Can a Geographic Sort Improve Hot Deck Donor Imputation in the Canadian Census?

Darryl Janes
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6, Canada

Abstract

The quality of donor imputation relies upon the content of the chosen donors. The search for donors can be computationally expensive for an imputation system, so it is beneficial to find good donors quickly. A good donor usually has certain characteristics in common with the record requiring imputation. The Canadian Census Edit and Imputation System (CANCEIS) imputes data at the dwelling, family, or person level, and uses a ripple search in stages to find its donors. In the Canadian Census, data is sorted geographically so that nearby households or persons are the first to be considered as donors in hopes of improving imputation quality. This paper will examine the effect of the geographic sort (GEOSORT) in the Canadian Census, and what impact it may have on hot deck donor imputation results.

KEY WORDS: CANCEIS, Donor Imputation, Geographic Sort

1. Introduction

Edit and Imputation (E&I) is a key process for the Canadian Census in order to attempt to improve the quality of census data. The order in which data is sorted or indexed on the database can greatly impact the results of donor imputation. A series of programs were developed for the 2006 Canadian Census to allow the database to be rearranged geographically. As a result, households that were physically close to each other would usually be listed close to each other on the database. This paper will show how this geographic sort process, known as *Geosort,* was developed and how it could help with the edit and imputation process in the Canadian Census.

1.1 Introduction to Edit and Imputation in the Canadian Census

A substantial amount of effort is put into detecting and correcting invalid and inconsistent data before results are made available to the public. Invalid data are any values deemed unacceptable as defined for their subject matter area (i.e. age = 200). For many census variables, non-response is entered into the database as some value that will be considered invalid and later imputed. Inconsistent data are those where some pre-determined edit rules (or *edits*) are violated. For example, if we have the rule that a mother must be at least 15 years old, then a 5-year-old mother would be a case of inconsistent data.

1.2 Introduction to CANCEIS

The Canadian Census Edit and Imputation System (CANCEIS) is software developed within Statistics Canada to correct invalid and inconsistent census data. Janes and Bankier (2004), and Benjamin (2006) describe some of the advances in CANCEIS as well as some of the challenges faced during the 2006 E&I process. For the first time in 2006, CANCEIS was used to perform edit and imputation on nearly all data in the Canadian Census, so the software had to evolve to handle the challenges of these data.

We will define a “record” as simply a collection of data for a dwelling, family, or person. Good records that do not have invalid or inconsistent data will be known as *passed records* because they pass the edits. Records with problems with their data will be called *failed records*, and will require imputation. Although CANCEIS can perform deterministic imputation, many of the corrections are done by applying *hot-deck donor imputation*. This occurs when data is borrowed from good records called *donors* within the same data set to correct invalid or inconsistent data in the failed records. Donors are usually a subset of the passed records.

2. The Importance of Donors

2.1 CANCEIS Donor Imputation

CANCEIS applies the Nearest-Neighbour Imputation Methodology (NIM) during the E&I process. The methodology is described by Bankier (1999), but a few of the relevant points will be provided below. Tables 1 and 2 below use a very small example involving only 5 records (persons) and 3 variables to show how donors are used in CANCEIS.

Table 1: Before Donor Imputation

Person #	Age	Sex	Marital Status
1	36	Male	Married
2	3	Male	Married
3	20	---	Single
4	15	Male	Single
5	22	Female	Single

We see in Table 1 that person 2 is only 3 years old and married, and we are also missing the sex of person 3. Since problems exist in records 2 and 3, the remaining persons (1, 4, and 5) can be considered as *potential donors*. We know that person 3 is 20 years old and single, so person 5, who is 22 and also single, would make a good donor. Therefore, we use this donor's sex as the missing sex for person 3. Similarly, person 1 would be a good donor for person 2. Here are the imputed results:

Table 2: After Donor Imputation

Person #	Age	Sex	Marital Status	Status
1	36	Male	Married	Donor for #2
2	*36*	Male	Married	Imputed
3	20	*Female*	Single	Imputed
4	15	Male	Single	Non-donor
5	22	Female	Single	Donor for #3

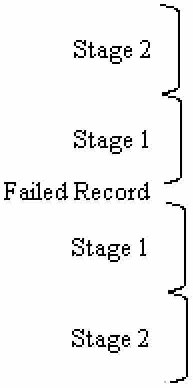
In practice, there are usually more variables, often resulting in many edit rules. Data sets usually consist of tens of thousands or even millions of records, depending on the stratum being processed. For example, if data is stratified by the number of persons in the household, then we can expect many more records in a stratum of 2-person households than a stratum of 8-person households. As the number of records and the number of edit rules increase, the search for donors can get more complex and time consuming.

2.2 The Search for Donors

As we did in the example in tables 1 and 2, we try to find donors that are similar to the failing records. If CANCEIS were to evaluate every potential donor in the data set, it could take hours or even days. There needs to be a balance between the quality of the donor and the time used for searching. CANCEIS is designed to find the best *suitable* donors within a limited search area. It is up to the user to determine what characteristics make a donor suitable. The user must also set up appropriate variable weights, matching criteria, and other CANCEIS parameters.

CANCEIS imputes one record at a time as if the records were listed one after another in a file. Potential donors are evaluated by starting at the failed record and working outward in an alternating fashion (forward, backward, forward, etc), called a *ripple search*. The closest potential donors in the list constitute the first *stage*. The next group of potential donors makes up the second stage, and so on. Stages usually contain a few hundred or thousand potential donors. Figure 1 illustrates how sets of potential donors are evaluated.

Figure 1: Stages in CANCEIS donor search.



The stage sizes and the number of stages, which are both controlled by the user, have an enormous impact on processing time. Since very good donors can be found within 2 stages in most applications, we can usually limit the donor search to the potential donors closest to the failing record, and thus greatly reducing time required to evaluate donors.

Limiting the donor search to only a few stages often means that less than 1% of the potential donors in the data set are examined for any failed record, so it is important for the sake of imputation quality that there is at least one suitable donor in the group. In order to increase the chances of having suitable donors available in the search area, we would like to order the data set so that the ripple search includes potential donors that are similar to the failing record, but not having the invalid or inconsistent data.

2.3 Restricting the Search to Neighbourhoods

A long-standing assumption during donor imputation is that households within the same neighbourhood as the failing household are generally the best donors. Like the proverb, "Birds of a feather flock together", we often make the same general assumption about the people living on nearby streets. In many cases, any two neighbours of close physical proximity have dwellings of approximately the same size, age and type. Furthermore, the same schools, recreation, shopping, and job opportunities are usually available to both. Sands and Griffin (2006), and Thibaudeau (2002) discuss experiences in donor imputation within neighbourhoods, and Collins et al. (2006) discuss how city neighbourhood crimes are related to income and employment.

Hot-deck donor imputation is performed on many variables over 23 subject matter areas in the Canadian Census. We believe that many of these variables would actually have better imputation by searching for donors within the immediate neighbourhood rather than having a somewhat random search over a large region. Therefore, we will assume for the moment that the majority of the variables will benefit by arranging the database so that the donor search focuses on neighbourhoods or areas that are physically close to each failing record being imputed.

2.4 Geographic Composition and Coding

Canada is made up of 13 provinces and territories (PR), with each of these being made up of Census Divisions (CD), which are equivalent to counties. Each CD is made up by Census Subdivisions (CSD), which are municipalities. Each CSD is made up of many Dissemination Areas (DA), and these are built from DA Blocks. Table 3 lists each of these areas in their hierarchical order.

