GEOCODING ADMINISTRATIVE DATA

Nelson Kopustas, Douglas Norris, and John Leyes, Statistics Canada

1.0 INTRODUCTION

An important strength of the development of administrative records for statistical purposes is the potential ability to derive *small area data* on an annual or more frequent basis. A prerequisite of doing this is the ability to geocode the administrative records to the desired level of geographic detail. Traditionally, most census and survey data on individuals are collected and tabulated on the basis of the individuals' place of residence. The geographic coding of place of residence is generally done by field staff prior to data collection. With administrative records, the situation is different. Generally, the only source of geographic information on an administrative record is the mailing address that is used by the administrative agency to correspond with the individual. Therefore, the tabulation of data from administrative records depends on the ability to use the mailing address to geocode the records to the desired classification system.

In Canada, the postal code is an integral part of the mailing address. The postal code is a six character alpha-numeric code that efficiently summarizes the mailing address and lends itself to computerized manipulation and processing. A by-product of recent work on the development of administrative records for statistical purposes has been the development of an infrastructure for geocoding administrative records to small geographic levels. In the U.S., work on the development of small area population and income estimates from Internal Revenue Service records has also been concerned with the geocoding of administrative records using the zip code.

The main purpose of this paper is to discuss the concepts underlying the relationship between residence location and mailing address; the limitations and problems in using the mailing address are identified; there is a discussion of the techniques for geocoding administrative records; finally, the issue of geographic coding is discussed from the point of view of three national Canadian administrative files - the tax records, unemployment insurance records and family allowance records.

2.0 CONCEPTS

The process of allocating individual administrative records to place of residence contains a number of assumptions and transformations related to the concepts of mailing addresses, postal codes and geographic areal systems. In this section these basic concepts are discussed and in the next section transformations between the concepts are considered.

2.1 Mailing Address

The mailing address is generally the sole interface between an individual and an administrative agency. Most social statistical data are tabulated on the basis of where individuals live, that is their residence address. In using administrative records the only source of geographic information is a mailing address and therefore an important question is the extent to which a mailing address corresponds to a residential location.

In using the mailing address as an indicator of residence address, two problems are encountered. First, some persons may not use the mailing address of their usual residence but rather may use some third party or "care of" address such as an accountant, parent, post office box or business address. A second problem is that, in rural areas, the mailing address corresponding to a usual residence may lack specificity in that the mailing address may simply identify a post office and therefore cover a large geographic area crossing municipal or other boundaries. Similar situations arise for rural routes out of a city or in new suburban housing developments where households temporality receive their mail at one central location, often a set of mail boxes located near the entrance to the new development.

Mailing address can be classified according to the mode of mail delivery. The main types of mail delivery are:

- **Letter carrier delivery (street address).**
- **Rural route delivery.**
- **Suburban service delivery (one location for a whole subdivision).**
- **General delivery.**
- **Post office box delivery.**
- **Rural post office (combination of general delivery and rural route delivery).**

These different types of mailing address can be summarized in terms of their geographic specificity as follows:

2.2 Postal Code

The Canadian postal code is a six character code starting with a letter and alternating numbers and letters. The first letter signifies a province or part of a province (e.g., M for Metropolitan Toronto, V for British Columbia). The first three characters are called the Forward Sortation Area (FSA) and designate a service area. Postal codes can be identified by the mode of mail delivery as indicated in the previous section. In urban areas with door-to-door mail delivery, the postal code identifies a block-face or a single large apartment building. In rural areas the postal code simply identifies a rural post office. All persons in rural areas served by a single post office have the same postal code and the codes are easily identified as all having "0" as the second digit (e.g., K0M 2R0). Currently, there are about 600,000 distinct postal codes in Canada. Approximately 6,000 of these are rural codes that cover 25% of the population and the remainder are urban codes including rural routes out of cities, suburban service, etc.

Postal codes are quite stable over time in the sense that the great majority of dwellings and business locations never change postal codes. There are however some exceptions, for example when a rural area or suburban area is converted to door-to-door carrier delivery. This would result in a number of block face postal codes replacing a single postal code. In fact, the situation may be further complicated if only part of the rural or suburban area is converted since these codes con-

Type of Address	Non-specific Address	Specific Street Address
Residential	Rural routes Rural post office Suburban service	Letter carrier- Residential
Convenience	Post office box General delivery	Letter carrier- Business

tinue to exist but refer to different geographic areas. Postal codes may also be retired but are never re-used for a different area.

