A MULTIVARIATE APPROACH TO RESPONDENT LOCATION A PRELIMINARY ANALYSIS OF RESULTS

Pierre Daoust and Larry Li
Statistics Canada

ABSTRACT

For the 1991 Census of Agriculture, an automated geographic editing system was implemented to locate and link respondents and establishments to enumeration areas(EA). This paper outlines the methodological approach behind the Census of Agriculture Geographic Edit System (CAGE) and provides a preliminary analysis of its performance.

KEYWORDS

Geocoding, geographic referencing, data linkage, automated coding

1.0 INTRODUCTION

Carried out simultaneously to the Census of Population, the 1991 Census of Agriculture collected data on farms and farm operators in Canada. In this process, up to three operators could be reported per farm, although only one of these operators completes the questionnaire. These operators either live on the farm (resident operators), or do not live on the farm (non-resident operators). The status of residence was collected for each operator. Data was also collected on the geographical location of the farm headquarters and on the operators' residences. The farm headquarters is defined as the operator's residence for resident operators and as the location of the farm main buildings or main gate for non-resident operators.

A land description, corresponding to concepts known to the operators, was provided to locate the farm headquarters. Examples of types of land description are Meridian-Range-Township-Quarter(& Section) in the Prairie Provinces, and County-Concession-Township Lot in eastern provinces. Operators sometimes provided a place name instead of a land description when they either did not know the land description or did not understand the question as formulated on the questionnaire. The place name might represent, for example, a municipality. Throughout this paper, the term land desription will be used although it could represent either an actual land description or a place name.

Addresses, postal codes and telephone numbers were gathered to locate operators' residences. It should be noted the operators declaring themselves as resident would not necessarily provide the address, postal code or telephone number of the farm headquarters. They might provide, instead, an address of a second residence.

Like many other surveys and censuses, it is important to link the respondent data to the correct geographic location. For the Census of Agriculture, the farm headquarters and/or an operator's residence are linked to an enumeration area (EAs). An EA represents an area assigned to a Census enumerator for complete enumeration. There were approximately 45,000 EAs in the 1991 Census. Identification of the EA location of farm headquarters was necessary for data processing, data validation and data publication activities. Identification of operators' residence locations was also required to link the farm operators to the respondents in the Census of Population. This linkage enables analysis of various characteristics of the operators, such as their age, sex and ethnic origins.

The EAs corresponding to the farm headquarters and to the residence of the operator completing the questionnaire, known as the drop-off EAs, were initially identified by Census enumerators. For resident operators, which represented the majority of the agricultural operations, this identification process was simple and basically error free. For non-resident operators, however, the identification of the EA corresponding to farm headquarters was more difficult. For some of these cases it was impossible to determine the correct EA based on the reported data. Finally, the EAs associated with the residences of operators not completing a questionnaire were not assigned by enumerators due to the complexity of such a procedure.

Once the data collection activities were completed, all questionnaires were transferred to the central office of Statistics Canada in Ottawa. A series of edits were performed to detect invalid or missing EAs associated with farm headquarters and operators' residences. EA identifiers could be invalid because of data processing errors which might have been introduced during transcription and capture. All cases failing these edits were flagged and sent through an automated geographic edit system known as the Census of Agriculture Geographic Edit (CAGE) system. Cases not resolved by CAGE were edited manually.

The problems of linking the location of farm headquarters or operators' residences can be generically viewed as an imputation problem of trying to find the location of respondent based on common geographic descriptors, such as address, postal code, telephone number and land descriptions. These problems are very similar to those found in migration coding, place of work coding, address register construction and reverse record check operations.

In the following discussion, Section 2 describes the general approaches to solving these problems and in particular the methods implemented in the CAGE system. Section 3 presents a preliminary analysis of results obtained for the 1991 Census of Agriculture. Finally, conclusions and recommendations are made in Section 4.

2.0 COMPETING APPROACHES AND SYSTEM PROCESSES

2.1 Competing Approaches

In tackling the geographic linkage problem, two types of approaches have emerged. Most common are univariate approaches based on postal codes or addresses (Nadwodney, 1989; ESRI, 1989; Mapping Information Systems Corporation, 1989). Univariate approaches based on land descriptions have also been used for rural areas. More recently, multivariate approaches have become prominent in a variety of census and survey processes (Hansen, 1991; Norris and Coyne, 1991; Yergen, 1987; Drew, Armstrong and Dibbs, 1987). Indeed, many of these multivariate approaches have been implemented in record linkage or text coding systems such as GRLS V1(also called CANLINK, Statistics Canada 1989A) and ACTR (Statistics Canada, 1989B). The multivariate approaches generally are more tolerant of errors in the input data. They also achieve a higher rate of correct matches but require substantially more computing resources.

The CAGE system differs from most of record linkage and automated text coding systems in a number of ways. Most importantly, the CAGE approach is based on the observation that the area which is serviced by a given postal code is often somewhat different than the service area of telephone exchanges in the same locale. The same holds true for other types of service areas, such as those associated with a land description or an address. Since the service areas of various geographic descriptors overlap, but are not totally coincident, a respondent having a unique combination of the various geographic descriptors can only be found in the area of intersection of all the descriptors. This subtractive approach of CAGE is somewhat different than the additive point-scoring approach which is typically taken by most record linkage systems.

