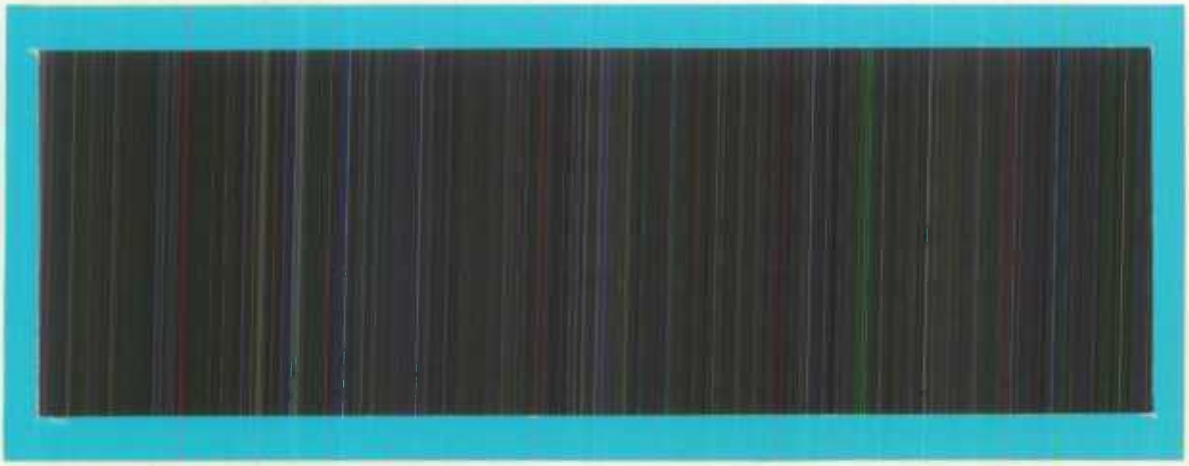
Statistics Canada Statistique Canada

NOT FOR LOAN
NE S'EMPRUNTE PAS



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

11-613
no. 92-05
c. 1

Canada

c1

WORKING PAPER NO. SSMD 92-005 E

METHODOLOGY BRANCH

OVERVIEW OF PRE/POST CENSUS ADDRESS REGISTER CONSTRUCTION

GURUPDESH S. PANDHER

NOT FOR LOAN
NE S'EMPRUNTE PAS

SSMD 92-005 E

STATISTICS CANADA STATISTIQUE CANADA
JUIN JUNE 15 1992
LIBRARY
BIBLIOTHÈQUE

Development and Analysis Section
Social Survey Methods Division, Statistics Canada

April 27, 1992

STATISTICS STATISTIQUE
CANADA CANADA
DEC 15 1992
LIBRARY
BIBLIOTHÈQUE

TABLE OF CONTENTS

List of Figures and Illustrations

Overview of Pre/Post Census Address Register Construction

Abstract

The Address Register (AR) project was undertaken by Statistics Canada as part of its continuing research into improving census methodology for the 1991 Census of Population and Housing. Originally envisioned in the 1960s and having gone through various developmental and small-scale testing phases over the past three decades, the AR project culminated in the creation of a national address register for urban areas in Canada. The primary objective of the 1991 AR project was to reduce census undercoverage in urban areas attributable chiefly to multiple-dwelling residential sites (Drew et al., 1990 and Royce, 1986). There are also several potential usages of the AR after the census once the original AR has been geo-coded (transcribed census geography information) and updated with address information obtained through census field operations.

This report seeks to describe the methodology behind the various operational steps used in constructing the AR in the pre-census phase of the project along with a description of the post-census AR data capture and update methodologies. An important consideration kept in view while writing this report was to provide a comprehensive description of the various operations and steps required in building the AR. Details of various processes are given, where possible, in a conceptual way so as not to clutter the report with an unnecessary degree of technicality. Although the report was originally motivated by the need for documented information on AR construction by the Labour Force Survey (LFS) Redesign in its planned evaluation of the geo-coded AR as a potential frame for the LFS, it may prove useful to those interested in becoming familiar with the AR project without getting lost in its intricate details.

Étapes pré/postcensitaires de la construction du registre des adresses

Résumé

Statistique Canada a entrepris le projet de registre des adresses (RA) dans le cadre de ses recherches permanentes pour améliorer la méthodologie du recensement en prévision du Recensement de la population et du logement de 1991. Le projet de RA, qui remonte aux années 1960 et qui a fait l'objet de divers travaux préparatoires et essais à petite échelle au cours des trois dernières décennies, a culminé avec la création d'un registre national d'adresses pour les régions urbaines au Canada. L'objectif premier du projet de RA pour 1991 était la réduction du sous-dénombrement dans les régions urbaines attribuable principalement aux immeubles à logements multiples (Drew et al., 1990 et Royce, 1986). Le RA se prête aussi à plusieurs applications possibles après le recensement une fois qu'il a été géocodé (que l'information géographique du recensement y a été transcrite) et mis à jour par l'ajout de l'information sur les adresses obtenue grâce aux opérations régionales du recensement.

Le présent rapport entend décrire la méthodologie qui sous-tend les différentes étapes opérationnelles de la construction du RA à l'étape précensitaire du projet, de même que les méthodes de saisie et de mise à jour des données du RA après le recensement. Le rapport a été rédigé avec le souci constant de décrire de façon exhaustive les différentes étapes et opérations ayant mené à la construction du RA. La description détaillée des différents processus est présentée, dans la mesure du possible, sous l'angle des concepts pour ne pas alourdir indûment le rapport avec des détails techniques. La décision de produire le rapport a été dictée à l'origine par la nécessité de fournir par écrit aux responsables du projet de remaniement de l'Enquête sur la population active (EPA) de l'information sur la construction du RA, information qui devait leur servir à évaluer le RA géocodé en tant que base de sondage possible de l'EPA. Il reste cependant que le rapport pourra être utile à ceux qui désirent se familiariser avec le projet de RA sans s'attacher à ses moindres détails.

1. Foreword and Implications for Potential Use of AR as LFS Frame

This report seeks to provide an overview of Address Register (AR) construction with the aim to facilitate an understanding of its potential role in the Labour Force Survey (LFS) Redesign. Knowledge about the operations used to construct the AR, it is hoped, will allow a quantified assessment of the strengths and limitations of the AR with respect to LFS Redesign applications. The description may also prove useful to those interested in becoming familiar with AR creation without getting lost in its intricate details.

An important consideration kept in view while writing this report was to provide a comprehensive understanding of the various operations and steps required in building the AR. This is, however, not done at the sake of readability. Details of various processes are given, where possible, in a conceptual way so as to not clutter the report with an unnecessary degree of technicality.

Upon reading this report it will become clear that a decision on the potential role of the AR as a LFS frame will, besides other factors, rest on concerns about AR construction with respect to the quality of the addresses and the telephone numbers used; the reliability of the record linkage methodology employed; and time-cost considerations of AR updating cycles. Evaluations on various aspects of the project are needed in order to quantify these concerns so that their implications on the use of the AR as a LFS frame may be assessed. Some of these issues and concerns, which require further investigation, are mentioned below.

Readers entirely unfamiliar with the AR project may find it more profitable to skip the remainder of this section for now and go to the body of the report.

- Contribution of AR Source Files

Four administration files consisting of income tax, municipal, hydro, and telephone files provide the starting bank of addresses. In assessing the degree of contribution from each source for future planning, it will be informative to determine which files contribute what fraction and vintage

of unique addresses to the AR.

- Updating Telephone Number

The procurement date of the administration files used ranges from May 1989 to October 1990 depending on file-type and worksite and is an important determinant in the reliability of addresses and telephone numbers present in the AR. It should be kept in mind that although the updated post-census AR will reflect address information gained during census field operations, the telephone numbers in the AR will be the same as those obtained from the original source files (in cases where they existed). A mechanism that periodically updates existing telephone numbers in the AR will need to be implemented. It should also be determined what fraction of the total AR records have usable telephone numbers and which files provide the highest quality numbers.