Table 3: Geographic Areas in Canada

Area	Quantity
Provinces and Territories (PR)	13
Census Divisions (CD)	288
Census Subdivisions (CSD)	5,418
Dissemination Areas (DA)	54,626
Dissemination Blocks (DABLK)	478,831
Total	539,176

Each one of these geographic areas is assigned a code of a particular format. For the rest of this paper, any two geographic areas of a certain type that share a common boundary of non-zero length will be considered to be *adjacent*. Areas that only share a single common point will not be considered adjacent. Once a code has been assigned to one geographic area, the next area in the coding sequence will usually be adjacent. In fact, some geographic areas are coded according to a serpentine back-and-forth fashion from southeast to northwest. Because of this, if we sort the database by increasing PR code, then by CD code, CSD code, DA code, and DABLK code sequentially, the database should automatically be arranged so that the donor search will be kept within the neighbourhood.

2.5 Why not Sort the Database by Codes?

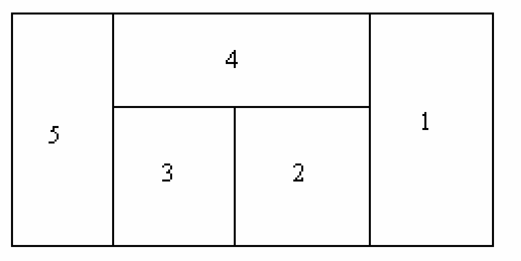
Although the coding of geographic areas appears to be an adequate method for keeping households physically close together also close on the database, there are a few flaws that may reduce its effectiveness. For example:

(a) There is no guarantee that the assignment of codes will consider adjacency for all levels of geographic areas.

(b) Area splits and amalgamations between censuses without recoding could considerably disorganize the sort order.

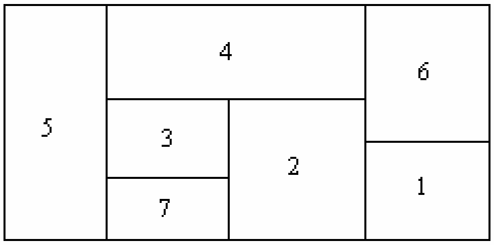
Figures 2 and 3 illustrate how a split could cause a problem. In Figure 2, we see areas coded sequentially from 1 to 5. The order of these areas is acceptable.

Figure 2: Before an area split



Now suppose that area 1 is split into two parts, now labeled 1 and 6, and suppose that area 3 has been split into two parts now coded 3 and 7. We can see that traveling to the individual areas in sequential order in Figure 3 would be inefficient because sequentially numbered areas are no longer adjacent.

Figure 3: After splitting areas 1 and 3



(c) A southeast to northwest serpentine sort may be effective for each level of geographic area, but ineffective when we consider multiple layers.

Figure 4 shows an upper layer geography sorted nicely from southeast to northwest, but Figure 5 adds a sub-layer geography, which is also sorted southeast to northwest. When each of these areas is also sorted from southeast to northwest, we immediately see that the end of one area is often nowhere near the start of the next area.

Figure 4: Upper level SE to NW sort

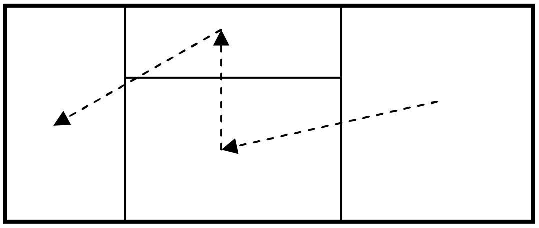
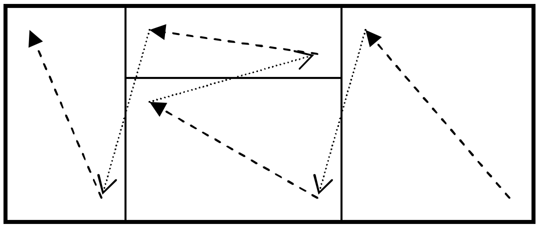