Postal codes on one set of administrative records, the individual income tax records, have been checked for accuracy. A sample of 1,500 records were recoded manually and a comparison with the original postal code indicated differences in 4.3% of the cases. However, in the majority of these cases (70%) the differences were in the last three characters of the postal code that may have affected the block face coding but probably not the census tract nor the municipality. Therefore in all but about 1% of the cases, the postal code is quite accurate. Furthermore in the cases where the postal code and mailing address do not correspond, there is some evidence that data processing operations (e.g., key entry) are responsible for most of the differences.

2.3 Geographic Area Systems

The Standard Geographic Classification (SGC) has been developed by Statistics Canada as a national system of subprovincial areas based on municipalities. The SGC code is hierarchical. The municipalities are grouped into counties or census divisions (depending upon the province) which are then aggregated to provinces or territories. There are 5,710 census subdivisions (e.g., municipalities) and 266 census division or counties and 12 provinces or territories. For example, the code 3537039 is the code for the province of Ontario(35), Essex County (37), City of Windsor(039).

Of course with small area data, geographical area systems abound and there are many different, often overlapping systems. For example, in one province the statistical agency has discovered 48 area systems that are used with little or no relationship between them. The key to geographic coding, therefore, is to use as small as possible a building block to permit aggregation to a diverse set of areas. The postal code is a potential candidate for such a building block and this will be discussed in a later section. At Statistics Canada the work with the postal code has emphasized census divisions, Federal Electoral Districts and to a lesser extent, census tracts.

3.0 TRANSITIONS

Thus far, the concepts and structures of mailing addresses, postal codes and area systems have been examined. This section will concentrate on the transitions between these entities.

3.1 Mailing Address to Postal Code

As noted above the postal code is an efficient summary of the mailing address. Furthermore, as the postal code has become more accepted as part of a person's address this item is generally supplied by the individual(1). In fact, recently, a surcharge was instituted for business mail without a postal code.

There are nevertheless a subset of records that may not contain a postal code. For these cases software has been developed to assign postal codes to mailing addresses. This software structures the mailing address and matches it to a similarly structured master list of address ranges and their corresponding postal codes. Because many local addressing conventions exist, the problem is difficult to solve in a general way. Not only may the address be difficult to structure, but the resulting structure may not match the "official" address. In spite of these drawbacks, the current software developed by Statistics Canada assigns postal codes to approximately 85% of the addresses.

The six character format of the postal code makes it very efficient for data processing compared to an unstructured thirty to sixty character address. In most cases the information loss in using only the postal code is minimal. For example, single street addresses are grouped into block faces; all apartments in a building are grouped into one code (in the case of small apartments, these may in fact be grouped with other dwellings on the block); post office boxes may also be grouped. Clearly, these groupings are insignificant for any statistical applications. In rural areas the loss is somewhat greater since a number of distinct rural routes that might be indicated in the mailing address, all have the same rural postal code. Other information loss occurs since "care of" addresses can no longer be identified and small business locations may have the same postal code as residential dwellings on the same block.

3.2 Postal Code to Geographic Areas

3.2.1 Conceptual Considerations

In considering the mapping between postal codes and geographic area systems, the key question is "To what level of detail can each type of postal code be allocated in terms of residence?"

In the case of urban areas with door-to-door carrier delivery, there is virtually no limit to the level of geographic coding since the block face postal codes can be mapped to tracts, enumeration areas or even block face centroids.

However for the remaining types of delivery modes the specificity with which a postal code can be mapped is much less certain. For example, for a postal code corresponding to rural routes that run out of towns or cities, the routes may cut across the municipal boundaries of cities, counties, etc. Clearly, it is not possible to allocate precisely such postal codes although the codes may be allocated to larger geographic areas such as census metropolitan areas, economic regions, etc. The situation is similar for post office boxes and business addresses where the individual probably lives within commuting distance of the mailing address. An exception is the case of care-of addresses where the problem is perhaps the most difficult since the address may be that of a relative if the individual is temporarily away, or, in the case of a tax return, the address may be that of an accountant, generally a local accountant but in some cases perhaps an accountant hundreds of miles away.

