Geocoding of Destination Addresses from Travel Surveys

Noel Villamor
Computer Programmer
University of Melbourne

A.J. Richardson
Director
University of Melbourne

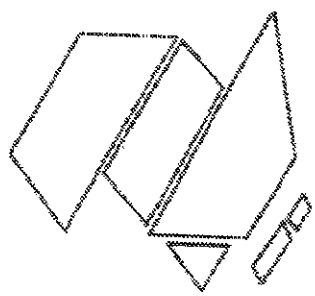
Abstract:

In 1992, the Transport Research Centre designed and conducted the South-East Queensland Household Travel Survey for the Queensland Department of Transport. This survey took the form of a mail-back questionnaire survey which was sent to approximately 20,000 households in Brisbane, the Gold Coast and the Sunshine Coast. In this survey, destination locations were recorded by respondents by means of four variables: street address number, street name, suburb, and nearest landmark (or cross-street). Taking into account the location information supplied by respondents, and the capabilities of the MapInfo® GIS program and associated databases, the paper outlines the development of a hierarchical process of geocoding based on full address matching, cross-street matching, landmark matching, sampling along a street and sampling within a suburb. The paper then outlines some of the problems encountered, and gives an indication of the accuracy obtained from the various geocoding methods.

Contact Author:

Noel Villamor
Transport Research Centre
University of Melbourne
PARKVILLE VIC 3052

Telephone: (03) 344 4074
Fax: (03) 344 7036



1. INTRODUCTION

In 1992, the Transport Research Centre (TRC) designed and conducted the South-East Queensland Household Travel Survey (SEQHTS) for the Queensland Department of Transport (QDOT). This survey took the form of a mail-back questionnaire survey which was sent to approximately 20,000 households in Brisbane, the Gold Coast and the Sunshine Coast (see Figure 1 for study areas). Valid responses were received from approximately 13,000 households. This yielded information on approximately 40,000 people and 150,000 trips. Further details on the survey design and procedures are contained in the project report (TRC, 1993) while details of the response characteristics are described in Richardson and Ampt (1993).

This paper describes the methods used for geocoding of locations in the SEQHTS project, outlines some of the problems encountered, and gives an indication of the accuracy obtained from various geocoding methods. The interested reader may wish to compare the experiences in the SEQHTS project with those reported for the 1991 Sydney Travel Survey (Yeomans, 1992).

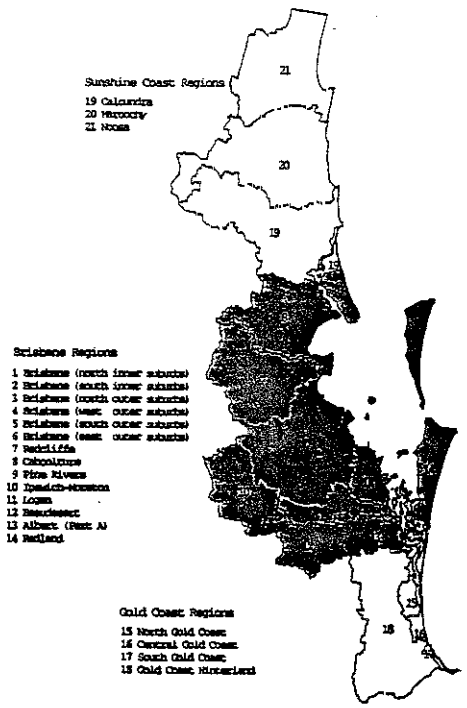


Figure 1 The SEQHTS Study Areas

2. GEOCODING OF ADDRESSES

A major feature of the SEQHTS survey design was that geocodes were to be assigned to each sampled household address and to each of the trip destination locations in the SEQHTS travel data. A geocode is a pair of longitude and latitude values which is based on an earth coordinate system. The process of assigning longitude and latitude values (or X and Y coordinates) is called geocoding.

MapInfo®, a Geographic Information System (GIS) software package, was used as the basic platform for the geocoding task with ERSIS Australia Pty. Ltd., the Brisbane distributor of MapInfo databases, supplying the electronic reference maps of the study area.

In adopting MapInfo, it was expected to be found wanting because it could only geocode full street addresses (i.e. when street number, street name and suburb name or postcode are given). This is a problem because the SEQHTS survey instrument allowed for respondents to provide the nearest cross-street (two intersecting streets) or a landmark in lieu of a full street address. There is also the problem of geocoding partial and/or inaccurate address information provided by the respondent, an occurrence which cannot be totally avoided in a mail-out mail-back questionnaire survey.