- Attrition Rates

The original pool of addresses used to construct and geo-code the AR goes through at least six major data processing steps, namely, the FILTER, PAAS, PCVERIFY, DEEXACT, unduplication and AR/AMF, linkage as shown in Figures 1, 2, and 3. Although addresses lost due to bad fields at each step are small, for data quality reasons these attrition rates need to be quantified and major contributing factors determined.

- AR Record Linkage Methodology

As explained in Section 5 (AR Construction and Unduplication), the record linkage methodology employed to determine the linkage threshold value used to identify address duplicate pairs from unique address pairs in the AR unduplication phase and to distinguish AR/AMF matches from non-matches during the AR/AMF linkage is conservative in rejecting true non-matches (has less statistical power) and more subjective than a modified approach proposed by the author (see Section 5). The result of using the former approach is that more address replicates and addresses with incorrect blockfaces will be present in the final geo-coded AR. The extent of these problems needs to be quantified and, perhaps, more research is required to determine the most optimal linkage

strategy -- both methodologically and time-wise -- for determining linkage thresholds.

In a production environment, requiring the processing of over 90 worksites each containing an average of approximately 50,000 records, by staff with no prior exposure to , there was an overwhelming need to use the simplest and most efficient approach to record linkage. These constraints, however, should not preclude further initiatives to investigate the statistical reliability of the linkage methodology used which might lead to improvements ensuring a more statistically rigorous linkage methodology in a production setting. In this regard, the author proposed a modified linkage approach which is statistically more reliable and less subjective.

- Cost, Time-frame, and Sychronization of AR Updating Cycles

AR creation and maintenance depends on information from two sources: 1) administration files, which serve as an address bank and 2) Geography Division's AMF, which serves as a vehicle to geo-code the AR addresses. In order to maintain the AR over a period of time, the AMF updating cycles -- which at present are driven primarily by the needs of a "foreign" division -- will need to be synchronized more closely with future AR updating cycles. Furthermore, cost and time-frame estimates of these activities also need to be estimated for feasibility and planning purposes.

2. Introduction and General Overview

In its entirety the AR project can be seen as comprised of two components: the pre-census AR and the post-census AR. The primary product of the pre-census AR project will be a geo-coded EA-based listing of urban addresses across Canada. The objective of the pre-census AR project is to reduce census undercoverage in urban areas attributable chiefly to multiple-dwelling residential sites (see Drew et al., 1990 and Royce, 1986). EA-based address listings produced by the pre-census AR project will provide census enumerators with a priori information on separate individual dwellings within multi-dwelling residential sites reducing the number of hidden or "camouflaged" dwellings which might otherwise be overlooked during census enumeration.

There are several potential usages of the post-census AR once the original AR has been updated with address information obtained through census field operations. Some of these have been enumerated by Drew et al. (1990) and Royce (1986). For one, it may be used to geo-code (census geography information transcribed to AR addresses) the census updated bank of addresses in the AR. Another usage, relevant to the LFS, is its possible use as a sampling frame once the geo-coded AR has been updated and expanded utilizing information obtained through census enumeration. There are also a number of other potential uses in applications which require information not just on the address, but also, on it's census geography.

Operationally, the pre-census component of the AR project is divided into three phases that will be explained in this report. These are as follows:

i) Construction of AR from administration files.
ii) Geo-coding of AR using Geography Division's AMF (Area Master File).
iii) Census AR list production by EA (Enumeration Area).

There are two further post-census phases to the AR project:

iv) Post-census data capture of AR census listings.
v) Post-census AR update.

As the name suggests, the pre-census AR Project was contingent upon completion of phases i), ii) and iii) which were earmarked for completion by April 1991 to allow for operational implementation by census day, June 4, 1991. Phase iv) followed next as modified EA-based census listings were returned upon completion of census field work. Work on the detailed planning of phase v), the post-census AR update, will commence after implementation of the preceeding step and will depend heavily on specifications for the post-census AR data capture methodology which have been devised.

3. Glossary of Terms

Before proceeding further, definition of terminology used extensively in this document is provided next.

EA: The census enumeration area is a geographic entity corresponding to the workload of an individual census enumerator (census representative). It is the basic building block of census geography which is used to construct bigger geographic constructs like CSD (Census Subdivisions). In urban areas it is a closed area consisting of and defined by street block networks. An EA will typically have around 300 dwellings.

CSD: Collection of contiguous EAs grouped to form a Census Subdivision conforming to existing municipal boundaries.

AR Worksite: In small urban areas a worksite may be a collection of CSDs; in large urban centres (eg. Toronto), it may be part of a CSD. The worksite breaks up urban areas into manageable chunks for AR data processing. Egs. Guelph (GPH00), an area of Toronto (TOR03).

Blockface: One side of a street segment bounded by two physical features or geographic structures (eg. intersections, rivers, railway tracks, other EAs). Technically, a blockface is defined by the following information:

bf_id: Unique blockface identifier.

street name, direction and designator.

bf_seq: Blockface sequencing number used for EA map generation.

st_num_low: Lower bound of the street number range for the blockface.

st_num_high: Upper bound of the street number range for the blockface.

zxy coordinate: Census geography location co-ordinate.

Address: Consists of information identifying a unique dwelling. Consists of the following fields: province, street name, street number, street designator, street direction, postal code and, if available, the telephone number.

4. Administration Source Files

Four files provided the bank of addresses used in the creation of the AR. The name, source, and vintage dates of theses files is given by the table below.

File Name	Source	Vintage Date
Tax	Revenue Canada	12/89
Telephone	Telephone Companies	12/89 - 5/90
Hydro	Utility Companies	5/89 - 10/90
Municipal Assessment	Municipalities	5/89 - 5/90

The tax file was obtained from Revenue Canada on a country wide basis and was broken into smaller manageable worksites to facilitate data processing. For this reason, all worksite tax files have the same date. The remaining files, consisting of municipal, hydro and telephone, were obtained at a more local level directly from provincial and municipal sources. Therefore, the period these addresses reflect varies from worksite to worksite with ranges given in the table above. More detailed information on file procurement cost and data quality measures such as the coverage,

uniqueness, and reliability of address and telephone numbers in the various files; address attrition rates at various processing steps; and results from investigations into the record linkage methodology employed will become available from evaluation studies planned after the census.

5. AR Construction and Unduplication Methodology

This phase entails the combination of the various source files into a pool of addresses which are then unduplicated to produce a concise AR list. This process is depicted in Figure 1. All processing is done at the worksite level. As a first step in this process, addresses from the different files, which may be structured by a provincial, municipal, city or some other regional configuration, are separated and recombined to create AR worksite files.

