Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

A MULTIVARIATE APPROACH TO RESPONDENT LOCATION

L.Li, G. Deecker and P. Daoust[1]

ABSTRACT

This paper reports on the development of a multivariate approach to the problems of locating respondents and data linkage. The paper describes how addresses, postal codes, telephone numbers, place names and legal land descriptions can be used for locating respondents thus linking data to the appropriate geostatistical area. A strategy for combining these spatial elements to improve locational specificity and robustness is detailed. Decision rules to enhance fault tolerance and for tie breaking are outlined. Finally, the potential utility of the multivariate approach to various survey operations and topics for further research are put forth for consideration.

KEY WORDS: Geocoding; Geographic referencing; Data linkage; Automated coding.

1. INTRODUCTION

In census and survey operations, defining the whereabouts or location of a respondent is necessary for correct linkage of the data, data quality evaluation, reconciling the results between different censuses and/or surveys, and a variety of other tasks. The same requirement is true for conversion of administrative data to geostatistical reporting units for statistical report preparation.

Traditionally, the tasks of geographic editing and data linkage have been time consuming, manual processes. Beginning in the late 1960s, tools such as Statistics Canada's Postal Code Conversion File and Area Master File and the U.S. Bureau of the Census' DIME, and now TIGER, files have enabled automation of parts of these tasks. Univariate approaches have been most often used, with the postal code being a common choice as the locational key (Wilkin 1988a; Wilkin 1988b; Nadwodney 1989). More recently, multivariate approaches have attracted interest (Norris and Kirk 1989; Schneider 1987; Yergen 1987), as they offer potential for better spatial resolution and more robustness (*i.e.* tolerance to missing or erroneous data).

This paper reports on research at Statistics Canada on the development of a multivariate approach to the problem of respondent location and data linkage. The paper begins with a discussion of how different geographic elements can be used to locate respondents and link their data to an appropriate geostatistical unit, and then how they can be combined to improve the respondent location process. Alternative approaches are reviewed. Decision rules to resolve uncertainties arising from improper responses are detailed. Figures are provided to illustrate the spatial resolution of univariate and multivariate processes. Finally, the potential utility of the multivariate approach to various survey operations and items for future research are put forth for consideration.

[1] L. Li and G. Deecker, Geography Division, Statistics Canada, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A OT6.
P. Daoust, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A OT6.

2. GEOGRAPHIC ELEMENTS FOR RESPONDENT LOCATION

2.1 Addresses

The respondent's address is one of the most commonly available geographic element in surveys and administrative files. Responses typically reflect the postal address of the respondent. The characteristics of postal addresses are detailed by Deguire (1988). Generally, in urban areas, the postal address refers to a dwelling, although super mailboxes, postal boxes, general delivery and other forms of nonlocationally specific addresses do exist. In rural areas, postal addresses are less informative. Rural route addresses, such as 'RR3, Almonte, Ontario', provides lower spatial resolution. Use of postal boxes and general delivery addresses, which only link the respondent to the post office, is also more prevalent.

In urban areas, addresses can be used in combination with street network files, like Statistics Canada's Area Master File or the US Bureau of Census' DIME and TIGER files, to locate respondents (Yergen,1987,221-230). The process is carried out by matching the address of the respondent to the address range and street name which is attached to each street segment in the street network file. Today, many commercial Geographic Information System (GIS) software packages, like MAPINFO (Mapping Information Systems Corporation 1989, 4-63) and ARC/INFO (ESRI 1989), include specific programs for this task.

In rural areas, use of postal addresses for locating respondents is much more difficult. This is due to the poor spatial resolution/specificity of rural addresses, and the lack of street network files for rural areas in Canada.