To summarize, in mapping postal codes to geographic area systems, most urban codes can be mapped with high accuracy but rural codes, and codes for rural routes are much less specific and these can only be assigned to larger regions. These codes should be assigned only to geographic areas where there is a reasonably good chance that the mailing address and residence address of the individual both lie in that geographic area. In addition to this problem, the use of non-residential mailing addresses causes problems in mapping postal codes, although again it is probable that even these can be assigned to large regions with little loss of accuracy.

3.2.2 Implementation

The tool to facilitate the allocation of data with postal codes to geographical areas has come to be known as a "conversion file". The conversion file for all of Canada has one record for each postal code. That record contains the postal code and the geographical codes for areas to which the postal code has been assigned. On the Statistics Canada conversion file all postal codes have been assigned to census divisions and Federal Electoral Districts. For some census metropolitan areas, postal codes have also been assigned to census tracts. Many other conversion files have been created for specific sub-national regions. For example, files have been constructed to map postal codes to school districts and provincial electoral districts in Manitoba, community associations in the City of Calgary and census consolidated subdivisions in the Province of Newfoundland. The Province of British Columbia has developed what is perhaps the most extensive file that maps postal codes to a large number of area systems including municipalities, health districts, forestry districts, school districts, enumeration areas, census tracts, etc.

The Statistics Canada conversion file was originally created by matching the postal code master file (containing address rang-

es and postal codes) to 1976 Census street index files (containing address ranges and geographic codes such as census divisions). The postal codes that did not match were coded using a Forward Sortation Area (first three characters of the postal code) to census division table and a manual follow-up. Annual updates to add new areas have been accomplished by Statistics Canada using the latter two methods. The manual process and in particular the updating is very dependent on the acquisition of up-to-date, good quality maps.

In creating any conversion file a number of assumptions must be made as to how—and, in fact, if-to map certain types of postal codes. For postal codes corresponding to street addresses the process is straightforward and mapping can be done to a very fine level of detail. At the other extreme the postal codes corresponding to rural post offices, rural routes and general delivery cover a much larger area. The simplest approach is to code these to larger areas in which the post office is located. For example, for the files maintained by Statistics Canada these codes have been assigned to census divisions and counties and federal electoral districts but they have not been assigned at the census tract level and in many cases could not be assigned to municipalities. The remaining types of codes that correspond to post office boxes, commercial office buildings and large volume mail receivers are more problematic, since these are known to be non-residential postal codes. Again it might be assumed that individuals live within a large region containing the point of mail delivery. However, in mapping codes to smaller geographic areas such as census tracts or enumeration areas such codes are probably best left unassigned. In a set of census tract tabulations done using income tax records, business codes were assigned to the census tract level and the result was substantial overcoverage in the downtown business district.

The assignment of rural and other non-specific codes even to larger regions involves some mis-assignment in the case of those codes that cross boundaries. For any region there will probably be some persons living in the region but receiving postal services outside the region and conversely others who are living outside the region but receiving their mail from a post office in the region. Conceptually, other more elaborate mapping procedures could be developed, for example a postal code might be allocated to two or more geographic areas on a proportional basis. A practical problem with this approach is that the boundaries for most rural codes are difficult to document and to relate to more standard area systems. There is also the question of what to use and how to estimate the allocator.

Once a conversion file is created based on assumptions about how to map different types of codes, it is desirable to evaluate the accuracy of the conversion file. A thorough evaluation of a conversion process would require individual records to be coded using postal codes and independently by more traditional census or survey methods. The two coding techniques could then be compared. As yet this approach has not been attempted, however it has been possible to do some indirect evaluation to identify areas where the net misallocation problems are "large".

The technique compares aggregate counts of children derived on the basis of postal codes from the administrative records of the Family Allowance Program to Census counts. The Family Allowance program is virtually universal for children under 18 years and therefore counts of children should closely approximate counts of children from the Census of Population. In fact, comparisons for June 1981 for the population 1-14 show the two sources give virtually identical counts at the national and provincial levels. However, at the subprovincial level the counts from Family Allowance are derived using postal codes while the Census counts are based on traditional census geography on place of residence. Therefore, aside from intra-provincial differences in census or family allowance coverage, differences in counts at the subprovincial level, particularly larger differences are probably due to postal code boundary problems.

Comparisons between the June 1981 family allowance counts and census counts of the population 1-14 years for census divisions have been done. The results are shown in Table 1. Differences of less than 3%, or perhaps even 5%, could be due to many different factors. However, differences of more than about 5% are probably in large part due to postal coding problems.