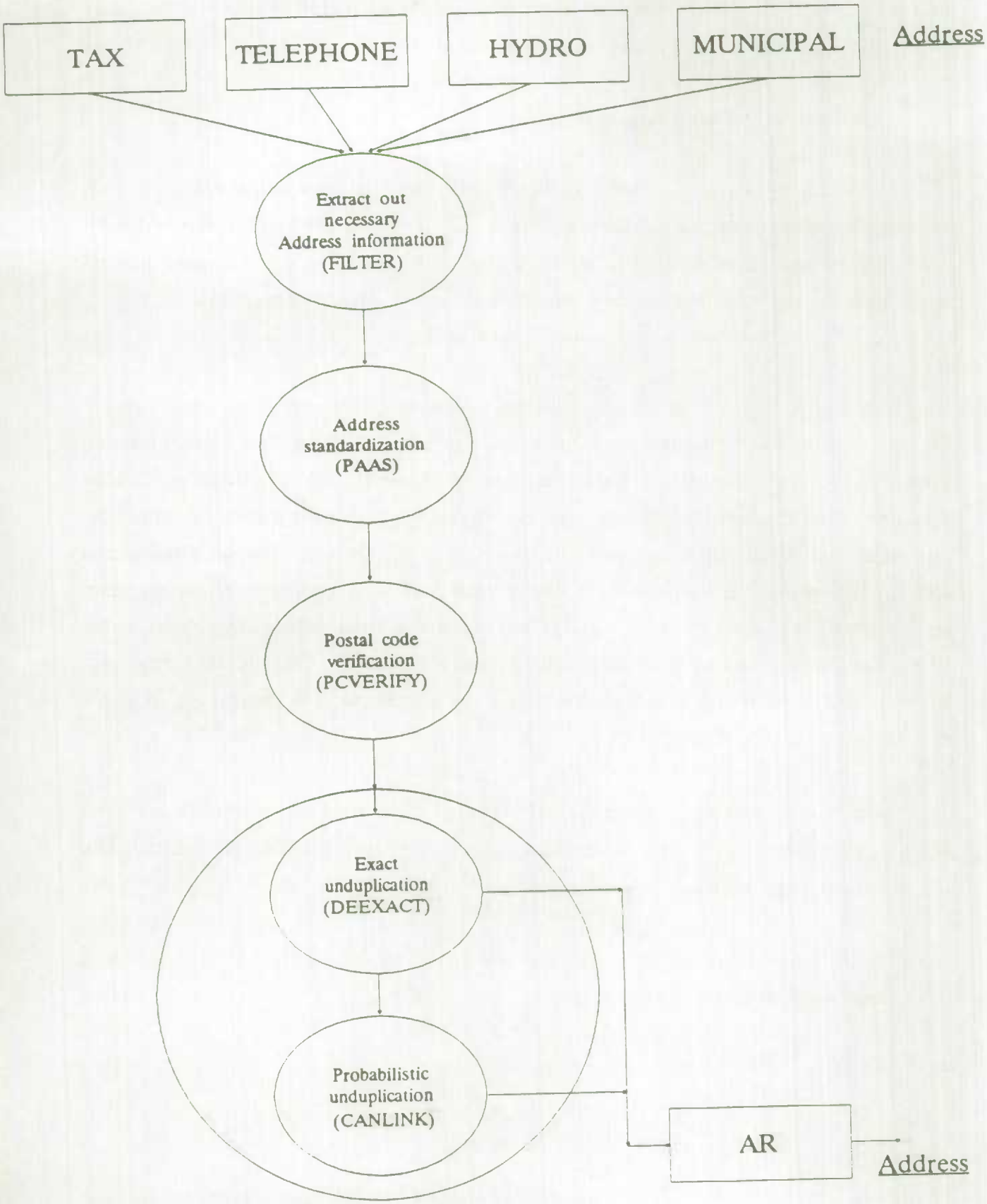
These files then proceed through a filter step which extracts the address information needed from the various files. Next, these address records are sent through PAAS (Post Address Analysis System) software which breaks up the address into its sub-components (eg. street name, street designator, street number, street direction, postal code, etc.). Whereas at the start of the process each address is in free-format, dependent on the origin of the file, after PAAS the free-format address has been analyzed and separated into its basic components. Furthermore, these components are recombined to produce a standardized ASK key to facilitate direct address comparisons. After PAAS, the addresses in a worksite are processed through PCVERIFY (Postal Code Verification) software. In this step the postal code it checked and, if discovered to be incorrect, changed; if it is missing, one is supplied if possible.

The final stage of this phase is the unduplication of identical addresses. In cases, where address replicates are found to exist, addresses with the latest vintage date are kept. Unduplication is performed through two steps:

a) Exact unduplication: Duplicate addresses are reduced by DEEXACT software which does exact internal file address matching by using the ASK key.

(Figure 1.)

A.R. CONSTRUCTION



b) Probabilistic Unduplication: Not all true duplicate addresses will have been removed by DEEXACT due to the very stringent exact match (on ASK key) condition used. It is plausible that spelling variations abbreviations or missing information on certain address fields will give rise to different ASK keys, while in fact these addresses represent the same dwelling. An attempt is made to further reduce these replicates through software.

Before describing the linkage methodology employed, it is important to note that the approach of using frequency weights (as termed in the documentation) was not employed in record linkage applications of the AR Project. The theoretical development of this approach was devised by Fellegi and Sunter (1969) and was further studied by Armstrong (1990). This approach is based on the likelihood ratio method of finding a linkage threshold value (critical region) for separating matched record pairs from non-matches.

The linkage methodology employed in all AR record linkage applications is based on the "global weights" approach as termed in the documentation. In this method, all address pairs within each pocket (homogeneous subdivisions of the worksite for linkage purposes) are compared using global weights derived from the empirical frequency distribution of address fields used in the comparison such as municipality, street name, designator , direction, postal code and postal qualifier. Selected fields are individually compared for complete and partial matching and assigned sub-weights depending on the degree of concurrence. These weights are then added to obtain a total linkage weight for each address pair comparison. This total weight is then checked against a worksite linkage threshold value, and if found to exceed the threshold, deemed a match (a duplicate). The threshold value is obtained by examining a sample of linkage results over a wide range of total weights to determine a score which appears to accept true matches (above threshold) and reject false matches (below threshold). The null hypothesis here is that the address pair compared consists of unique addresses (not a match) while the alternative hypothesis is that the address pair is a duplicate (a match).

Although, the global weights approach used does not guarantee any of the uniformly optimal statistical qualities theoretically granted by the frequency weights approach, it does, however, offer an efficient and simpler method of record linkage which does not require

intricate knowledge of . In a production environment, requiring the processing of 90 worksites, containing an average of approximately 50,000 records each, by staff with no prior exposure to , there is an overwhelming need to use the simplest and most efficient record linkage approach. These constraints, however, should not preclude further initiatives to investigate and quantify the statistical reliability of the linkage methodology used which might lead to improvements ensuring a statistically rigorous linkage methodology in a production setting. In this regard, the author proposed a modified global weights record linkage approach which is statistically more reliable and less subjective.

The global weights method of finding the linkage threshold is not totally satisfactory for two reasons. First, the sample of 100 to 500 linkage pair results examined are selected without the use of a sample design and the selection of the sample is determined entirely by the unknown internal selection mechanism of . Secondly, the determination of the threshold based on visually comparing fields and relating them to the assigned linkage weight relies to a great extent on the subjective judgment of the user and not on a quantifiable methodology.

To address this problem a method was proposed by the author that enables users to find the linkage threshold with less subjectivity by using a "Linkage Weight Range Field Agreement Probability" (LWRFAP) table (see example on next page). This table utilizes information on all address linkage results instead of relying on a sample and furnishes information, for important selected fields, on the probability distribution of observed matches, partial matches and disagreements for bands of total global weight ranges. This methodology was not applied to all sites but was used as a quality assurance measure on some sites. It was found that the original global weights method was more conservative (had less power) but was not seriously out of line. It failed to unduplicate some record links when they were in fact true duplicates with the effect that more replicates were not reduced. This problem is less serious than false matching and its consequent elimination of unique records when in fact they are truly unique (type I error).

WORKSITE: MNN00 (Montréal/North)

Linkage Weight Range Field Agreement Probability (LWRFAP) Table: Example of method to find the linkage threshold using information on all worksite linkage results.