Another method of using postal addresses for respondent location is to convert the address to a postal code, and then use tools such as Statistics Canada's Postal Code Conversion File (Geography Division 1989) to link the respondent to an Enumeration Area or several Enumeration Areas. Conversion of postal addresses to postal codes is usually done in a two step process. First, the raw addresses are analyzed to derive an address search key (Deguire 1988; Yergen 1987, 221-231). The second step takes the address search key and retrieves a postal code from the Postal Code Master File, a file of all postal codes from Canada Post. The whole process is similar to going to the post office and looking in the postal code directories to find a postal code for a given address.

2.2 Postal Codes

The postal code, a six character, alphanumeric code is found on a large majority of survey responses and administrative files. It has been widely used as a locational key for survey processing (Nadwodney 1989), data analysis (Wilkin 1988a) and a host of other applications (Maloney 1988; and Nadwodney,1989)

As noted above, a postal code can be used with Statistics Canada's Postal Code Conversion File (PCCF) to link respondents to an Enumeration Area or Enumeration Areas. The PCCF serves as a look-up table to facilitate the conversion.

The spatial resolution of postal codes vary significantly between urban to rural areas. In urban areas, a six digit postal code can be as precise as a block face, a large apartment building or even one floor of an office building (Canada Post 1983, 7). In rural areas, a postal code represents a service area, which can encompass parts of several geostatistical areas. Table 1 shows the comparative resolution of urban and rural postal codes by province.

TABLE 1: AVERAGE NUMBER OF EAs or PART EAs PER POSTAL CODE BY PROVINCE

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	1.04	1.04	1.03	1.03	1.04	1.04	1.03	1.04	1.03	1.02	1.05
Rural	4.31	5.13	6.16	4.92	3.90	4.31	2.99	4.29	3.92	5.18	2.14

2.3 Telephone Numbers

A telephone number is composed of three components, a three digit area code followed by a three digit exchange and then the four digit local. The area code delineates large regions. For example, all of Manitoba is served by the area code, 204. A telephone exchange delimits a relatively small area, usually served by a single telephone switching station, and often equal to a municipality in size, but not necessarily coincident with the municipal boundaries. The four digit local is unique to each telephone customer, except for those served by party lines.

For locating respondents, the area code and exchange are the most useful at this time, since they delimit specific service areas, and each exchange is unique within its respective area code. The local which may facilitate linkage of the respondent to a household holds future promise, but data unavailability and cost limits its present utility at the national level.

In contrast to postal codes, telephone exchanges provide relatively better spatial resolution in the rural area than the urban area. Table 2 shows the average number of EAs for an exchange by province.

TABLE 2: AVERAGE NUMBER OF EAs PER TELEPHONE EXCHANGE BY PROVINCE[2]

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	---------- Data Not Available ----------										
Rural	5.84	4.18	7.96	7.88	---	7.13	5.29	5.22	5.13*		2.30

* for Nova Scotia and P.E.I. together.

2.4 Place Names

A place name, denoting the place of residence of the respondent, is often available from survey or administrative files. Where such places cannot be located as part of an address search or postal code based search, use of the place name directly for respondent location can be tried. The keys to making place names useful for respondent location lie in the development of a place name reference file, to provide the XY coordinates or the appropriate geostatistical unit for each place, and an appropriate parsing and text matching strategy.

A place name reference file, with approximately 160,000 entries, is being compiled at Statistics Canada from previous census data, names associated with standard geostatistical areas, unincorporated places and other relevant sources (Norris and Kirk 1989, 12-13). An additional source which requires further investigation is the Toponomic File from the Geographic Services Division of Energy, Mines and Resources Canada which incorporates official place names from provincial gazetteers.

Place name matching has some similarities to the problems of automated text retrieval, automated coding and record linkage. The large body of literature on automated text retrieval (Ashford and Willet 1988; Saffady 1989; Stanfill and Kahle 1986; Bouchard 1979) offers insights on useful approaches and methodologies including compromises between recall and precision and methods for incorporating fault tolerance into searches. Means of search optimization are examined by Sellis (1988), Wu and Burkhard (1987), Ramamohanarao and Lloyd (1983) and Bouchard (1979).