The second difference is the incorporation of edit rules and imputation procedures to ensure that a minimum level of reliability is met in the selection of EAs for farm headquarters. The imputations procedures were used to select the most probable EA, when the subtractive approach and the edit rules did not result in a unique choice.

The implementation platform of the CAGE system represents another major difference from the other systems. CAGE is implemented on microcomputers with an interactive user interface which allows the edit staff to query the system. This was an important asset for the 1991 Census of Agriculture. In contrast, the record linkage systems at Statistics Canada are typically mainframe based and batch mode oriented. CAGE is more user-friendly, and is far less costly to run. It should be noted however that record linkage systems, GRLS V2 (Statistics Canada, 1991), are being developed for microcomputer applications.

2.2 Description of System Processes

As was mentioned in the introduction, CAGE was executed after a number of edits had been applied to the Census of Agriculture data base. For each questionnaire, these initial edits flagged those EAs for farm headquarters or operators' residences which were invalid (or missing). In cases where more than one person operated a given farm, it was assumed that the first operator was the person associated with the drop-off EA. Consequently, the EAs for the second and/or third operators were treated as missing, and were submitted to CAGE for resolution.

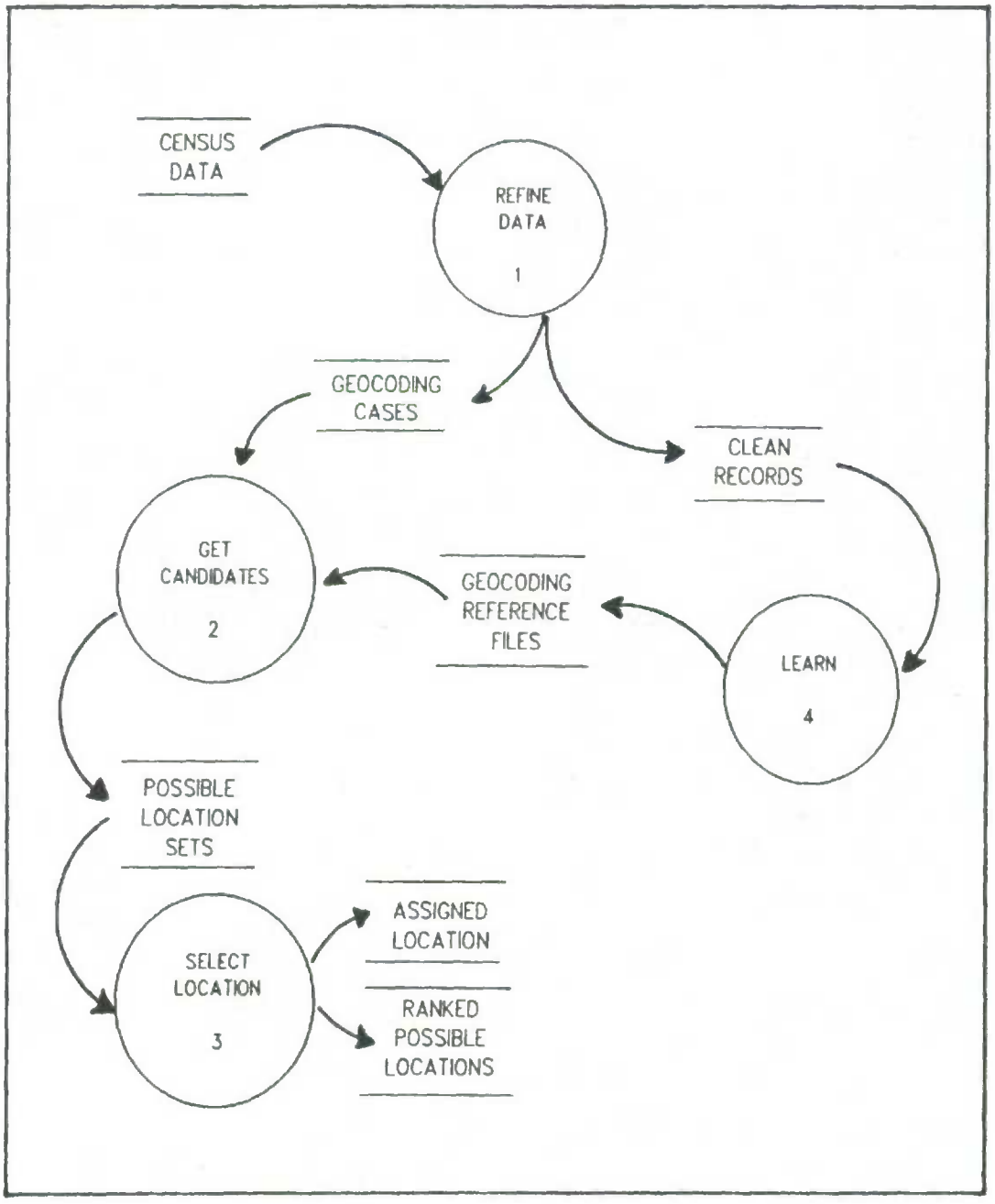
For the CAGE processes, the following data items from the Census were used: the drop-off EA; the headquarters EA; the address, postal code, telephone number and residence status of each operator; and the land description of the farm headquarters. Given these inputs, the task of CAGE was to assign one EA value for each invalid farm headquarters, and a list of possible EAs for each invalid EA for operator's residence. Moreover, CAGE flagged all the questionnaires for which a required edit was not resolved, so that they can be forwarded for manual editing.