Address Comparison Field	Total Weight Range	Missing Count (n_m)	Other Count (n_o)	% Missing (n_m) (n_m/n_m+n_o)	% Agree (n_A) (n_A/n_o)	% Partial Agree (n_{PA}) (n_{PA}/n_o)	% Disagree (n_D) (n_D/n_o)	Choosen Range
MUN	15-40	0	110	0	100	0	0	
	41-60	0	8,870	0	99.95	0	0.05	
	61-80	0	1,600	0	92.91	0	7.09	x
	81-100	0	140	0	99.31	0	0.69	
	101-125	0	2,840	0	100	0	0	
	126-150	0	800	0	100	0	0	
	151-175	0	21,850	0	100	0	0	
POSTAL 1	15-40	110	1	99.12	100	0	0	
	41-60	8,860	6	99.93	66.67	0	33.33	
	61-80	1,552	127	92.44	100	0	0	x
	81-100	0	144	0	100	0	0	
	101-125	0	2,838	0	100	0	0	
	126-150	0	800	0	99.75	0	0.25	
	151-175	0	21,852	0	100	0	0	
POSTAL 2	15-40	113	1	99.12	0	0	100	
	41-60	8,863	6	99.93	66.67	0	33.33	
	61-80	1,552	130	92.44	95.28	4	0.79	x
	81-100	0	140	0	97.22	2.78	0	
	101-125	0	2,830	0	97.46	2.54	0	
	126-150	0	800	0	100	0	0	
	151-175	0	21,852	0	100	0	0	
STNAME	15-40	0	110	0	0	100	0	
	41-60	0	8,870	0	99.88	0.12	0	
	61-80	0	1,680	0	99.05	0.83	0.12	x
	81-100	0	140	0	.69	2.78	96.53	xx
	101-125	0	2,800	0	2.43	97.57	0	
	126-150	0	800	0	51.82	48.18	0	
	151-175	0	21,852	0	100	0	0	

x Denotes total weight ranges where the disagreement rate for the field declines to "acceptable" levels and continues to decreases over subsequent ranges.

xx For the STNAME field, the range 61-80 was chosen over 81-100 despite the fact that 96.53% of the record pairs in the latter group are non-matches. Closer examination of the linkage results for this group of 140 records revealed that the very high disagreement rate was due to an anomaly and that the threshold may less conservatively fall in the lower 61-80 range. This conclusion is consistent with results from the other three fields in the table.

6. Geo-coding of the AR

The purpose of this step is to link AR addresses to the EA (census enumeration area) they fall in and to assign geographical coordinates which connect the dwelling with its census geography. Address listings by EA will be used by census enumerators as a "dwelling blueprint" during the 1991 census in order to reduce undercoverage.

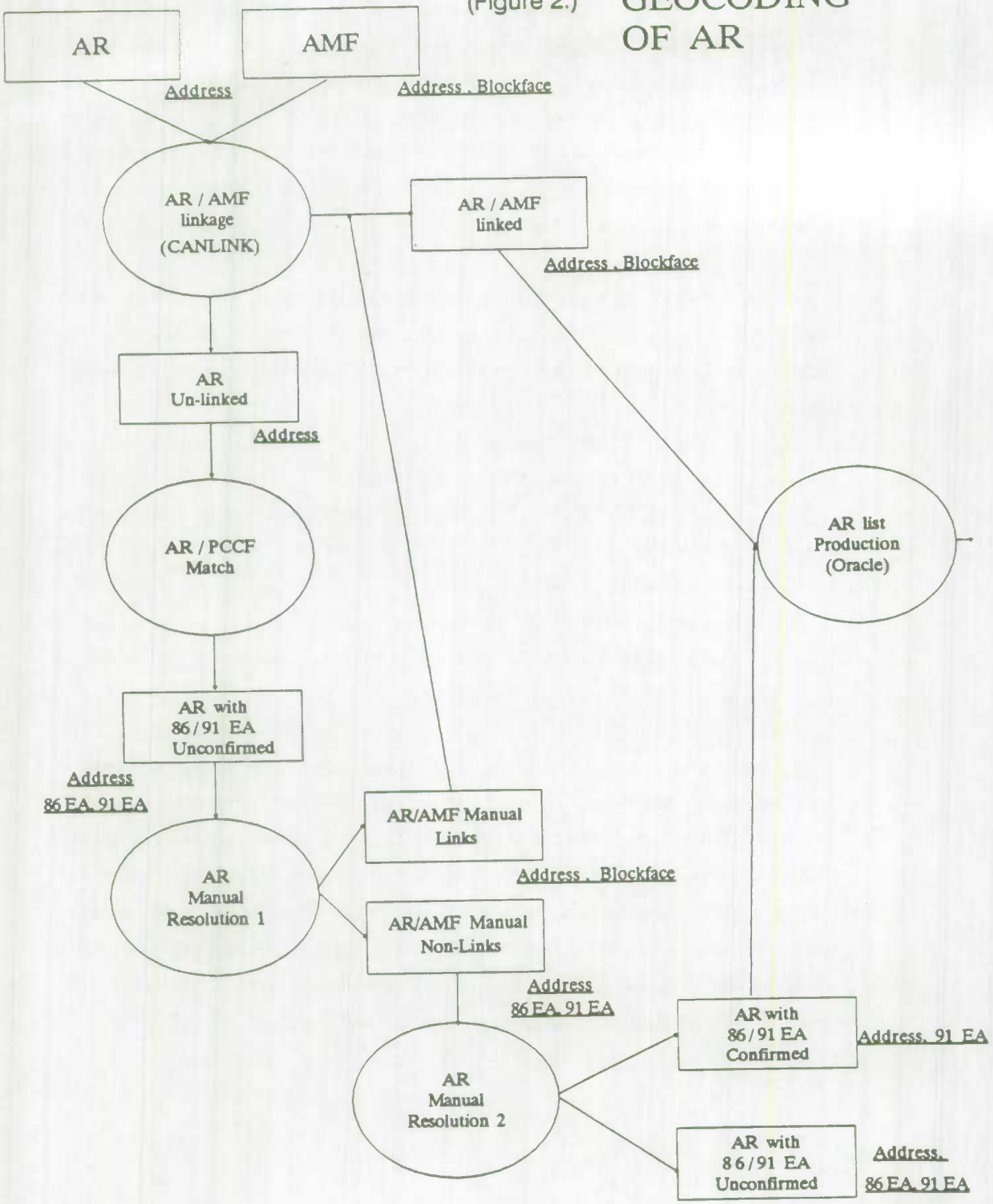
The mapping of AR addresses onto enumeration areas (EAs) by the assignment of census geography blockface information is accomplished through the AMF (Area Master File) maintained by geography division. Besides other information, records in the AMF contain blockface information consisting of bf_id, street name and designator, st_num_low, st_num_high, and zxy coordinates (see Glossary, Section 3). Each record in the AMF represents a range of dwellings falling on one side of an uninterrupted street segment, defining a blockface. Hence, the AMF has the potential to map an address in the AR to its corresponding census geography (blockface) information, represented in particular by the AMF fields consisting of bf_id and the zxy coordinate. The process of assigning AMF blockface information to addresses in the AR is called the geo-coding of the AR and is depicted graphically in Figure 2.

The complete geo-coding process consists of two automated operations followed by two manual operations. The two automated steps utilize record linkage and DEEXACT exact matching software, respectively, to link the AR/AMF and PCCF/86 EA/91 EA files, processing approximately 87% of the AR address. An additional 10% of AR addresses not assigned a blockface during the automated procedures are salvaged in the two manual operations, leaving 3% of AR address records unresolved.