Place name matching, however, differ substantially from text retrieval from free text documents in that the uniqueness of place names is context sensitive, (a name may only be unique within its locale; sometimes towns

[2] These averages should be treated as approximate estimates, as some Enumeration Areas included more than one exchange, thus were counted more than once. Further the data source for the telephone information was the 1986 Census of Agriculture which does not have many respondents from urban Enumeration Areas. Thus, it does not have full coverage of all telephone exchanges in the urban areas.

with the same name occur quite close together) and aliases are common. In Canada, the process is further complicated by the presence of English, French and aboriginal names.

The work of Norris and Kirk (1989), and the US Bureau of the Census (Schneider 1987) provides useful insights into automated place name coding. Norris and Kirk, using the Automated Coding and Text Recognition system (ACTR) (Development Division 1989) which incorporates an entropy based algorithm for match determination, have achieved success rates of approximately 80% with nation wide samples. Schneider (1987, B-3 to B-6), in some of their preparatory research for their 1990 census reported varying success rates, approximately 87 % for Los Angeles and 44 % for Mississippi.

The current research builds upon the general approaches used by Norris and Kirk (1989), Schneider (1987) and Yergen (1987). The process is as follows: Input names are first parsed, then matches are sought based on the significant words within the name string. Direct matches are then sought for each of the input variables. For variables which are not directly matched, closest matches are retrieved. An intersection set is then defined by cross comparison of the candidates which were retrieved for each input variable. If multiple candidates remain in the intersection file a set of rules can be applied to eliminate unsuitable candidates from the set by considering dissimilarities between the characteristics of the places being matched. A distance decay function is also used to determine the likelihood of each candidate being the place of the respondent, if appropriate.

The spatial resolution of locating respondents through place names varies greatly with the nature of the question which is posed to the respondent. For national applications, such as for migration studies, location of subjects to a municipality seems to be the desired resolution (Norris and Kirk 1989). For census data quality studies and data editing, more specific locations, such as an individual Enumeration Area, may be desired.

In 1986, Canada had a total of 6009 legally incorporated municipalities. Approximately 41% of them encompass only one Enumeration Area. 17% of the municipalities encompass two EAs, 9% includes three EAs, and 33% includes four or more EAs. From these figures, it is easy to understand why municipal names, which are often the same as postal office locations, serve as a useful key for geographic data linkage.

Unincorporated places which are generally sub-municipal entities are noted as part of the census field process. They provide an additional source of information for determining the location of respondents for data linkage.

From some of our tests, municipal and unincorporate place names have yielded deterministic mapping of respondents to an EA in up to 70% of the assignable cases in rural areas, given reasonably error free inputs. In urban areas, where there is a predominance of one to many relationships between a municipality and EAs, deterministics assignments at the EA level are very few.

The results of our research and the experiences of Norris and Kirk (1989) and Schneider (1987, B-7a - B12) indicate that the factors responsible for unsuccessful matches or mismatches include: abbreviations, spelling errors, alias names, entries in to incorrect fields, incomplete names, historical names which have since changed, and ambiguous or non-unique names.

2.5 Legal Land Descriptions

The nature of legal description of land units varies significantly from province to province, within some provinces, and also between urban and rural areas. The heterogeneity reflects the colourful history of Canadian settlement, with the interplay of French and British influences in the east, the more orderly settlement of the Prairies and the strong dominance of the rugged physical environment in British Columbia. This heterogeneity makes it very difficult to formulate a generalized question to obtain legal land descriptions, and to devise a standard structure for capturing respondents' answers for national surveys and censuses.

For example, township-range-section data from the Prairies and the B.C. Peace River District are numeric descriptors with occasional alphabetic modifiers. The descriptions for all other provinces are alphanumeric. Responses from Quebec, the Maritimes and British Columbia are extremely variable. Variations which can be expected include answers which do not follow the expected structure and sequencing of variables, inclusion of street addresses, property identification numbers, town names, seigneury names, lot numbers from plan of subdivisions and others.