In addition, however, the reference maps provided by ERSIS proved to be a major problem when they were found to have serious deficiencies. Foremost of these problems was that there were no suburb boundary files provided for the Sunshine Coast and the suburb boundary files for most of the Gold Coast areas were provided late. Figure 2 shows the extent of the areas for which suburb boundary files were eventually provided. Note that all of the Sunshine Coast is missing, while there are missing suburbs in the Gold Coast and even in Brisbane. As a substitute, however, postcode boundaries were provided and these were used in the geocoding of household addresses in areas not covered by suburb boundary files.

Another deficiency was that a number of streets were missing on the reference maps, especially in newly developed areas. Finally, the database of landmarks was provided late, and was then found to be severely deficient. As a result, the TRC was forced to do a substantial amount of extra work in the preparation of geocoding reference files, and this resulted in added costs to Queensland Transport and severe delays in the completion of the geocoding task. Nonetheless, the use of geocoding has resulted in far better definition of locations than has occurred in previous surveys, and will prove invaluable in subsequent analysis.

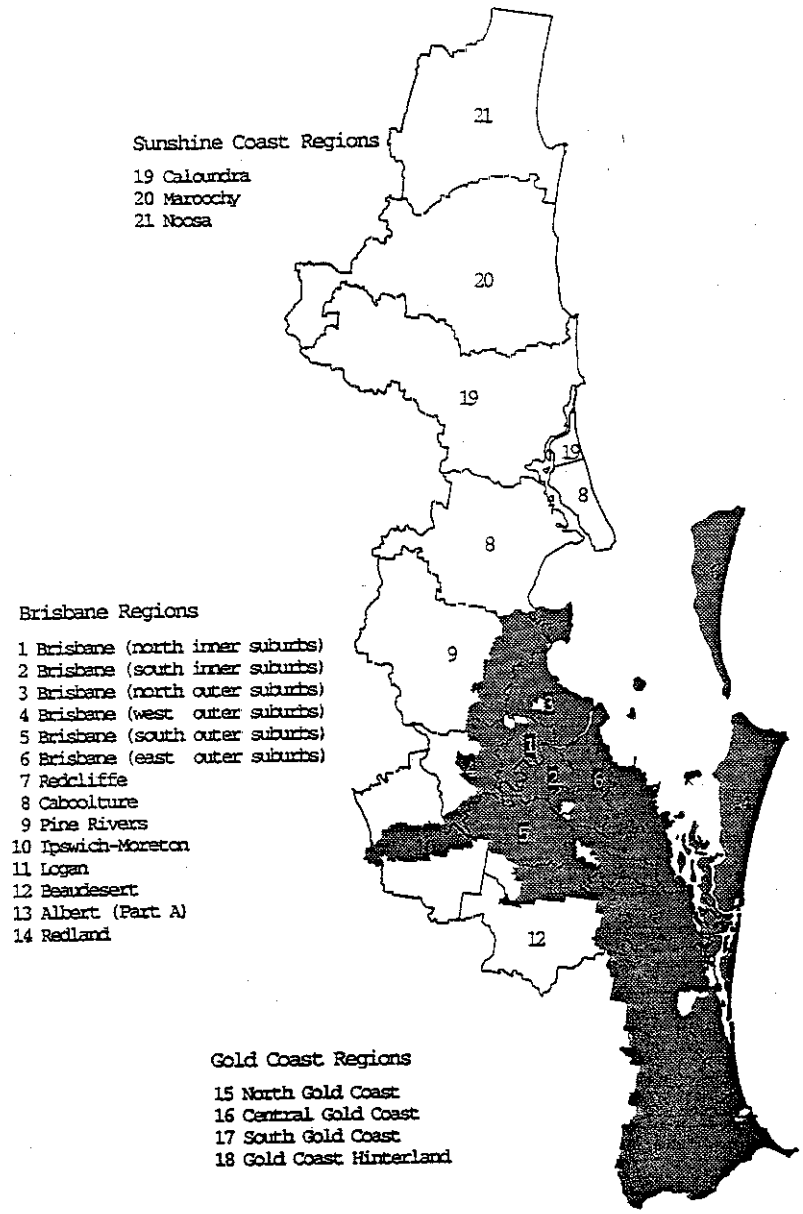


Figure 2 **Areas Covered by MapInfo Suburb Boundary Files**

3. OVERVIEW OF THE GEOCODING PROCEDURE

The conversion of geographic information about home addresses and trip destinations into machine-readable format (geocodes) has been one of the most time-consuming parts of data coding and data entry for the SEQHTS survey. However, such geocodes are extremely useful for the plotting of trip information, for calculation of distances between destinations, and for aggregation of nearby destinations into zones for use in origin-destination matrices. In past travel surveys, destination locations have often been coded directly to rather aggregate traffic zones, with the result that considerable information has been lost about the precise location of destinations. In more recent times, however, the emergence of widely-available Geographic Information Systems has meant that destination locations can now be converted to an x-y coordinate system, with potential accuracies of ±10 metres.