6.1 AR/AMF Record Linkage

At this stage of the AR project, the AR consists of address records containing information on fields such as the province, street number, street designator, street direction, postal code, and if avalable, a telephone number. Furthermore, address fields, including the postal code, were analyzed by PAAS and PCVERIFY software to yield a standardized

(Figure 2.) GEOCODING OF AR



address (ASK key) and a corrected postal code for each AR record. The AMF, on the other hand, contains blockface information on street segments collectively specified by the blockface fields bf_id, bf_seq, st_num_low, st_num_high, and zxy coordinate . Conceptually, the geo-coding operation can be seen as matching each address in the AR to its corresponding street segment on the AMF and adding the blockface information to the AR address. The final outcome of this step is an AR/AMF linked file which contains address information and its corresponding blockface information for each individual dwelling in the AR successfully matched to the AMF.

The mapping of each address in the AR to its correct street segment (within each EA) represented by the AMF blockface is accomplished through software. The following three pockets (record groupings within which worksite record comparisons are performed) were used:

i) STEET NAME.
FSA: First 3 letters of the postal code.
ODD-EVEN FLAG: Flag for street number being even or odd.

ii) PCODE: Postal Code.
ODD-EVEN FLAG

iii) NYSSIS-NAME: Abbreviated street name.
ODD-EVEN FLAG: Indicates whether the street number is odd or even.

Within each pocket, the municipality, street name, designator, direction, postal code and postal code fields are compared and assigned global sub-weights coherent with the result. The total global weight obtained by cummulating these sub-weights is then compared with a threshold value determined from reviewing a sample of comparisons for the worksite. If the total global weights exceeds the threshold, the comparison is deemed a match and the blockface information from the AMF record is merged with the AR address and written to the output AR/AMF linkage file. Record comparisons yielding a total weight less than the threshold value are deemed non-matches and proceed to the next step, the AR/PCCF match as shown in Figure 2.

6.2 AR Non-linked/PCCF Match

In order to facilitate the manual 2 operation, AR records not geo-coded in the AR/AMF linkage are linked with Geography Division's Postal Code Concordance File (PCCF) using DEEXACT exact matching software. The 1990 version of the PCCF provides a mapping of 1986 enumeration areas to their corresponding postal codes. The other Geography product used was the 1986/1991 EA Correspondence File (86/91 EA File) which provides a mapping of 1986 EAs to 1991 EAs. First, postal codes common to both AR addresses and PCCF records are used to obtain the identity of the 86 EA the address belongs to. Subsequently, possible 91 EAs in which the address may be located are determined by looking up the 86 EA in the 86/91 EA file.

A complicating feature of this procedure is that in many cases the 91 EA ascertained may mot be unique. This complication arises from, firstly, the overlapping of postal code zones over multiple EAs, and secondly, the splitting of 86 EAs over/into multiple 91 EAs. The fact that parts of two or more EAs can have the same postal code means that, in these cases, the postal code by itself is not suffficient to resolve the EA identity of the address. Changes in EA boundaries from 1986 to 1991 adds a second source of ambiguity since an 86 EA may be transformed to a 91 EA totally intact or may be broken into smaller units and dispersed among many 91 EAs. One-to-many transformation of addresses from 86 to 91 EAs lead to the problem of correctly identifying the 91 EA in which the AR address is located. An attempt is made in the manual 2 operation (described fully in the next section) to resolve these ambiguities.

Upon completion of the AR/PCCF linkage, each non-geo-coded AR address will be supplemented with information on the 86 EA and 91 EA that the dwelling belongs to. More precisely, AR records will now have the following information: address, EA 86, and EA 91. Since the purpose of the AR project is to produce address listings of dwellings by 91 EA, residual non-geo-coded AR records not

assigned an AMF blockface during the AR/AMF link will be assigned their respective 91 EA's during the manual operations described in the next section.

6.3 Manual Operations

The manual 1 and manual 2 operations described in this section attempt to salvage AR addresses unsuccessfully linked to AMF blockfaces during the AR/AMF linkage as a result of the "type II errors" and the AMF updating problems discussed below. The manual 2 operation is, however, contingent on the AR/PCCF match described in Section 6.2.

Failures in AR/AMF match may arise as a result of the following two reasons.

i) "Type II" Errors

An AR address may not match to its AMF blockface, although in fact the address falls in the street segment defined by the blockface. In statistical language this situation is a akin to possible reasons leading to the occurrence of type II errors. In this context, the null hypothesis is the event that an AR address does not fall in any AMF blockface (non-match), while the alternative hypothesis is the event that an AR address falls within a AMF blockface (match).

These non-matches may occur when address information of an AR address is at variance with the address range information of an AMF blockface. This may result from misspellings or use of alternative spellings in the AR for the same street name in the AMF and/or mistakes/omissions in other address fields. Note that a mistake in one or more address fields does not necessarily imply an immediate non-match. Only if it happens that the total global weight derived from the sum of individual field comparisons during the record linkage falls short of the threshold match value will AR/AMF address pair be considered a non-match.

ii) AMF not Up-to-date

Non-matches will also occur if the AMF street segment within which the AR address falls is not present in the AMF. The AMF is maintained by Geography division with various

periodic update cycles. It is plausible that certain street network information, especially for high growth urban areas, is not up-to-date in the version of the AMF being used for geo-coding.

6.3.1 Manual 1 Operation

This operation is intended to salvage those AR address records not linked to the AMF blockface due to "type II errors" described above. In these cases, the AR/AMF record pair is a true match, but for various reasons, such as spelling variations, errors or omissions in the individual address fields, comparisons failed to yield a cumulative global weight that exceeds the threshold level, hence rejecting the match.

The operation is performed by first printing (by worksite) a combined list of unique AMF and non-linked AR records for each unique combination of street name, street designator, street direction and postal code. Each record in this listing is flagged as belonging to either the AR or the AMF. The street information for each AR and AMF record is compared within each unique combination of street name, street designator, etc. and a decision is made as to whether any of the AR addresses fall within one of the street segments specified by the AMF records in that street information combination. In cases where these comparisons lead to a match, AR matched address fields which differ from the AMF address fields are replaced by the latter.

The batch of non-geo-coded AR addresses corrected during the manual 1 operation are then re-linked to the AMF through in order to assign them an AMF blockface. Approximately 3.4% of the total AR addresses were salvaged in this step. The total here refers to the total number of AR records valid at the start of the AR/AMF linkage.