In spite of the heterogeneity of the data, the legal land description provides a powerful means of locating respondents in rural areas, since it is an extremely specific descriptor of the location of the respondent, and is commonly known in many rural locales. At the most precise level, an individual ownership parcel can be identified, however, both the compilation cost and the data volumes of the required reference file would be very high for all of Canada. At a more aggregate spatial level, the township concession lot or township-range-section represents a more practical scale for national applications. Township concession lots or its equivalents in eastern Canada are typically 100 to 200 acres. In the Prairies, a township section covers approximately 640 acres or a square mile.

To use the legal land descriptions for respondent location, dual strategies were tested. For the Prairies and the B.C. Peace District, where responses usually adhere to the expected data format, direct matching on the full section-township-range and meridian was often successful. When a direct match is not possible, a partial match without the section and/or the range was attempted.

For some of the other provinces, particularly Quebec and British Columbia, where the input is much more variable, the input position of the response can not considered to be a significant factor in match evaluation. For example, if Oxford was reported in the "County" variable, a match on Oxford County would not be given precedence over Oxford Township. The two competing candidates would be equally considered and other tie breaking rules, incorporating a distance decay function and characteristics of the places, would be applied to eliminate extraneous candidates from consideration.

3. MULTIVARIATE STRATEGY

The above discussion outlines five univariate approaches to the respondent location problem. Combining all or several of the approaches holds promise to make the process more robust, as well as improving the spatial specificity of the assignments by taking advantage of the strengths of one assignment path to cover the weakness in another.

Several researchers have tested multivariate strategies for respondent location or similar problems (Norris and Kirk 1989; Schneider 1987; Yergen 1987; Drew, Armstrong and Dibbs 1987). The current research extends the general approaches used by the above-noted authors to include use of telephone exchanges and legal land descriptions. Error tolerance is introduced into the individual candidate identification process, with the incorporation of rules for checking and refining missing or erroneous telephone area codes and the province designator in the address. Further, use of a separate variable as a probability indicator for tie breaking amongst final candidates is tested for locating farms and farmers for the Census of Agriculture.

The first step in the process is to refine the input data to ensure essential components are available to maximize its utility/suitability for subsequent processes. The refinements are carried out by taking in a respondent's postal address, postal code, telephone number and legal land description. The input data is refined by correcting missing province data, which is critical for address processing, by using the postal code and telephone area code to retrieve the province from a reference file. Telephone area codes are verified against a list of all known area codes from the telephone companies. Incorrect or missing area codes are corrected by using the post office name and province to retrieve the appropriate area code for the respondent from a reference file. Addresses are parsed, and a postal code retrieved for each address using the PCODE program (Research and General Systems Subdivision 1987). Separately reported legal land descriptions are parsed to remove noise and facilitate identification of the most significant components for subsequent matching.

The second step is to retrieve candidate locations/EAs where the respondent may be found via each of the input locational elements - postal code, address, telephone exchange and legal land description. This is accomplished via a series of straight look-ups on a postal code to EA file (the PCCF), telephone exchange to EA file, and county-township-concession-lot to EA file.

Having retrieved up to four sets of candidate locations/EAs, a cross checking process is invoked to identify the candidates which are common to the highest number of the input sets. The resultant intersection set defines the potential locations where the respondent should be found. This cross checking process is useful for weeding out erroneous input, since the process requires that several independent geographic elements point to the same potential locations before they are included in the intersection set.

An indication of the spatial resolution of such a multivariate process is shown in Table 3.

TABLE 3: AVERAGE NUMBER OF EAs ASSOCIATED WITH UNIQUE COMBINATIONS OF POSTAL CODE AND TELEPHONE EXCHANGE NUMBER BY PROVINCE[2]

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	-- Data Not Available --										
Rural	2.65	2.80	2.82	2.81	2.07	2.61	1.99	2.49	2.28	2.93	1.50

Note: Estimate based on data from the 1986 Census of Agriculture, which under represents rural nonfarm areas.