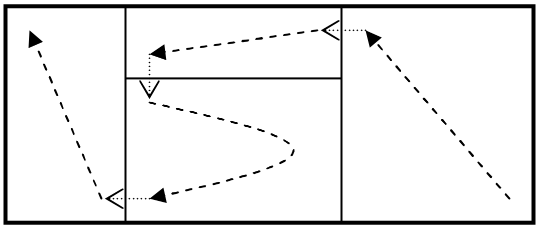
Figure 5: Dual-layer SE to NW sort



For someone who had to canvass all households in the areas in Figure 5, the multi-layer southeast to northwest sort would be very inefficient because of the distance traveled from the end of one area to the start of the next area.

A better solution would be to minimize the transition distances so that the end of one area is adjacent to the start of another area, and to remove the constraint of a southeast to northwest directional sort. Figure 6 shows an improvement of the situation in Figure 5.

Figure 6: Improved Sort Method



In Figure 6, we have modified the upper layer sort and significantly shortened the transition between adjacent areas. The first and fourth areas continue to sort southeast to northwest, but the second area heads southwest, and the third area starts in the northwest and ends in the southwest.

While the illustrations in Figures 2-6 are very simple, they suggest that we could benefit from a separate sort procedure that does not rely solely on geographic codes.

3. The 2006 Geographic Sort (Geosort)

For the 2006 Canadian Census, it was decided to build a series of programs (in C and SAS) that would create an index variable to be stored on the database. This *Geosort* variable would allow a database sort by geographic areas down to the DA Block level. Here are some points that were considered:

- Sort by PR-CD-CSD-DA-DABLK by physical proximity, not by codes.
- The Geosort index variable would be unique at the DABLK level, and numbered sequentially from 1 to 478,831.
- Keep adjacent geographic areas together as much as possible.
- Do not restrict the direction of the sort.
- The last area of one upper level area should be adjacent to the first area of the next sorted upper level area whenever possible.

The program requires information about each geographic area at all levels in order to know its geographic positioning and that of neighbouring geographic areas. Boundary and polygon information was obtained for each basic block throughout the country and then aggregated as required to have geographic information for all 539,176 areas. For each area, we would know the latitude and longitude or the Lambert projection coordinates of its central location. We would also know the size of each area, the areas to which it is adjacent, and the length of the corresponding boundaries.

3.1 Start and End Points

Provinces and territories are the first areas to sort. The codes for these are already in an adequate order, so no adjustment is necessary. The next layer to be sorted is the group of Census Divisions within each province. First we identify a pair of adjacent census divisions at each provincial boundary. This way, each province and territory has both a start and end CD such that the end CD is adjacent to the start CD of the next province. The choice of CDs is the only manual operation, since all remaining sorting and lower level operations are done automatically.

The start and end areas of the lower level geographies are only determined after the upper level geography has been completely sorted. For example, the start and end CSDs within the CDs are only determined after completely sorting the CDs. Suppose we know the sort order of the CDs such that they are labeled CD1, CD2, etc. Each CD will be assigned a start CSD and an end CSD. The start CSD of CD1 is the one on the far edge away from CD2. The end CSD of CD1 is the one that is the farthest one from the start CSD and also adjacent to CD2. The start CSD of CD2 is the one that shares the longest boundary with the end CSD of CD1. The end CSD of CD2 is the farthest from the start CSD of CD2 and adjacent to CD3, and so on. Figure 8 illustrates this.

3.2 Sorting Between Start and End Points

Once the start and end CDs have been chosen, there are several ways that the remaining CDs could be ordered. The goal is to minimize the total distance traveled from the start CD to the end CD while covering all CDs in between and considering adjacent geographic areas. Simply taking the start CD and finding the closest remaining CD repetitively until all CDs are used is a simple technique, but it reduces the quality of the sort toward the end. It was determined that the more complex *Minimum Spanning Tree* technique, was a better alternative and would be used.