The geocoding procedure used in the SEQHTS survey consisted of a series of geocoding methods applied in a hierarchy to obtain a likely geocode for a set of address information. The accuracy of the geocode is dependent on the geocoding method used. Therefore, the more reliable methods were attempted first. Figure 3 shows the various geocoding methods and their level in the hierarchy (methods nearer to the top of the page generally give the more accurate geocodes).

The degree of accuracy of the geocoding depends on two factors; the accuracy with which the respondent can supply the locational information, and the accuracy with which MapInfo can use that information to generate a set of coordinates. For example, a respondent might know that they went shopping at the Coles supermarket in Chermside. From their point of view, this is the most accurate description of their destination. However, whether MapInfo can geocode this location correctly will depend on what information it has about the location of Coles supermarkets. If all Coles supermarkets are in the landmarks datafile, then this should provide a very accurate geocode. However, if they are not in the landmarks file, then the very accurate locational information provided by the respondent will be of little use, unless an alternative method of locating Coles supermarkets can be found.

It would be possible, for example, to look up the Yellow Pages (or the Telecom Business Finder CD-ROM database) and find that the Coles supermarket in Chermside is on the corner of Gympie and Webster Roads. This information, in that form, is still not very useful since MapInfo needs a street name and number to find a geocode. However, as will be described later, the TRC has written a special program module which finds geocodes based on the specification of cross-streets. Therefore, the accurate locational information supplied by the respondent can eventually be converted into an accurate geocode. On the other hand, the information that MapInfo is most accurate in working with (i.e. full street name, number and suburb) is often not easily supplied by the respondent. For example, very few people who visit the Coles supermarket in Chermside would know the street number of that supermarket, even if they knew what street it was on. If they provided only the street name, then we would be forced to select

a random position along the street within the suburb - providing a less accurate geocode than that provided by use of the shop name.

It is obvious that not all of the methods in Figure 3 can be successfully applied to each address because of differing input requirements. On the other hand, some respondents provided more information than was required, allowing for two or more equally reliable geocoding methods to be applied. An example is when a cross-street or landmark is given together with a full street address. In such a case the geocode obtained using the full street address is preferred.

In the actual computer implementation of the geocoding methods, four program modules were developed for the SEQHTS project. These are:

- geocoding using MapInfo;
- geocoding using a cross-street database;
- geocoding with the assistance of a street directory; and
- geocoding by sampling.

In addition, an interactive spelling checker program was developed to automate the correction of spelling errors/mismatches of street names and suburb names. Spellings were checked against a dictionary created from the electronic reference maps provided by ERSIS.

As shown in Figure 3, the geocodes (or X and Y coordinates) are stored in a file called ADDRESS_XY.DBF. The ".DBF" extension indicates that the basic data were held in a database file, using the FoxBASE+/Mac database program on an Apple Macintosh (this program is similar in many respects to the dBASE program on IBM-compatible platforms). The address information extracted from the various SEQHTS data files, as represented by the file TRAVEL_DATA.DBF, are also kept in this file. Geocoding was performed on the household addresses in the household data file, on the start-of-day locations in the person file, and on the destination locations in the stop file. A separate file of address information is necessary as spelling changes will be made to it to maximise matching success during geocoding.

The next few sections of this paper will discuss how addresses are prepared to make them suitable for geocoding and then details of the four geocoding program modules mentioned above will be provided.