In respondent location applications, a definitive location is often required. For migration studies, the required resolution may be a municipality. For data quality studies, EA accuracy may be desired, and prioritization of EAs in which the respondent may be found is useful for reducing the search space for in-depth manual searches or further automated matching on specific respondent characteristics, such as sex and birth date. Therefore, if multiple candidates are identified in the intersection set, a mechanism/process is required to determine the likelihood of the respondent in each of the candidate locations. The process which has been adopted brings in an associated variable as an indicator of the probability of the respondent being found in the given location. For example, for locating a given farmer, the number of farms in each of the candidate EAs may be used as probability indicator to rank the EA candidate for more stringent searches.

In the future, the performance of the process should be improved by the expansion of the rule set to consider other key characteristics of the respondent in order to better define specific candidate locations with the same characteristics. In the case of the Census of Agriculture, the expanded rule set may consider the crops which are being grown by the farmer and other characteristics such as farm size.

4. MACHINE LEARNING TO IMPROVE SYSTEM PERFORMANCE

Compilation of a reference file of legal land units (township-concession-lots) for Canada is a time consuming and costly process. As part of our research, a process is being developed to mimic the on-the-job knowledge acquisition capability of staff engaged in manual editing.

The process assimilates previously unknown land unit descriptions and their associate EAs from clean records; checks the unknown parts of the geographic descriptors against existing information and keeps a running tally of the number of times each of these relationships is reported. Once the tally exceeds a threshold value, which is being empirically defined from past census data, the given relationship is entered into the land unit to EA reference file to improve its coverage.

Test results indicate that this process works well for areas with relatively clean data, and where there is a relatively large number of respondents present within each of the areal units for which the spatial relationship is to be learned. In areas where the land descriptions are highly variable and responses differ considerably, such as in the province of Quebec, the required number of repetitious responses needed to elevate a particular spatial relationship to the "knowledge base" is seldom met. Further research to improve understanding of the characteristics of the responses, leading to the development of improved parsing and phrase substitution strategies would likely improve the performance of such as "learning" process.

5. CONCLUSIONS

Many areas of survey taking, collection and data analysis need to locate their subjects and link their data to a set of geostatistical areas. The current research has demonstrated that a multivariate approach can improve the spatial resolution of locational edits and improve the robustness of the process. The approach, which is being developed, draws from strategies and methods which are used in automated text retrieval, record linkage and the work of others in multivariate geocoding. Applications have been implemented for the Census of Agriculture geographic edit, and for quality studies within the Census of Population and Housing.

At this time, the implemented systems are still in production, thus definitive results on their performance are not yet available. Preliminary indications are that they are performing very close to expectations. Analysis of these results is planned for the near future.

From our experience so far, a number of opportunities for future research have been identified. Further research is required into the derivation of postal code boundaries for rural areas. Our current lack of detailed knowledge of place names and land description across the nation are also impediments to better system performance. Consultation with field personnel to examine other sources of geographic information and maps which may be available, and the integration of such information into the automated process would also be helpful. Focused research to better our understanding of the relationship between respondents and characteristics of various geographic locations and appropriate distance decay functions for various interactions are needed for expansion of the rules for elimination of extraneous candidates. On this latter topic, adoption of some of the concepts and approaches inherent in donor imputation may prove valuable. Investigation of artificial intelligence capabilities may also hold promise for development and enhancement of rules in real time leading to overall improvements in the processes.

ACKNOWLEDGEMENTS

The majority of this research was conducted under the Statistics Canada's Internal Sabbatical Program. Additional funding and data was also provided by the Census of Agriculture. Heartfelt thanks are due to Catherine Cromey, Marcelle Dion, Dave Dolson, Doug Drew, Rennie Molnar, Mary-Jane Norris and Peter Schut for their advice and guidance. The hard work of the research and system development team under Danny Wall and Joe Gomboc were also instrumental to the achievements. The authors gratefully acknowledge all of their support and contributions.