The results show that of 260 census divisions, the difference in the counts was greater than 10% in 17 cases (7%), between 5% and 10% in 30 cases (11%) and between 3% and 5% in 33 cases (12%). Eight of the seventeen cases with deviations in excess of 10% were in Manitoba, and this reflects the sparsely populated nature of central Manitoba (i.e., areas served from post offices in other census divisions) and boundary problems near Winnipeg. Further analysis will determine the importance of size and other factors in explaining the discrepancies. Similar comparisons will also be carried out for other area systems, in particular federal electoral districts and census tracts in selected metropolitan areas. Although the comparative technique is crude, it is useful in identifying large geocoding problems.

Note that the differences reported in Table 1 are net differences and clearly mask larger differences that have to some extent been offsetting. However, even the net differences are in some cases quite large and, if one is interested in absolute counts, perhaps some adjustment or combining of regions should be considered.

4.0 SOME APPLICATIONS

In using the mailing address as a geographic indicator of residence address, the accuracy of the data will depend in part on the congruence between the residence and mailing addresses. One factor that will affect data quality, is the extent to which individuals use their postal code and also the extent to which they use third party or convenience addresses. Clearly, these factors may vary with the administrative record set under development.

The incidence of missing postal codes has been tabulated for three main federal level administrative files - the individual income tax records, unemployment records and family allowance records. The incidence of missing postal codes is approximately 8% for the tax records although the use of geocoding software to assign postal code from mailing address reduces the incidence of missing postal code to about 2%. For family allowance and unemployment insurance records, the postal code is more complete, with less than 1% missing.

Part of the explanation for the low level of missing codes is the policy on the part of administrators to follow-up and include postal codes wherever possible. In these two programs, cheques are periodically mailed to recipients and the correct postal code allows for more efficient delivery.

These three administrative data sets have also been tabulated by type of postal code as discussed above. In the case of the monthly unemployment insurance records two months were considered to allow investigation of the possible effects of seasonality. The results are shown in Table 2. The main finding is that for all three record sets, nearly 98% of the records have a domestic postal code corresponding to an urban residential code, a rural route or rural post office. However, a sizable proportion of these are rural codes that may cover large geographic areas. The incidence of rural codes ranges from 31% for the tax file to 44% for the February unemploy-

TABLE 1.
Differences Between June 1981 Family Allowance Counts and 1981 Census Counts for the Population 1-14 Years, Census Divisions by Province.

Province	Number of Divisions	Absolute Percentage Difference		
		3.0-4.9%	5.0-9.9%	10% +
Newfoundland	10	2	0	0
Prince Edward Island	3	1	1	0
Nova Scotia	18	3	5	1
New Brunswick	15	5	2	3
Quebec	76	8	5	2
Ontario	53	6	7	0
Manitoba	23	1	5	8
Saskatchewan	18	3	3	1
Alberta	15	2	2	0
British Columbia	29	1	0	2
TOTAL	260	32	30	17

TABLE 2
Distribution of Administrative Records by type of Address

Postal Code Type	Percent			
	Taxation	Family Allowance	Unemployment Insurance	
			February	September
Domestic				
Urban Residential domestic address	60.4	57.8	50.5	59.1
Urban Residential apartment	6.5	3.4	4.0	5.6
Rural Post Office	27.4	33.0	39.8	29.6
Rural Route/Suburban Service	3.6	4.5	4.2	4.1
Convenience				
Business	0.9	0.1	0.2	0.2
Post Office box/ general delivery	1.2	1.2	1.3	1.4
TOTAL	100.0	100.0	100.0	100.0

TABLE 3
Distribution of "Care-of" Addresses on Tax File by Type of Address

Postal Code Type	Percent
Domestic	
Urban residential - domestic address	36.2
Urban residential - apartment	2.5
Rural Post Office	21.2
Rural route/suburban service	2.9
Convenience	
Business	26.8
Post Office box/general delivery	10.4

ment insurance file. This high rural incidence in the unemployment insurance file reflects high winter unemployment and the inclusion of special off season benefits for fishermen.

On all three files the incidence of third party convenience addresses is small, approximately 2%. Note, however, that the incidence of business addresses is highest in the tax file reflecting the use of tax service agencies and accountants. The incidence of convenience addresses as reflected in Table 2 should also be considered as a lower bound since some third party addresses may be residential or small businesses with a residential type postal code.