6.3.2 Manual 2 Operation

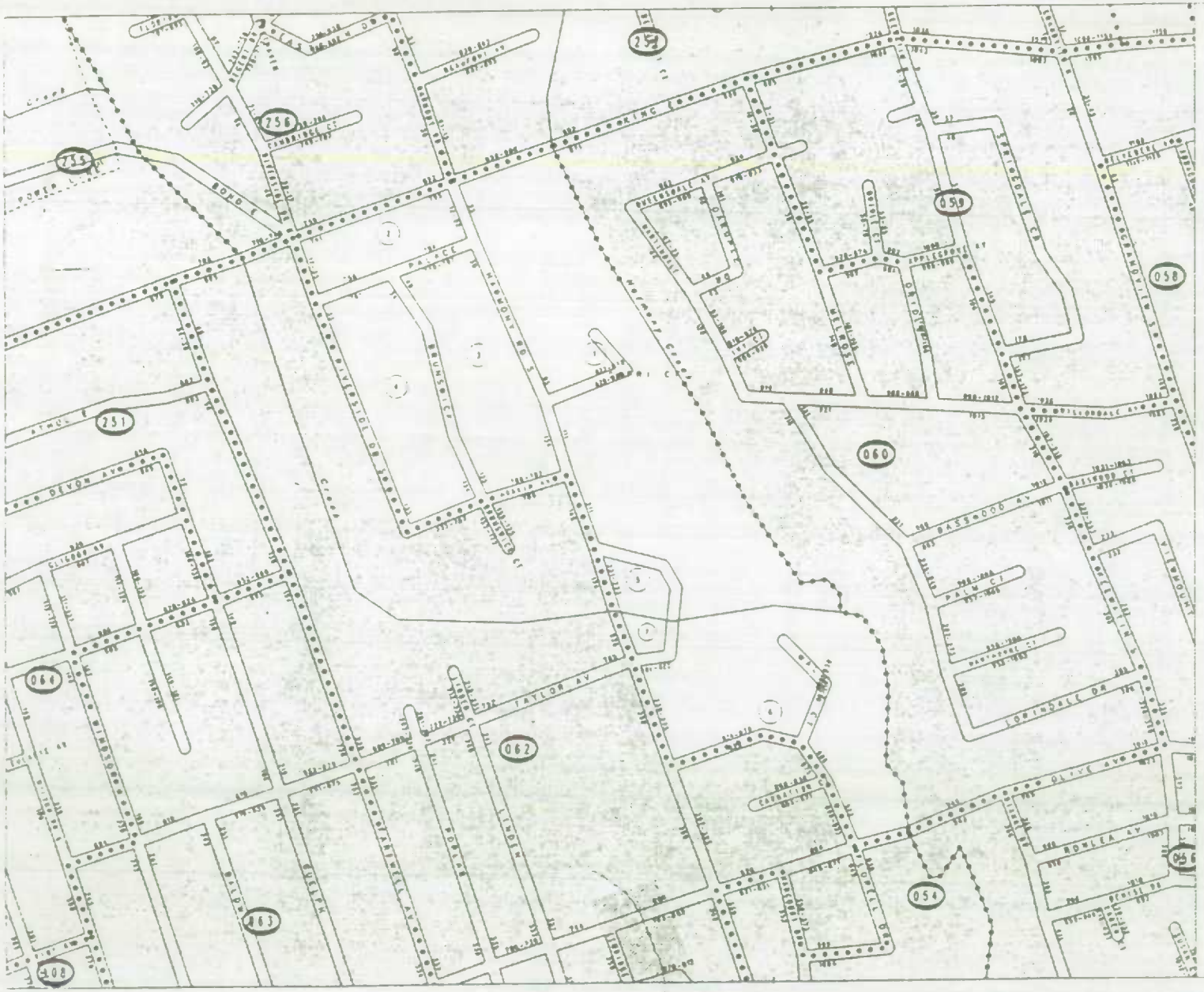
Residual AR records not geo-coded during AR/AMF linkage and not resolved during the manual 1 operation are handled here in a last attempt to assign these non-geo-coded AR records to their respective 91 EAs.

Since the previous manual operation attempts to recover AR non-links caused by "type II errors", a majority of the records processed at the manual 2 stage are associated with AMF update problems. As mentioned in section 5.2, the AMF represents the status of street networks at the time that the most recent updates to the AMF were performed. Hence, some AR addresses being sought in the AMF may not be present and hence will not be linked to a blockface.

The manual 2 operation makes use of the 86 EA and 91 EA information attached to the non-linked AR records during the AR/PCCF match. Recall that at the completion of this step, the AR records have address, EA 86 and EA 91 information. To implement this operation a listing of residual AR addresses is produced which contains information on EA 86, EA 91 and address. In many cases, changes in census geography from 1986 to 1991 may cause a 86 EA to be split among many 91 EAs. In these cases only one of these 91 EAs is listed. Current EA maps, city maps, and postal code directories are used to determine which 91 EA a particular address falls into. This process typically involves finding the AR address on the city map, then locating it on the 91 EA map displaying street range information. Other aids such as the postal code directory may be used to resolve the location of the address in the EA street network (a sample 91 EA map is attached on the following page).

In situations where AR addresses from the listing have been identified on the EA map, the correct 91 EA is entered on the listing. This information is then used to update the residual batch of non-geo-coded AR addresses through Oracle software, which will now also include the identity of the EA the address falls in.

91 Enumeration Area (EA) Map for EA = 061 in FED = 35057



1991 CENSUS RECENSEMENT CANADA

FED-CEF 35057
CCD-DCR 02
EA - SD 061
CT - SR 008.02

1 : 5000
0 120 240 36

CMA - RMR
OSHAWA

CSD - SDR
OSHAWA, C

Prod 31 05 90

7. Census AR List Production

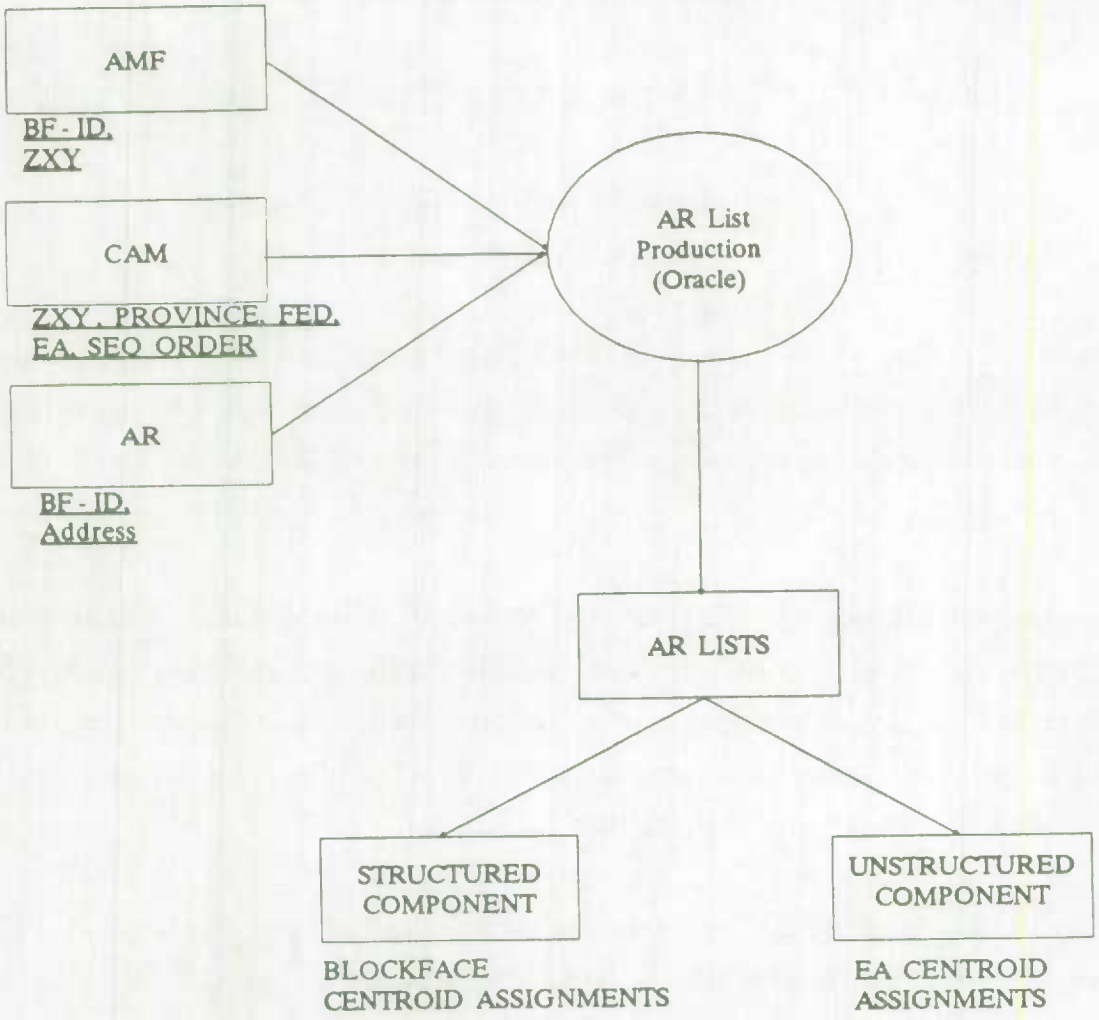
The final phase of the pre-census AR project is directed at producing urban address listings by EA. Addresses in each EA listing are sequenced using AMF blockface information in dwelling-contiguous order so as to minimize the leg work required of census enumerators in covering households in their assigned EA. These lists are prepared by making use of Geography Division's Computer Assisted Mapping (CAM) facilities which contain street network sequencing information which enable the construction of EA maps maintaining the contiguity and sequencing of dwelling and streets within the EA. A graphical summary of the list production process, by EA, is given in Figure 3.