CAGE processed independently each questionnaire in the input file. Initially, the system read the edit indicators to determine which EAs were failing an edit. Each edit failure was treated separatly thereafter. Based on the operator's residence status and the type of edit failure being treated, CAGE determined which geographic descriptors (adress, postal cade, telephone number and land description) could be used. When an EA for an operator's residence failed

an edit, CAGE used all geographic descriptors, with the exception of land description for non-resident operators. When an EA for a farm headquarters failed an edit and at least one resident operator was reported, CAGE used the address, postal code and telephone number of the first operator identified as being resident, and the land descriptions. When only non-resident operators were reported, CAGE used only the land description.

For the remaining processes, CAGE always used all geographic descriptors. In cases where the system identified that some geographic descriptors could not be used, it simply treated the descriptors as missing. These processes have been detailed by Li (1990). The following excerpt describes briefly the operations of the four main modules, as illustrated in Diagram 1: (1) Refine Data, (2) Get Candidate Locations, (3) Select Location, and (4) Learn.

DIAGRAM 1. The main System Modules of CAGE



2.2.1 Refine Data

Given an input record with the operator's address, postal code, telephone exchange number and land description corresponding to the farm headquarters, the system began by refining the input data. Entries for the province and telephone area code were checked and repaired by consulting definitive reference tables. The address was parsed and linked to a postal code by the PCODE software (Statistics Canada, 1987). The land description was standardized via a set of standardization rules which are similar to rules in text recognition systems.

2.2.2 Get Candidate EAs

All possible EAs, called candidate EAs, associated with each geographic descriptor were retreived independently from a set of reference files. Retrieval of EAs with the derived postal code (from the address), the telephone exchange and the reported postal code was straight forward. The look-ups for EAs based on land descriptions required a more elaborate matching strategy which is briefly described below.

First, significant words were retrieved from names reported for each component of the land description. For eastern provinces, these four names are typically for county, township, concession and lot. All the significant words were then used to retrieve possible EAs from a large reference file. On the combined set of possible candidates, the most exact matches were identified for the four components, with consideration of the type of the component, and then without consideration of the type of the component. For example, a match of a county name to a county name was preferred to a match of a county name to a township name. The extent of the match constraints varied depending on the province. The candidate EAs were then compared with each other to determine their logical consistency. For example, the township must be found within the given county. Inconsistent candidate EAs were eliminated.

Candidate EAs not eliminated in this module were passed to the next module "Select Location".

2.2.3 Select Location

This module worked with the four sets of candidate EAs from the "Get Candidates" module. Each set represented the possible EAs where the subject, the farm headquarters or the operator's residence, could be found based on address, postal code, telephone number and land description. The first task of the "Selection Module" was to consider the possible EAs and select those common to a maximum number of geographic descriptors. The second task, applied to the EAs selected in the first task, was to weed out erroneous EAs and select one EA for farm headquarters or to produce a ranked list of EAs for the operator's residence.

The process started by searching for the most restrictive intersection set amongst the four candidate sets. If no common EAs were found in all four candidate sets, then those which were common to three of the four sets were retrieved, and so on. If common locations were not found amongst any of the candidate sets, the system selects from the sets in a user-specified order.

When the subject was the farm headquarters, the system eliminated all EAs which failed one of the edit described in Appendix 1. The final imputation of an EA was done using a random procedure which assigned to each candidate EA a probability of selection proportional to the number of farms it contained. When the subject was the operator's residence, the EAs of the intersection set were ranked by the number of farms. In this module it was assumed that likelihood of finding either a farm headquarters or an operator's residence within an EA is proportional to the number of farms within the given EA, as enumerated in the 1986 Census of Agriculture. In other words, the farm headquarters and operators' residences for which EAs had to be imputed would have the same geographic distribution as the distribution of all farms in 1986 Census.

2.2.4 Learn

Currently, a reference file of land survey units for all of Canada is not available. However, it seems reasonable to try and learn the land description to EA relationships from the census records. The process was modelled after the way a person would acquire such knowledge.

When an EA which is associated with a farm headquarters, passed the geographic edit for farm headquarters, the known portions of the land description were checked against available information in the reference file to weed out erroneous data. Valid information was then passed onto the learning process, where the land description and associated EA were stored. As the same relationship was reported by other operators, the confidence of the relationship being correct was increased. When enough operators have reported a given relationship then it was reasonable to believe that the relationship was true. The relationship was then written to a secondary reference table and used in the "Get Candidate" process.

3.0 ANALYSIS OF 1991 CENSUS OF AGRICULTURE PRODUCTION RESULTS

Data processing using CAGE to determine EAs for farm headquarters and operators' residences for the 1991 Census of Agriculture has just been completed. The following is a preliminary analysis of the results for a portion of the data. From the cases where an EA was be found successfully, the precision of the results was assessed. Precision was defined in terms of number of EAs contained in the final sets from which an EA could be chosen. The accuracy of the results was also of interest, but unfortunately could not be investigated in this study since there was insufficient time to embark on the time intensive process of verifying the true location of the subjects. Furthermore, from the failure cases, i.e. when no EA could be determined, the underlying reasons for non-resolution were examined, and noted as possible topics for future research.