The following summarizes the algorithm of the Minimum Spanning Tree Method adapted into the Geosort process. Assume that there are *n* geographic areas (polygons) to sort, and the start and end areas are known and fixed in positions [1] and [*n*] respectively. Each of the *n* polygons has a central coordinate (x_i , y_i), where x is a Lambert X coordinate and y is a Lambert Y coordinate.

1) Calculate the distance between all pairs of adjacent polygons.

$$D_{ij} = ((x_j - x_i)^2 + (y_j - y_i)^2)^{½}$$

2) For each polygon, locate the closest adjacent polygon. The connection will form one of the spanning tree branches.

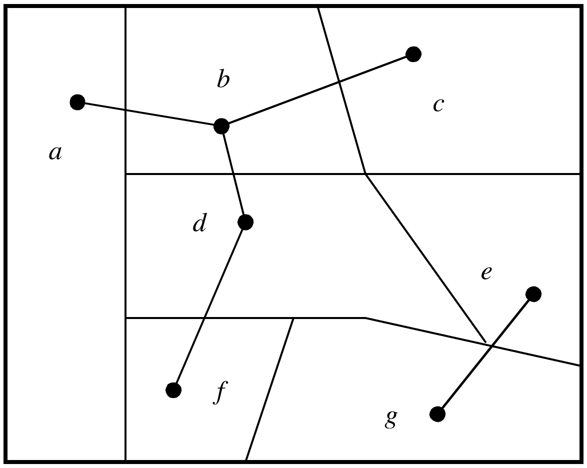
3) Beginning at the start polygon, follow the tree branches, working toward the end polygon.

4) Whenever there is an intersection (choice of branches), choose the path that gives the shortest possible path including subsequent paths. Once the end of the path is reached, back up to the last intersection with a path that hasn't been taken, and take that path.

5) If there is a gap between branches, jump from the end of the first branch to the closest point of the closest disjoint branch.

The following is an example of the spanning tree. Suppose that there are 7 polygons, labeled *a*-*g*, as shown in Figure 7 with their central points represented by the large dot. We see that *a* is adjacent to *b, d,* and *f,* but it is closest to *b*. So the combination *(a,b)* is a branch in the minimum span tree. We also observe the following branches: *(b,d), (c,b), (d,b), (e,g), (f,d),* and *(g,e)*. These branches are drawn as lines on Figure 7.

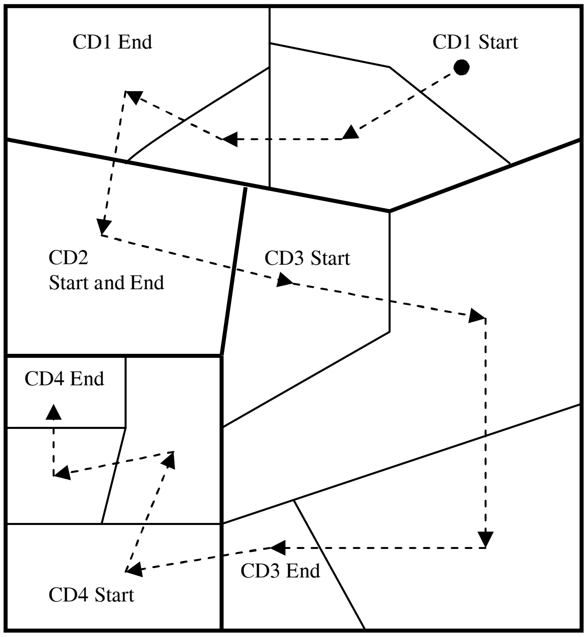
Figure 7: Example of minimum spanning tree branches



Suppose we start at polygon *a*. We move to *b* and have to make a decision between *c* and *d*. Since going through *d* means that we must go through *f* as well, and this is longer than going through *c*, we go through *c* first, then *d,* then *f.* We then jump from *f* to *g* and finally *e.*

Figure 8 shows four CDs (with bold boundary lines) that have been sorted and labeled CD1 to CD4. It also shows how CSDs (thinner boundary lines) within these CDs have been subsequently sorted.