Prior to list production, the usable addresses processed thus far fall into two categories. The first category, the structured list, contains AR addresses which were assigned blockface (geo-coding) information from the AMF as a result of the AR/AMF linkage and the manual 1 operation. Approximately, 87% of the addresses in the final AR fall into this category. The second category, the unstructured list, consists of residual addresses which failed to be assigned blockfaces at both the automated AR/AMF linkage and the manual 1 operation but were salvaged during the combined AR/PCCF linkage and the manual 2 operation by assignment to their corresponding 91 EAs (10% of AR records).

Excluded from both the structured and unstructured lists are those addresses which were not assigned blockfaces, or at the very least 91 EAs. Although the proportion of addresses not appearing on the EA listings fluctuate from worksite to worksite, a crude running estimate of the overall rejection rate is 3%. More accurate figures will become available upon completion of the AR evaluation scheduled after the 1991 Census.

The AR list production process is carried out using Oracle software, and is depicted in Figure 3. In this step, information described in the following table from three files is used.

(Figure 3.) AR LIST PRODUCTION BY EA



SOURCE	INFORMATION LOADED
AMF	BLOCKFACE information as described in Section 3 including BF_ID and ZXY Coordinate.
CAM EA	FED, EA, SEQ ORDER, ZXY Coordinate
AR/AMF Linked File	ADDRESS and BLOCKFACE information. including BF_ID.

During this operation address records in the AR/AMF linked file are assigned to their corresponding EAs and then listings of addresses falling into each EA in a dwelling-street contiguous order are produced using the blockface sequencing information of the CAM EA file. This forms the structured portion of the EA listings.

In the structured lists, assignment of EA to addresses in the AR/AMF linked file is performed by first matching the AMF and AR/AMF linked file by BF_ID and then matching by ZXY coordinate to the CAM EA file to assign EAs to addresses. Next, addresses with common EAs are collected and blockface sequencing information contained in the CAM EA file is used to prepare an ordered street-dwelling contiguous list of addresses for each EA.

Finally, the unstructured EA lists, containing AR addresses manually assigned 91 EAs after failing assignment to an AMF blockface during the AR/AMF Linkage, are appended to the structured lists in street name order.

8. Post-Census AR Data Capture

All potential uses of the AR after the census such as the geo-coding of the census database or the use of the AR as a LFS frame will depend heavily on the quality and dependability of the AR. The quality of the AR can be improved immediately after the census by incorporating into it new and revised address information obtained as a result of the reconciliation of AR and census visitation record (VR) listings during the census. Urban addresses missing in the original AR but discovered to exist during the census will be incorporated into the AR along with corrections to existing addresses to produce the most up-to-date snapshot of urban address geography across Canada.

Computer software needed to perform the post-census AR data capture was developed by Headquarters Operation Division (HOD). Specifications for these are laid out in the report prepared for HOD entitled "Specifications for AR Data Capture" (see References). A sample of addresses from a simulated AR census listing used to test the AR data capture programs is given on the following page. The typed information is what existed on the listing before the census while the hand written information was entered by census enumerators during the AR/VR reconciliation operation. This operation is described in detail in the "Procedures Manual: Address Register Reconciliation" and all changes to AR address listings during the census stem from this step.

Briefly, the purpose of the AR Reconciliation operation is to assign household numbers (column 5 in AR listing) to AR addresses in cases where the AR address is confirmed to exist in the real world during census enumeration and consequently entered on the VR listing. AR addresses which are not present in the VR are investigated in the field after the initial visit and, if validated, assigned a household number. Entries in columns 6 to 10 of the AR listing result from this process. The converse case, of an address being present in the VR but not in the AR points to an address missed during AR construction. These addresses are directly entered in the blank space at the end of the AR listing and assigned a household number in concordance with the VR listing. The assignment of household numbers from the VR to the AR listing will make it possible to later link the geo-coded pre-census AR to the census database.

The purpose of the AR data capture operation is now clear: it captures all new and revised

EA-based AR listing Modified during AR/VR Reconciliation (mock-up)

ADDRESS REGISTER
REGISTRE DES ADRESSES

Protected
Protégé

PROVINCE 35
FED - CÉF 213

EA - SD 003
VN - NV 2

Block No / No d'îlot	Address - Adresse: Civic No / No de voirie	Street / Rue	Apt No / No d'app	Hhld No / No de ménage 48-50	Not Listed at Drop-off / Non inscrit à la livraison 51	Field Follow-up Required / Suivi sur place requis 52	Invalid - Non valide: Duplicate / En double 53	Outside EA / En dehors du SD 54	Other / Autre 55	AR Ref No / No de réf du RA	Telephone Number / Numéro de téléphone
1	2	3	4	5	6	7	8	9	10	11	12
30	8	CAMBERLEY CR								1028975	4164519018
~~30~~	8	CLAYPINE TRL			✓	✓		✓		8035656	4164537496
~~30~~ 30	7	CAMBERLEY CR			✓	✓	✓			4028969	4164521177
30	7	CAMBERLEY CRESENT	B24		✓		✓			2028994	
30	7	CLAYPINE TRL		50		9				0035655	4164543419
30	6	PAUL MILLIAM GT								1096605	4164561941
~~30~~	6	CAMBERLEY CR		123						7028963	4164597840
45	6	CLAYPINE TRL			✓			✓		3035654	4164505537
48	5	CAMBERLEY CR					✓	✓	✓	3028955	4164521628
48	5	CLAYPINE TRL			✓	✓	✓	TYLO Problem		3035649	4164520866
48	4	CLAYPINE TRL			✓	✓	✓ ←	✓		4035644	TYLO Problem 4164573022
~~48~~	3	~~CLAYPINE TRL~~ Maywood		249						6035638	4164538972
~~48~~	2	CLAYPINE TRL			✓	✓	✓		✓	7035633	4164519831
~~60~~	2	PAUL MILLIAM GT		6BC						2096600	4164528657
70	2	SMENNEN- DR		AO	✓	✓			✓	4110657	
80	1	CLAYPINE TRL				✓				0035622	4164574668
14	86	Woodwind Pl. G	16M	522							
14	89	Le Heron Bank	E19	530							
014	73	Pickfield-Estenan Dr N	110	535							

address information for urban areas obtained during census field work along with assigned VR household numbers. For the sake of space, the AR data capture methodology will not be described here in great detail. Those wishing further details should refer to memoranda directed to HOD (Headquarters Operation Division) from June 1990 onwards listed in the References section at the end. The data capture methodology is dealt with in a summary fashion below.

The data capture operation consists of two phases: 1) the data capture of census information of existing addresses, and 2) the data capture of new addresses and their census information. In phase 1), census information (province, FED, EA, VR household number, and various flags) for existing addresses in the EA-based AR (pre-census version) along with changes/correction to address fields achieved through census field work (AR/VR Reconciliation) are captured. Phase 2) involves capture of addresses missed during pre-census AR construction but found to exist during census enumeration together with their census information. These two phases are described next in sub-sections 8.1 and 8.2.

8.1 Data Capture of Existing Addresses

The EA-based AR forms will originally contain information in columns 1 to 4 (see AR form attached). Information in the remaining columns, 5 to 12, will be transcribed from the VRs during the AR/VR reconciliation process and field follow-up mentioned in the previous section.