To further investigate third party addresses, "care of" addresses on the tax file were identified. A 2% random sample of all records for tax year 1980 was considered and the "care of" addresses were identified and analyzed. A total of 2.5% of all records were identified as "care of" addresses and the subset of these records containing valid postal codes (89.5%) were tabulated by delivery mode. The results are shown in Table 3. Nearly 63% of all "care of" addresses were what have been termed domestic addresses. An investigation of these records indicated that these were mainly tax accountants and tax service organizations that were located in a building that had the same postal codes as other residences on the block.

Further analysis revealed that in the 2% sample, three quarters of the postal codes on the "care of" addresses showed up only once and a further 10% showed up twice. This suggests that many of the "care of" addresses may be isolated cases of persons using a convenience address. For most applications these should not be a major problem. On the other hand, one address was estimated to account for over 10,000 taxfilers (a public trustee) and a further 34 addresses, each accounted for more than 1,000 taxfilers. These postal codes would result in more serious misallocation problems although their impact again depends on the applications and the size of the geographic area under consideration.

Recall that when using the postal code it is not possible to identify the "care of" addresses, in particular those corresponding to domestic postal codes. In urban areas, some editing might be done for postal codes for which there is an excessively large number of records. However, in rural areas or for rural routes, little can be done to isolate possible third party addresses.

5.0 CONCLUSIONS

In using administrative records for statistical purposes, the postal code provides the key to geo-coding the records. In using the postal code, a number of factors must be considered, in particular the use of convenience or third party addresses and secondly the problems of assigning codes to desired geographic classification systems. These factors may result in some geographic misallocation, although generally it is difficult to separate geographic coverage problems from other coverage problems that may be due to non-reporting or conceptual differences.

The brief investigation of three main administrative files suggests the use of third party address is not a major widespread problem, although it may be important in specific small areas, for example a downtown census tract. Of much more importance is the problem of assigning or mapping postal codes to small areas. Although two thirds of the records are in urban areas that can be easily assigned, the remaining records are in rural areas where the postal codes cover a large area often cutting across boundaries of interest. Some indirect evaluations of the accuracy of coding at the census division or county level indicated that even for this relatively large "small area", the problems are significant.

The impact of geographic coding depends on the application. For example, if one is interested in estimating the average income for a geographic area, bias may or may not be introduced, depending on the characteristics of the misallocated population. However, in most cases, the bias is likely to be fairly small. Since the data are consistent over time, postal coded data can also generally be used to identify trends although the estimation of levels may be problematic. If one is developing migration data, the use of third party addresses may cause more significant problems: if the third party moves all clients will also be moved. One must also be careful in comparing data (e.g., calculating rates or ratios) that are from different data sources that have not been coded from mailing address, for example relating counts from administrative data to census data. In this case, the geographic coverage bias may result in spuriously high or low values. A similar problem may arise if one is relating data from different sources that have differential bias due to postal coding. It has also been suggested that census data could be tabulated by postal code, where the postal code was obtained not from the respondent but rather by mapping standard geographic coordinates of residence address to postal codes. If this approach is adopted, the resulting census data may have a differential bias than say, tax records where the code is obtained from the mailing address, and one must be careful in relating data from such sources.

One possible approach to overcome some of the geographic coding bias is to introduce some type of benchmarking or correction of the data coded for mailing address. In general, such an approach would probably be a net adjustment for all differences i.e., coverage, conceptual and geographical.

A second possible approach would be to obtain additional data on residence. For example, for the development of small area population and income estimates, the U.S. Bureau of the Census periodically sponsors an additional question on the Federal tax return, asking each tax filer to indicate the name of the city and county in which he resides. The responses are then used to create a "coding guide," to allocate residential address. A drawback of this approach is the expense involved. An alternative to collecting additional data may be to link different record sets and use the address from the one that is more likely to refer to residence address. For example, the family allowance records are likely to have "cleaner" address information than the tax records.

In summary, the postal code provides a key to the development of small area statistical data from administrative records. However, the postal code has a number of limitations that must be considered. Perhaps the most important of these is the inability to provide a fine level of coding in rural areas. Moreover, the use of postal codes requires an ongoing effort to construct, update and evaluate conversion files and to assess the implications of third party and convenience mailing addresses.

FOOTNOTE

(1) Postal code directories are published by Canada Post to assist people in finding postal codes.