Figure 8: Sorting CDs and CSDs



By starting at the first CSD in CD1 (in the northeast), we can establish a sort order of the 2 CSDs within CD1 before reaching the end CSD of CD1 in the northwest. We then jump to CD2, which is comprised of a single CSD. CD3 is next, and the start CSD is the one adjacent to CD2. The sort continues until all 13 CSDs have been sorted from the start CSD of CD1 to the end CSD of CD4.

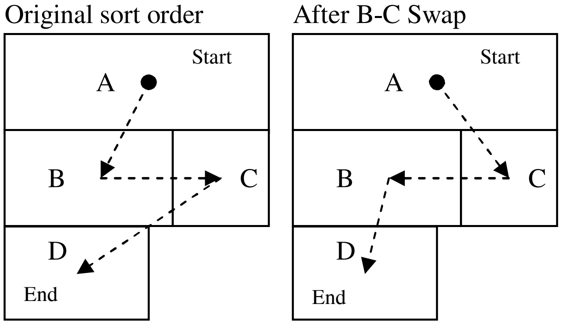
3.3 Successive Improvement Technique

The Spanning Tree technique is advantageous because it considers adjacency and generally works its way away from the starting area without backing itself into a corner too often. However, despite its generally satisfactory results overall, its algorithm is not perfect. In particular, it is not always efficient when 2 or more spanning tree branches are of similar length. As a result, a *Successive Improvement Technique* is applied after the spanning tree sort.

The premise of the Successive Improvement Technique is to locate sequences of geographic areas that would benefit from swapping two or more areas in the sort order. In other words, by swapping the order of two or more geographic areas, the distance traveled in sequence would be reduced. For the 2006 Census, it was decided to iteratively consider all sequences of four consecutive sorted geographic areas going from the starting geographic area to the end geographic area. This restricted the start and end areas from being swapped.

Figure 9 illustrates how a sort order might be improved by swapping the order of two geographic areas. Suppose that we have four geographic areas – the start and end areas have been identified as areas A and D respectively. It leaves us to determine if the sort order should be A-B-C-D or A-C-B-D. Since the central point of A is closer to B than to C, and since both B and C are adjacent to A, the system chooses B to follow A in the sort order. Therefore, we have an A-B-C-D sort. However, the Successive Improvement Technique would determine that by swapping C and B, the distance traveled between central points of an A-C-B-D sort would be less than the original sort.

Figure 9: Successive Improvement Technique

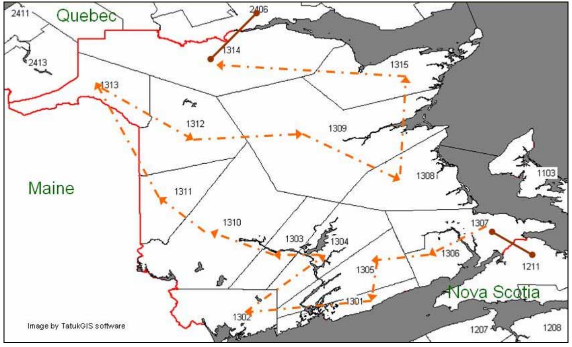


3.4 Results from the 2006 Canadian Census

The Geosort process was successfully completed in advance of doing the edit and imputation of all of the variables in the census. The resulting Geosort index variable was assigned to the database for the 478,831 DA Blocks. Nearly all subject matter areas used the Geosort variable in their E&I process.