This preliminary analysis focused on the province of Saskatchewan since it had an adequate number of questionnaires requiring CAGE editing, and it was one of the provinces which was considered to be a primary candidate for the CAGE processes. In addition, both the input data and reference files were of high quality. These conditions allowed the analysis to focus on the performance of the CAGE process free from confusion due to extraneous factors, such as incomplete or erroneous reference files. The analytical results would also be informative, as EAs failing the geographical edits are an important problem in Saskatchewan.

This analysis is based on the CAGE results of four samples of the questionnaires which failed the CAGE edits. For each type of edit failure (farm headquarters, operator 1, operator 2 and operator 3) a sample of questionnaires was selected amongst the questionnaires which failed the edit. The four samples of questionnaires were selected as independent systematic samples from those in the two strata: 1) those completely resolved by CAGE, and 2) those which had at least one unresolved edit. The four samples were drawn and pooled together to be resubmited to CAGE.

Due to the possibility of multiple edit failures for the same questionnaire, the edit failure counts are slightly higher than the number of questionnaires originally selected. For example, some of the questionnaires which were selected because they failed the edit for farm headquarters might have also failed the edit for the second operator. Nevertheless the final samples of edit failures were treated as if they had been selected under a stratified simple random sampling design. This assumption might not hold completely if the results from CAGE, when multiple edits failed, were different from the results when only one edit failed. In this case however, such a difference was considered too small to affect the analysis of the study results. The final sample sizes for the four samples of edit failures are presented in Table 1.

TABLE 1. Final Sample Sizes of Edit Failures

Strata	Headquarter	Operator 1	Operator 2	Operator 3	Net Number of Farms
Complete Resolution	133	59	183	73	311
Incomplete Resolution	261	32	170	59	271
Total	394	91	353	132	582

For example, in Table 1, 394 of the 582 questionnaires were selected to analyse cases where the edit for the farm headquarters failed. From the 394 questionnaires, 133 were selected in the first stratum where EAs were derived for all failing edits. 261 questionnaires were selected from the second stratum where CAGE could not resolve at least one of the required EAs.

From a preliminary analysis of the CAGE results, it was determined that the results were very similar for operator 1, 2, and 3. For this reason, these samples were treated as a group in additional analysis.

CAGE had some fundamental differences in the way it tries to determine EAs for operators' residences and farm headquarters depending on the operators' residence status (resident or nonresident). For this reason, resident operators were analyzed separately from non-resident operators. Similarly, farms with at least one resident operator were analysed separately from those with no resident operator.

Some of the important conclusions from the analysis of CAGE results are illustrated in the following figures and discussed below. The results that are presented pertain to resident operators, and to farms where at least one operator was a resident. In these situations the four geographic descriptors (address, postal code, telephone number and land description) could be used to assign an EA. Similar results were obtained for non-resident operators and farms operated by non-resident operators, but these results are not presented to avoid unnecessary duplication. In these cases, one can recall that the land descriptions were not used to locate operator's residence, and that operator's address, postal code and telephone number were not used to locate farm headquarters.

In the results, percentages were obtained by initially weighting the sample questionnaires according to the inverse of their inclusion probabilities. These weights were then adjusted by poststratification in each stratum, where the post-strata were defined according to the residence status. Post-stratification was used to compensate for the slight

distortion of sample representation in ignoring residence status in the original stratification. Quartiles were obtained by deriving the sample quartile in each post-stratum, and then by taking a weighted average of all these quartiles, based on the post-stratum population sizes.

Figures 3.1 to 3.4 describe the results of CAGE for locating operators' residences. In these tables TWP LOT represents land description. It can be seen from Figure 3.1 that all the geographic variables provided very good basis for geographic editing, with match rates of 80 percent or more for each component. With all the components together, in the CAGE multivariate process, a match rate of over 99 percent was obtained. This represents a marginal increase of approximately 2 percent over the next best performing component, the land description.