REFERENCES

Ashford, J., and Willet, P. (1988). Text Retrieval and Document Databases, Chartwell-Bratt.

Bouchard, D., (1979). Combined top-down and bottom-up algorithms for using context in text recognition. Unpublished MSc. thesis, McGill Univ., Dept. of Computer Science, Montreal.

Canada Post (1983). *Postal Code Manual*. Canada Post Corporation, Operational Services Directorate, Mail Collection and Delivery Branch.

Deguire, Y. (1988). Postal address analysis. *Survey Methodology*, 14, 2, 317-325.

Development Division (1989). *ACTR - Automated Coding and Text Recognition System Manual*. Statistics Canada, Methodology and Informatics Branch, Development Division, General Systems Subdivision, Ottawa.

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, 300-305.

Dueker, K.J., (1974). Urban geocoding. *Annals of the Association of American Geographers*, 64, 2.

ESRI (1989). *ARC/Info network manual.* Environment Systems Research Institute, Redlands, California.

Geography Division (1989a). *Detailed user guide, Postal Code Conversion File*, January 1989 Version. Statistics Canada, Ottawa.

Geography Division (1989b). *AMF user's guide*. Statistics Canada, Ottawa.

Giles, P. (1988). A model for generalized edit and imputation of survey data, *The Canadian Journal of Statistics*, 16, supplement, 57-73.

Hart, S.A. (1983). The development and use of postcodes for population information systems, in Jones H. (ed.) *Population change in contemporary Scotland*. Geo Books, Norwich, England. ISBN-0-86094-153-1.

Mapping Information Systems Corporation (1989). *MapInfo user's guide*. Mapping Information Systems Corporation, Troy, New York.

Nadwodney, R. (1989). *The canadian postal code system and postal code applications*. Geography Division, Statistics Canada, Ottawa.

Norris, M.J., and Kirk, J. (1989). Research and testing of an automated coding system for the mobility status variable using the ACTR System, analysis report. Internal report, Statistics Canada, Demography Division, 1991 Census Automated Coding Research Task. Ottawa.

Ramamohanarao, K., Loyld, J.W., and Thom, J.A. (1983). Partial-match retrieval using hashing and descriptors. *ACM Transactions on Database Systems*, 8, 4, 552-576.

Research and General Systems Subdivision (1987). *PCODE (Automated Postal Coding System) user guide*. Statistics Canada, Ottawa.

Saffady, W. (1989). *Text storage and retrieval systems*, Meckler Corp., London.

Schneider, P.J. (July 22, 1987). Memo to Robert W. Marx on "1986 Test Census: Analysis of Migration Coding and place-of-Work Workplace File Sources". US Bureau of the Census, Population Division.

Sellis, T.K. (1988). Multiple-query optimization, *ACM Transactions on Database Systems*, 13, 1, 23-52.

Stanfil, C., and Kahle, B. (1986). Parallel free-text search on the connection machine system, *Computing Practices*, 29, 12, 1229-1239.

Wilkins, R. (1988a). Using postal codes for analysis of socio-economic inequities in health outcomes, paper presented at the 14th Annual Health Administration Forum, University of Ottawa, Ottawa, August 1988.

Wilkins, R., and MacDonald, R. (1988). *Potential uses of postal codes with vital statistics data*. Health Division, Statistics Canada, Ottawa.

Wu, C.T., and Burkhard, W.A. (1987). Associative searching in multiple storage units, *ACM Transactions on Database Systems*, 12, 1, 38-64.

Yergen, W. (1987). 1990 test geocoding experience, *Building on the past-shaping the future, proceedings of URISA 25th Annual Conference*, Vol. II.

Dept. of the Environment (1987). *Handling Geographic Information: Report of the Committee of Enquiry chaired by Lord Chorley*, Her Majesty's Stationary Office, London.