Figure 10 shows how the CDs in the province of New Brunswick have been sorted. The codes have been provided to see how a sort by codes would look. We immediately see that the CD with the lowest code, 1301, is not bordering on the Nova Scotia boundary. Similarly, the CD with the highest code, 1315, is not bordering on the Quebec boundary. By not sorting by codes, we were able to choose 1307 as the first CD and 1314 as the end CD within the province, and then sort the remaining 13 CDs. The arrows indicate the sort order.

Figure 10: A Sort of CDs in New Brunswick



While the geographic sort gave satisfactory results, there were occasions where a better sort order could have been produced. In Figure 10, there were 2 instances where a non-adjacent CD was chosen (1313 and 1315). Although we try to minimize the frequency of such occurrences or reduce the size of the jump, they are often difficult to eliminate entirely, particularly without manual intervention.

4. Planning towards the 2011 Census

After the 2006 Census edit and imputation period ends late in 2007, planning for the 2011 Census begins. We will evaluate the results and experiences from the Geosort as part of the planning process. The Geosort system was put in place in 2006 with the assumption that donor quality would be improved by sorting the database so that donors would come from the same neighbourhood or nearby neighbourhoods. Since the 2006 CANCEIS imputation process is nearly complete, we will be able to use its results in order to answer several questions such as these listed below.

(1) Can we verify the assumption that good donors can be found in the same neighbourhood?

(2) How do the Geosort results compare to sorting geographic areas by codes?

(3) What changes should be made if the Geographic Sort is to be implemented in 2011?

In general, we want to see if imputation variances with the Geosort are less than those generated by a random sort or by a sort by geographic codes. If the assumption in (1) holds, we then need to verify that the Geosort results are better than a sort by geographic codes. Besides imputation quality, it will be interesting to compare statistics on the sort paths. Here are a few basic criteria that may be considered when trying to find a better sort path:

(a) Minimize the total distance (or squared distances) from the start area to the end area.

(b) Same as (a) above, but by considering a lag effect. For example, add the distances between the first and the third area, the second and the fourth, the third and the fifth, etc.

(c) Maximize the number of times that an adjacent area was taken as the next area in the sort.

(d) Maximize the total length of boundaries between adjacent areas in the sort.

Sorting the database by geographic code is a very simple and quick process. If it is found that the Geosort does not significantly improve imputation results, then the complex Geosort programs may need to be improved, or they may be replaced by a code sort.

One obvious way to improve the Geosort is to introduce auxiliary variables such as average income or shelter cost for an area. By minimizing the transition rate of these types of variables, there will be more continuity and homogeneity along the sort path. Another improvement would be to make the Successive Improvement Technique more complex and to consider more than four areas at a time. The improvements, however, should only be implemented if their added complexity is compensated by the increase in imputation quality.

5. Conclusion

The development of a geographic sort procedure for the E&I database was an important project in the 2006 Canadian Census. It has been used by nearly all subject matter areas in the donor imputation process, and it has even been used by the whole household imputation process that treats census non-response. Results from the Geosort are satisfactory, but it is a complex and time-consuming process. As a result, the donor imputation results from the 2006 Census will be used to determine the advantage of the Geosort as well as the role it will play in the 2011 Census.

References

Bankier, M. (1999). "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy.

Benjamin, W. (2006). “Enhancements to the 2006 Canadian Census Edit and Imputation System”, ASA Joint Statistical Meetings, Seattle.

Collins, K., C. Babyak, and J. Moloney (2006). “Treatment of Spatial Autocorrelation in Geocoded Crime Data”, ASA Joint Statistical Meetings, Seattle.

Janes, D. and M. Bankier (2004). “Minimum Change Edit and Imputation for the 2006 Canadian Census”, ASA Joint Statistical Meetings, Toronto.

Sands, R.D. and R.A. Griffin (2006). “2010 Census Count Imputation – Research Results using Spatial Modeling”, ASA Joint Statistical Meetings, Seattle.

Thibaudeau, Y. (2002). Model Explicit Item Imputation for Demographic Categories. Survey Methodology, 28, 135-143.