LOCATING RESIDENCE OF RESIDENT OPERATORS

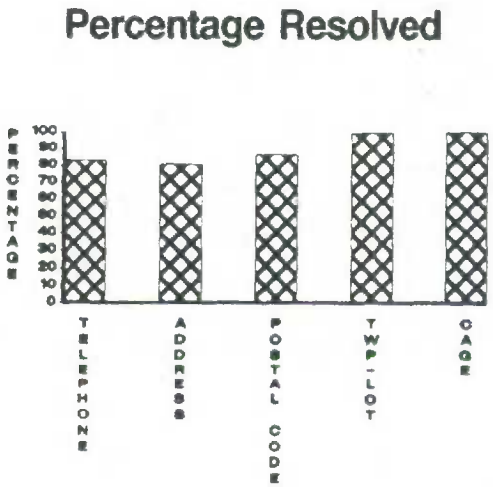


FIGURE 3.1

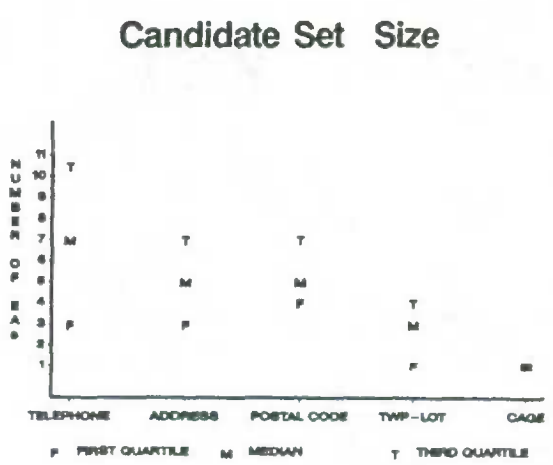


FIGURE 3.2

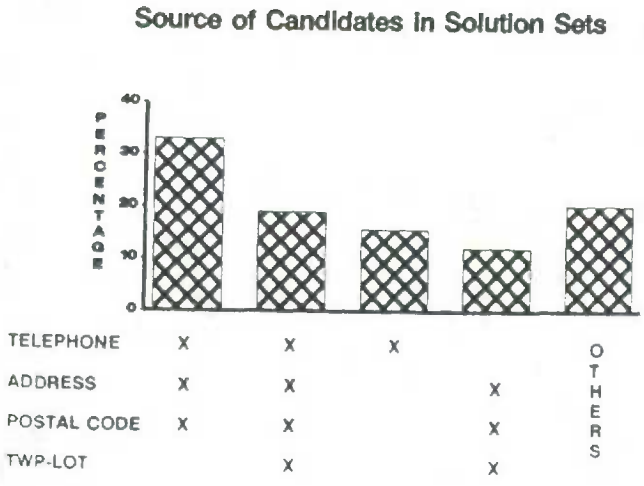


FIGURE 3.3

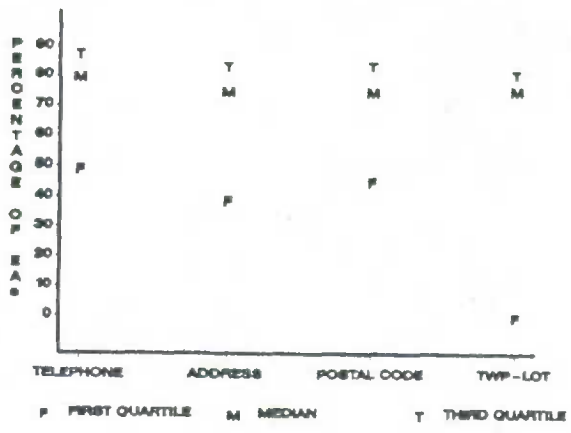


FIGURE 3.4

Figure 3.2 shows the size of the candidate sets which were obtained from the searches using each input variable. For example, the median size of the candidate sets for the searches using the telephone number included seven EAs. With all the components together in the CAGE process, the median candidate set size declined to one EA, representing a very significant improvement in precision.

Figure 3.3 illustrates the source of the candidate EAs, based on geographic descriptors, in the intersection set within CAGE. From the chart, it is clear that there was strong agreement between the candidate EAs retrieved from the various geographic descriptors. There was agreement amongst all the descriptors in approximately 20% of the cases. Agreement amongst the address, postal code and telephone number was noted in over 30% of the cases. The absence of agreement from the land description in these cases indicates a possible discrepancy between the respondents' claims of being a resident operator and the reality of their claims. It could also represent cases where the resident operators reported an address, postal code and telephone number different from the ones of the farm headquarters.

Figure 3.4 shows the percentage of EAs which were retrieved in each of the univariate searches, but were excluded from the solution set via the intersection mechanism in CAGE. At the median point, close to 80% of all the candidate EAs which had been retrieved through the univariate searches were excluded, representing a large reduction in the risk of incorrect EA assignements by CAGE over the univariate approaches.

Figures 3.5 to 3.7 show the results of finding the locations (EAs) for farm headquarters. It should be noted that the candidate EAs within the intersection set were not kept by CAGE and hence were not available for this study. Only the EAs, called winning candidates, finally assigned by CAGE from the intersection sets were available for analysis. Thus, the figures illustrates the characteristics of the winning candidates only. Figures 3.5 and 3.6 show, respectively, the success rate and the candidate set size for CAGE and for each univariate method. The candidate set size for CAGE was obtained by noting how many of winning candidates were chosen with probability 1, i.e. for which candidate set size was 1. Figure 3.7 shows the source of the winning candidates according to the univariate components.

Figures 3.5 and 3.7 indicate that CAGE was extremely precise at the cost of a lower match rate. This increased precision was obtained as a result of applying edit rules to further screen the EAs for the farm headquarters from the intersection set. Figure 3.7 again shows that multiple geographic descriptors contributed frequently in determining winning candidates. It should be noted that one of the edit rules required that the EA selected by CAGE for the farm headquarters had to be part of the set of candidate EAs which were retrieved using the land description. This is why land description appears in all combinations in figure 3.7.