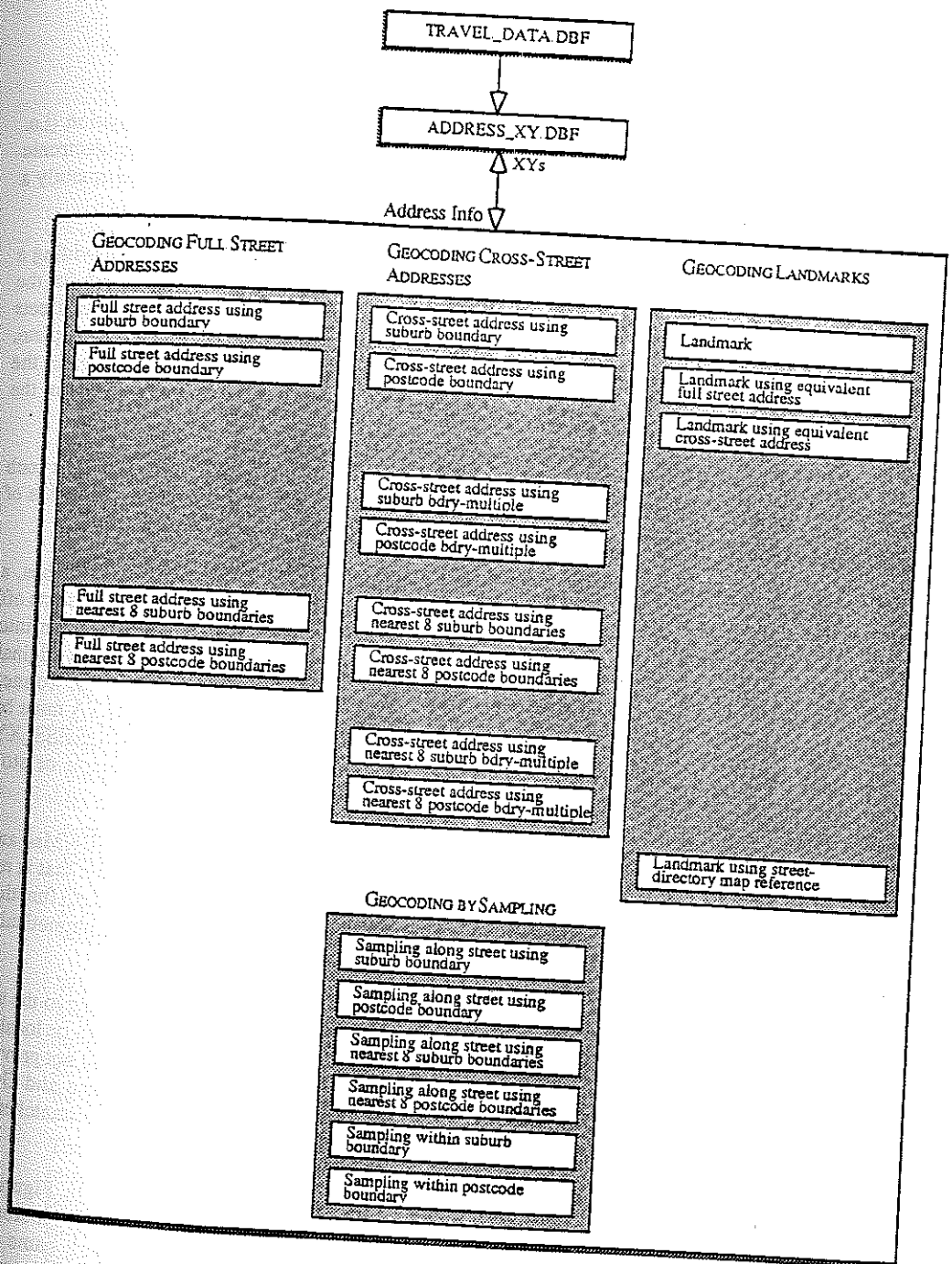


Figure 3 **A Hierarchy of Geocoding Methods**

Preparation of the address data

A crucial factor in geocoding is the success of matching the address information (i.e. street name and suburb name) provided by the respondents to that used in the electronic reference maps. Slight differences in spellings result in a mismatch and consequently a geocoding failure.

Steps were made to minimise spelling mismatches in the SEQHTS data by providing a pop-up dictionary of street names and suburb names in the data entry program for the travel data. The pop-up dictionary even went as far as displaying only those streets which belong to a specified suburb. However, as the dictionary was not really complete, some addresses were still required to be entered manually. Also, a few entries in the dictionary were discovered to be misspelt, giving rise to subsequent problems in the matching process.

The more common causes of spelling mismatches are variations in abbreviations such as Ter & Tce for Terrace, and Mnt & Mt for Mount, and reversals of combinations of names such as Wynnum West and West Wynnum.

Considering that there was so much address information to check for mismatches, a rudimentary interactive program was developed for the purpose. The program starts off by extracting the address records from TRAVEL_DATA.DBF and saving them into ADDRESS_XY.DBF. This latter database saves the spelling changes, with the original address information provided by the respondents left unmodified in the former database. Of course, a way of relating ADDRESS_XY and TRAVEL_DATA must be maintained to be able to attach the X and Y coordinates obtained using ADDRESS_XY onto TRAVEL_DATA. This was done via the unique household, person or stop identification numbers.

