several administrative sources. *Proceedings of the Symposium on the Statistical Uses of Administrative Data*, Statistics Canada, 181-190.

FELLEGI, I.P., and KRÓTKI, K.J. (1967). The testing program for the 1971 Census in Canada. *Proceedings of the Social Statistics Section, American Statistical Association*, 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.

ROYCE, D. (1986). Address register research for the 1991 Census of Canada. *Journal of Official Statistics*, 2, 4, 447-455.

ROYCE, D. (1987). Applications of an address register in the Canadian Census. *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada, 207-215.

ROYCE, D., and DREW, J.D. (1988). Address register research: Current status and future plans. Internal report, Statistics Canada.

STATISTICS CANADA (1988). Area Master File (AMF), User guide. Statistics Canada.

STATISTICS CANADA (1989a). Automated Postal Coding System (PCODE), User and retrieval guide. Statistics Canada.

STATISTICS CANADA (1989b). Generalized Iterative Record Linkage System, Concepts guide. Statistics Canada.

STATISTICS CANADA (1989c). Postal Address Analysis System (PAAS), User guide. Statistics Canada.

STATISTICS CANADA (1989d). Record linkage software, Reference guide. Statistics Canada.

STATISTICS CANADA (1990). *User's guide to the quality of 1986 Census data: Coverage*. Catalogue 99-135E, Statistics Canada.

STATISTICS CANADA (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide. Statistics Canada.

SWAIN, L., DREW, J.D., LAFRANCE, B., and LANCE, K. (1992). The creation of a residential address register at Statistics Canada. *Proceedings of the Symposiumon Spatial Issues in Statistics*, Statistics Canada.

VAN BAAREN, A. (1988). Report on the November 1987 address register test. Internal report, Statistics Canada.