In the cases where CAGE was not able to find at least one EA for a farm headquarters or an operator's residence, the reasons for failure were examined. For locating operators' residences, the key reasons for failure were missing or invalid data associated with the input data and insufficient information in the reference files. For locating farm headquarters, missing and invalid data and incorrect searching methods represented the minor reasons for failure. The major reasons for not retrieving an acceptable EA were invocation of the edit rules, described in appendix 1, such as the candidate EAs for the farm headquarters being too far from the drop-off EA.

LOCATING FARM HEADQUARTERS
OF FARMS WITH RESIDENT OPERATORS

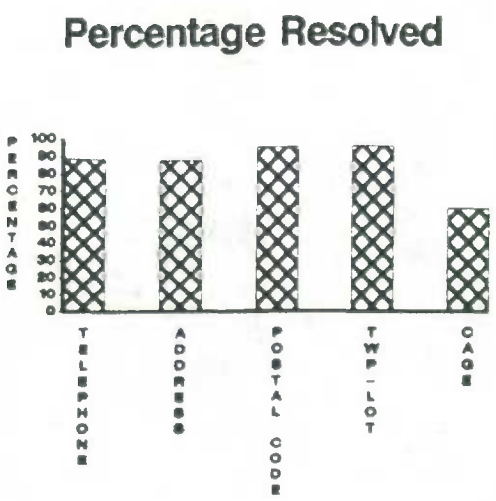


FIGURE 3.5

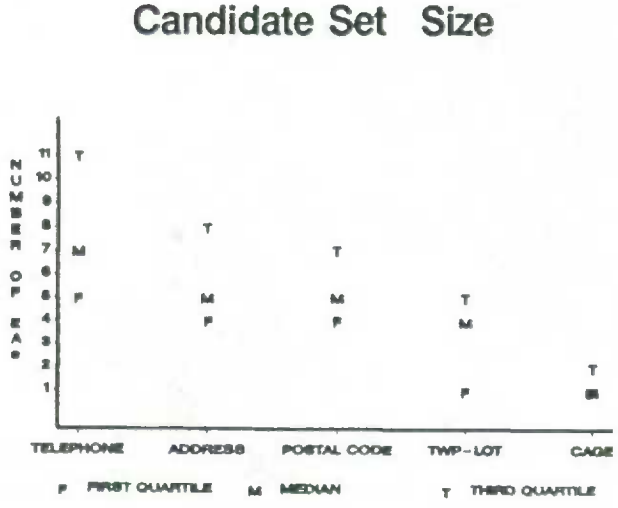


FIGURE 3.6

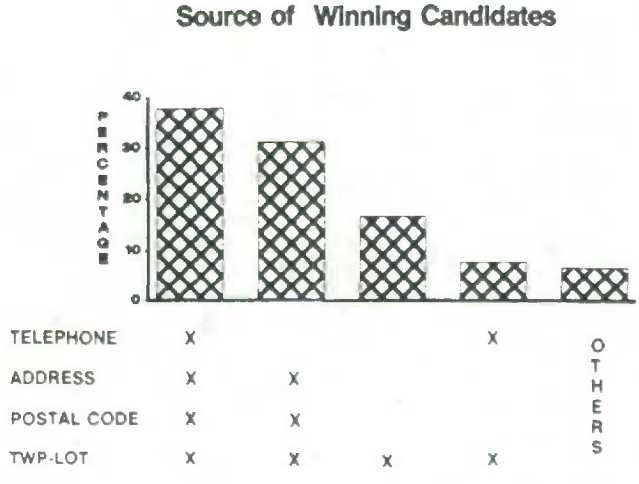


FIGURE 3.7

4.0 CONCLUSIONS AND RECOMMENDATIONS

The results of the analysis of the sampled questionnaires showed that the precision of CAGE for Saskatchewan was very good. The multivariate approach improved the precision of the locational assignments when compared to univariate approaches. In terms of edit success, the improvement was only marginal. The additional edit rules which considered such factors as distance decay and risk of incorrect assignments, were shown to improve even further precision but reduced considerably the success rate.

It is recognized that the performance of CAGE for Saskatchewan does not necessarily reflect its performance for the other provinces. It is expected that the results are similar with the other Prairies provinces. However, in the non-Prairie provinces, it is suspected that CAGE might not perform as well due to the more complex land descriptions, the lower quality response data and the lower quality reference files for land descriptions.

Although it was not possible to evaluate the accuracy of the locations provided by CAGE at this point, it has not been noted as a source of problems so far in downstream validation processes of the Census, except for a few cases in the province of Ontario. In addition, when CAGE was used as an interactive aid in the manual editing process, the staff involved felt that the locations provided by CAGE were very useful in determining headquarters locations. Unfortunately, due to resource restrictions, CAGE was not used on a large scale for this purpose.

From the experience of the 1991 Census of Agriculture, a number of issues were noted for future consideration. Further research into improving the precision of linkages between postal code service areas and EAs would be beneficial. Restructuring of the questionnaire to mnore closely mirror the way in which people describe the location of their farm in various parts of the country would help to improve the quality of the input data. Identification of additional geographic information sources, such as place name databases, which could be added to the reference database would also help to improve edit success, especially for farm headquarters. Refinement of the distance decay rules, perhaps with tailored distance thresholds to reflect the different perception of distance in different parts of the country would also be useful. Further research to improve the real time capture of information and geographic relationships from clean records would be advantageous. This may help reduce the cost of database improvement which is one of the keys to achieving high match rates in the linkage process. Finally, it is recognized that crop patterns and farming practices vary substantially from area to area. It may be possible to integrate these factors in the location determination process, perhaps through the adaptation of donor imputation principles to this process.