8.1.1 Data to be Captured

Data in the province, FED, EA and VN fields appearing at the top of the AR form (attached) along with data in columns 1, 5, 6, 7, 8, 9, 10 and 11 will be captured. Furthermore, corrections made to printed address information in the first four columns (consisting of BLOCK NO, CIVIC NO, STREET NAME and APT NO) during AR/VR reconciliation will also be captured where they apply. Record layout specifications for each of these fields consisting of name, position, size and data-type description of the flag fields can be found in the report entitled

"Specifications for Adress Register Post-census Data Capture" (see References).

8.1.2 Address Reconciliation Flag

The AR forms will originally have printed information in columns 1, 2, 3, 4, 11 and 12. The remaining columns, 5 to 10, are flagged (for numeric fields, flagged refers to a numeric entry) or left unflagged (blank) during the "Address Register Reconciliation" operation. An analysis of the reconciliation operation described in the procedures manual for this operation reveal that only 7 different valid configurations can arise among columns 5 to 10, if the reconciliation is done correctly. These possibilities, along with three additional groupings signifying problems of various sorts, are enumerated in Table 1 along with their corresponding "address reconciliation" flag values that will be assigned to each address record by the data capture software.

Table 1: Address Reconciliation Flag Values

		Column Number in AR Form						Reconciliation Flag Value
Case	Remark about Address	5	6	7	8	9	10	
1	In both AR & VR	X	-	-	-	-	-	1
2	Invalid	-	X	-	-	X	-	3
3	at "Classroom"	-	X	-	X	-	-	4
4	Field Follow-up	-	X	X	-	-	X	5
5	and Invalid	-	X	X	-	X	-	6
6		-	X	X	X	-	-	7
7	Field Follow-up and valid	X	X	X	-	-	-	2
8	Missed	-	-	-	-	-	-	8
9	Any Other	all other configurations (with correct data type)						9
10	Invalid Data	data in fields(s) is of incorrect data-type						0

In Table 1 above, only addresses with reconciliation flag values of 1 and 7 are valid usable addresses; addresses flagged as 2, 3, 4, 5, and 6 were determined to be invalid during the reconciliation process; addresses flagged as 8 suggest a missed address; addresses flagged as 9 represent all remaining possible configurations, and are all invalid; and, finally, addresses flagged as 0 indicate an incorrect data-type value in at least 1 field.

The reconciliation flag will help the post-censual update of the AR by identifying valid usable addresses and by providing statistical information on the occurrence of different types of problematic addresses.

8.1.3 Data-Change Flags

Aside from the automated coding of the reconciliation flag which identifies whether the address in the AR EA lists is usable, the "data change flags" identify addresses where corrections were made to either the BLOCK NO, CIVIC NO, STREET NAME or APT NO fields as a result of census enumeration. A zero valued data change flag signifies no change to the field; a value of 1 implies a changed field and a flag value of 2 represents a field incorrectly changed (either wrong data-type or length).

8.2 Data Capture of New Addresses

The previous section described the process through which VR household numbers for existing AR addresses along with modifications to these addresses observed during census field work will be captured. New addresses missed during the pre-census construction of the AR but added by hand to the AR listings during the AR/VR Reconciliation operation will be captured in this part of the operation. For new addresses, data to be captured include the province, FED, EA and VN fields appearing on the top of the AR form along with data in columns 1 to 5 consisting of BLOCK NO, CIVIC NO, STREET NAME, APT NO, and HOUSEHOLD NO.

9. Post-Census AR Update

The previous step, the post-census AR data capture, will result in the creation of two data captured files: one containing existing AR addresses with VR household numbers and the second containing new addresses missed during AR construction but caught during the census. This phase is directed at accomplishing the following objectives:

i) Assigns VR household numbers to geo-coded AR addresses. The presence of census VR household numbers to the AR makes it possible to link the AR to the census database.

ii) Update the pre-census AR by correcting existing address fields in cases where discrepancies were observed between the real address and the AR address during the census, and by augmenting the AR with new addresses missed during AR construction. Note that these new and revised addresses will also include their assigned VR household numbers. The geo-coded AR, updated by the census, makes possible the multiplicity of uses -- one being the use of the AR as a Labour force Survey Frame -- requiring knowledge of both urban addresses and their census geography.

The post-census AR update will be implemented once AR production activities for the 1991 Census are complete and will depend heavily on the parameters defined by the AR data capture methodology.

The first step to updating existing addresses in the AR (pre-census AR) will entail examining the reconciliation and data-change flag values to identify usable addresses and then merging/updating with the pre-census AR. Only addresses with reconciliation flag values of 1 and 7 are valid usable addresses; addresses flagged as 2, 3, 4, 5, and 6 were determined to be invalid during the reconciliation process; addresses flagged 8 suggest a missed address; addresses flagged as 9 represent all remaining possible configurations, and are all invalid; and finally, addresses flagged as 0 indicate an incorrect data-type value in at least one field, necessitating a closer look to ascertain address

usability. Data-change flag values make it possible to determine if valid changes were made to address fields. Flag values of 1 imply valid alterations to address fields and in such cases the pre-census AR addresses will be updated.

New addresses found as a result of census field work will next be added to the AR, yielding the most up-to-date bank of urban addresses in Canada. Before combining these new addresses with existing and revised AR addresses which are already geo-coded, the new addresses also need to be geo-coded in order to link these addresses with their corresponding census geography. This will again be done through record linkage of these addresses with Geography Division's Area Master File (AMF) using software. At the completion of this step, AR addresses will be assigned blockface information from the AMF. Note that although the EA of these new addresses is already known from the census, other geo-coding information such as the BF_ID (blockface identifier) and the ZXY coordinate of the blockface are not known until after the link. These geo-coded addresses which also contain census visitation record (VR) household numbers will be added to the pool of existing addresses in the census updated AR.

In this final form, the post-census geo-coded AR will, for each urban household contained in it, have information on the household's address, census geography (geo-coding) information, and the census visitation record (VR) household number. It is the presence of this combination of data for each AR address that gives rise to the proliferate uses -- only some of which have yet been envisioned -- of the post-census geo-coded Address Register.

Acknowledgements

The author wishes to thank Jack Gambino and Bryan Lafrance for acting as referees to this paper who, along with Doug Drew, helped shape this final report through the benefit of their comments on earlier drafts of this paper. I would also like to express gratitude to Bryan Lafrance and Kim Charland who were consulted for details on certain aspects of the AR project, and to Cristine Larabie, who typed this report.

References

Armstrong, J.B. (1990). **Weight Estimation for Record Linkage**. Methodology Branch Paper, Business Surveys Methods Division.

Drew, J.D., Armstrong J., Baaren, A.V., Deguire, Y. (1988). **Methodology for Construction of Address Registers Using Several Administration Sources**. Methodology Branch Paper, Social Surveys Methods Division.

Pandher, G.S. (June 25, 1990). **Cost & Time Frame for AR Data Capture Operation**. External Memorandum, Social Survey Methods Division.

Pandher, G.S. (October 5, 1991). **Specifications for AR Data Capture**, External Memorandum, Social Survey Methods Division.

Pandher, G.S. (October 15, 1990). **Estimate of Extra Cost and Number of Addresses Invalid in Extending AR Data Capture to Include Changes to Existing Addresses**. External Memorandum, Social Surveys Methods Division.

Pandher, G.S. (January 25, 1991). **Evaluation of AR Data Capture Program and Changes Required**. External Memorandum, Social Survey Methods Division.

Pandher, G.S. (January 25, 1991). **Specifications for AR Data Capture (updated)**. External Memorandum, Social Surveys Methods Division.

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

Ca 00S

Statistics Canada Library
Bibliothèque Statistique Canada

1010094661