ACKNOWLEDGMENTS

The authors would like to thank S. Cheung for his careful review of the preliminary drafts of this paper. His insightful remarks helped greatly to improve the paper.

REFERENCES

DREW, J. D., ARMSTRONG, J. B. and DIBBS, R., 1987; Research into a register of residential addresses for urban areas of Canada, Proceedings of the Annual Meetings of the American Statistical Association, Section on Survey Research Methods, San Francisco, California, Aug. 17-20, 1987, pp 300-305.

ESRI , 1989; ARC/Info network manual. Environment Systems Research Institute, Redlands, California.

HANSEN, K., 1991; Place of birth and migration coding, paper presented at the Annual Meetings of the Southern Demographic Association, October 10-12, 1991, Jacksonville, Florida.

LI, L., 1990; Automated Micro-Level Geocoding for Canada: An Expert System Approach, proceedings of 1990 GIS/LIS Annual Conference and Exposition, November 5-10, Annaheim, California, sponsored by ACSM, ASPRS, AM/FM Int. AAG and URISA.

Mapping Information Systems Corporation, 1989; MapInfo user's guide. Mapping Information Systems Corporation, Troy, New York.

NADWODNEY, RICHARD, 1989; The Canadian Postal Code System and Postal Code Applications. Geography Division, Statistics Canada, Ottawa.

NORRIS, M.J. and COYNE, S., 1991; Automated coding of mobility place name data for the 1991 Census, Proceedings of the Statistics Canada Methodology Symposium (to be published in 1992), November 1991, Ottawa, Canada.

Statistics Canada, 1987; PCODE (Automated Postal Coding System) user guide. Statistics Canada, Ottawa.

Statistics Canada, 1989A; Generalized Iterative Record Linkage System: Concepts. Systems Development Division, Statistics Canada, Ottawa.

Statistics Canada, 1989B; ACTR - Automated Coding and Text Recognition System Manual. Systems Development Division, Statistics Canada, Ottawa.

Statistics Canada, 1991; Generalized Record Linkage System: GRLS V2 Concepts. Systems Development Division, Statistics Canada, Ottawa.

Yergen, W., 1987; 1990 Test Geocoding Experience, in Building on the Past - Shaping the Future, Proceedings of the URISA 25th Annual Conference, vol. II.22

APPENDIX 1

This appendix provides a brief summary of the special edit rules which were applied to the EAs within the intersection set when an EA for a farm headquarters was required.

RULE 1: Every EA in the intersection set which was not found in the candidate set based on the land description would be eliminated. This rule implied that CAGE was used to improve on the univariate approach based only on land description.

RULE 2: The distance between each EA in the intersection set and the drop-off EA was calculated. If the distance obtained exceeded the prescribed threshold value, the EA would be eliminated. The threshold value was empirically determined using data from the past Census. This rule assumed that an operator can only travel a certain distance from his/her home to the farm headquarters in order to operate the farm.

RULE 3: For farms with at least one resident operator, all EAs except the drop-off EA were eliminated when the drop-off EA was in the intersection set. This rule was used because the drop-off EA probably represented the EA where the farm would also be found. It usually represents at least the residence EA of the first operator.

RULE 4: When the intersection set contained more than one EA, all EAs were eliminated if they belonged to different Census Consolidated Subdivisions (CCSs). These cases were rejected for manual review. CCSs represent the smallest areas for which the Census of Agriculture publish agricultural data. This rule was used to ensure that errors in determining EAs for farm headquarters would not affect the accuracy of the published data.

Note: RULE 2 and RULE 3 were applied only when the drop-off EA was valid.

Discussion

Alan Saalfeld
Bureau of the Census

Tying the Topics Together

After commending the authors of the two works on their fine efforts, my first role as discussant should be tying the two papers together to highlight common themes and methods. In some sense, I may have been selected to be discussant for this session because I am supposed to know something about each of the main areas—I direct research groups at the Census Bureau in both Geography and Confidentiality. However, my staffs have always operated independently of each other with very little cross-over of applications. My own initial reaction to these two papers is that they are totally unrelated. However, that is not completely true. Perhaps the common thread of my research and that of these two papers is the highly mathematical (as opposed to statistical) nature of the work. Both areas depend heavily on sound computer science principles and practices. Sound principles, revolving around analyses of computational complexity, tell us to seek to solve these problems with other than exact, closed-form, optimal solutions. Sound practices yield adequate heuristic methods for satisfactory solutions. Let us examine how these ideas apply to each of the two papers in turn.

3-D Cell Suppression

The authors of the paper on 3-D cell suppression strategies did not have time to present the background and history of the problem, but here are some of the highlights. The generalized cell suppression problem for 3-dimensional tables is so computationally difficult that one is forced to settle for approximate solutions to a simplified version of the problem. The authors and their colleagues have shown that even in dimension 1, the problem is NP hard. We know that the cell suppression problem may be modeled as an integer programming problem, but the complexity of all but toy sized IP problems makes it unreasonable to attempt a solution. A linear program formulation approximates the IP model and yields a solution which may fail to be optimal, but even linear programs do not run fast enough to be practical for very large 3-D tables. Linear programs may be used efficiently to check a relatively small suppression pattern for coverage because the size of the variable set in the checking LP is proportional to the number of suppressed cells.

So the problem that we CAN solve is to check that a solution protects the sensitive cells adequately. Not optimally, but adequately. Generating a potential solution to be checked is the art of the heuristics described in the paper. Implementing the checker has given the authors and all of us a tool to measure and compare other heuristics that may be developed in the future.

Multivariate Geographic Resolution

The paper presented by Larry Li proposes some new tools for editing for geographic consistency. My experience with geographic identifiers is that they are susceptible to numerous misclassification errors; and the user or agency responsible for maintaining the correct geographic codes does not have the means of guaranteeing or validating information. If multiple sources are used, the discrepancies cannot be resolved easily or systematically. This operational problem of determining ground truth lends a natural fuzziness to the procedures. This may be further complicated by pseudo-geographic frames: for example, phone exchange areas not mutually exclusive. The Bureau of the Census has tried to deal with a similar classification problem using the GTUB concept: a unique combination of geographic codes identifies a cell in a partition of space. Every point in space belongs to a unique cell that is the only cell carrying the complete set of geographic codes. This attempt to eliminate the matching problem by developing a comprehensive set of codes had its own shortcomings, but, at least in principle, such a comprehensive set would facilitate the proposed multivariate approach.

Two difficulties that were not addressed in the paper that require attention are (1) modeling errors in name/geocode data and (2) resolving matching difficulties. Errors that occur in naming and goecoding are far from being random. Their systematic nature may severely limit usual multivariate analysis techniques. Matching on geographic identifiers is still more of an art than a science for the same reason that perturbations in IDs are likewise not random.

Possible Future Research

Geographic hierarchies are used to develop table stubs. The additive relationships in such tables are quite complex; however, pure hierarchies along stubs present little additional complexity to 2-D tables. Even with hierarchies, 2-D table suppression is successfully handled using a network flow model or an LP. 3-D tables with geographic hierarchies along one dimension do not present much additional complexity. Non-hierarchichal geography is another more interesting and more complicated situation that invites future research.

We presently have a limited notion of a "cycle" structure in 3-D (unions and Boolean sums of cubes) that corresponds to the notion of a closed path in the network flow model applied to 2-D tables. If we could improve our understanding of 3-D cycles, then we could possibly identify better potential solutions to the 3-D problem to be tested by the checker.

Both papers deal with problems that may be attacked through partitioning techniques. Geography is a local characteristic; and most operations require only local update or local examination. Large tables may be protected (suboptimally, albeit) by concentrating on a group of dependent cells. Problems that yield to such divide-and-conquer approaches also are candidates for applying parallel methods and algorithms.

FLOOR DISCUSSION

Laura Zayatz
Bureau of the Census

The first paper, presented by Ram Kumar of the University of Maryland, discussed a method for protecting the confidentiality of data in three-dimensional, non-negative, additive tables by way of cell suppression. Government agencies which publish such data desire to provide certain protection levels to their data while minimizing the sum of the suppressed values. The method is based on mathematical programming and heuristic search. It results in a set of cells chosen for suppression as well as a lower bound for the objective function value which can be used to evaluate the quality of the given solution.

Larry Li of Statistics Canada presented the second paper which discussed a multivariate strategy to be used in the 1991 Census for the geographic editing process. The author described the spatial resolutions of various geographic elements and the quality of geographic data collected through questionnaires. Various editing methods were discussed, and spatial intersection and distance decay were highlighted as methods of improving the accuracy of geographic data. A preliminary analysis of results from the Census of Agriculture was given.

The chair, Brian Greenberg of the Bureau of the Census, remarked that the problems addressed in the two papers were alike in that they both require non-optimal solutions.

Alden Speare stated that he knew that overlapping geographical regions present a problem when using cell suppression as a method of disclosure limitation. He asked if the mathematical programming techniques that Ram Kumar described could be used to add noise to tabular data, rather than to find suppression patterns, and whether or not this would solve the problem of overlapping geographical regions. Ram Kumar stated that mathematical programming is not one of the better methods of adding noise to tabular data. Brian Greenberg stated that although it is possible to use mathematical programming techniques to add noise to tabular data, this does not solve the problem of overlapping geographical regions. Greenberg also noted that the Census Bureau uses cell suppression for all economic data except economic data from the 1990 Decennial Census.

Larry Cox of the Bureau of the Census asked if he was correct in understanding that the mathematical programming technique processes one primary suppression at a time. Ram Kumar stated that a large number of complementary suppressions are chosen at the start of the process, and then each primary suppression is examined individually to see if protection has been achieved. Larry Cox then asked if any attempt was made to choose complementary suppressions that would aid in the protection of more than one primary suppression. Ram Kumar said that, although the researchers have not yet done any work in that direction, they have given the idea some thought and plan to work on it in the future.