The interactive spelling checker program was implemented using FoxBASE+/Mac and has the basic features of a word processing spelling checker. It finds an item that is not in the dictionary and displays candidate dictionary items using the "soundex" function of FoxBASE+/Mac. Soundex is used to determine if two words are phonetically similar, that is, if they sound alike.

It was expected that only a few addresses would turn up as mismatches owing to the use of the dictionary pop-up during data entry. However in the case of suburb or locality names there were quite a number of mismatches. This is because ERSIS did not provide suburb boundary maps for Sunshine Coast, nor for some of the Brisbane and Gold Coast areas. Postcode boundary maps were provided, however, so that mismatches in suburb names were resolved by entering postcode numbers.

To speed-up the process in most of the geocoding methods shown in Figure 3, identification numbers are used instead of the actual names of streets and suburbs. A table of unique identification numbers for each street name and suburb name was created along with postcodes for each suburb. The identification numbers and postcodes are attached to ADDRESS_XY after the spelling changes have been made.

Geocoding full street addresses

The initial task in this procedure is to extract a unique listing of full street address records from ADDRESS_XY. A full street address is one whose street number, street name, and suburb name are given. Geocoding of full street addresses is done using MapInfo. MapInfo basically needs two inputs for geocoding: a street address (which consists of a street number and a street name); and a bounded area (known as a boundary file), such as a suburb or a postcode, to refine the search. MapInfo can geocode in both automatic and interactive modes. What was done usually was to run MapInfo in automatic mode first and then interactively to minimise processing time. In automatic mode, mismatches are skipped. In interactive mode, the user is given the opportunity to manually match each address that MapInfo was unable to match automatically.

It is quite common that respondents give incorrect suburb information and so the address cannot be geocoded. This, however, is often circumvented by assuming that respondents are likely to give a suburb not far from the correct suburb. Respondents often upgrade their suburb to a nearby, more socially distinguished, suburb. By using this assumption, success in geocoding can be improved by re-attempting to geocode using an increasingly larger boundary file. The methods shown in Figure 3 belonging to the full street address category use a larger boundary file as one goes down the hierarchy.

Postcode boundaries are generally larger than suburb boundaries and so they are used in the geocoding process after the suburb boundary. Larger boundaries are further defined using the nearest eight suburb boundaries and the nearest eight postcode boundaries. The number "eight" is chosen with the idea that if a suburb boundary is roughly square, then there will be four adjacent suburbs on each side of the square and another four on its corners. The nearest eight suburb and postcode boundaries were determined by comparing distances between boundary centroids. This was done only once, with the result saved in a database file for use by the appropriate geocoding methods.

It is expected that the probability of *correctly* geocoding an address diminishes as the boundary used becomes larger.

When geocoding a small file of full street addresses, all methods may be attempted in MapInfo before attaching the geocodes onto the address file. However, when the full street address file is large, it saved time if geocodes were attached to ADDRESS_XY after each method was applied and then the full street address file was compressed, by removing geocoded records, before attempting the next geocoding method.

The more recent version of the MapInfo software (ver. 2.0) allows for geocoding mismatches using the closest street number and using a match found in a different boundary as geocoding options. However, during the time that the SEQHTS survey addresses were being geocoded, this version was not yet available and so was not used. Even then the merit of these options still needs to be investigated. The latter of these two options, for example, might give a match on a boundary file that is very far from the area given by the respondent.

Geocoding cross-street addresses

As in the geocoding of full street addresses, a list of unique cross-street addresses was extracted from ADDRESS_XY to avoid unnecessary repetitions in geocoding. A cross-street address consists of two street names and a boundary (e.g. suburb or postcode).

As mentioned earlier, MapInfo does not have the capability to geocode cross-streets, at least as a standard function. A program was therefore written to fill this gap using a fairly straightforward procedure. A database of cross-streets with their coordinates was set-up from the reference maps provided by ERSIS with each record having the following fields:

street_one	- id number of the first street
street_two	- id number of the second street
x_coord	- longitude of intersecting point
y_coord	- latitude of intersecting point
subb_bdry	- id number of the suburb boundary
subb_mult	- number of multiples within the suburb boundary
pcod_bdry	- id number of the postcode boundary
pcod_mult	- number of multiples within the postcode boundary

Geocoding a cross-street address was just a simple matter of searching this cross-street database.

The last four fields of the cross-street database listed above are necessary because multiple occurrences of a cross-street in various locations are possible. To be able to identify which cross-street is pertinent, the cross-street database has to have a boundary field that qualifies each record. Searching a cross-street in turn must also have boundary information as part of the input. But this only partially solves the problem of multiples, as multiples may also exist within a boundary. A good example is a "court" type street where it intersects another street twice, with both intersections likely to be in the same suburb or postcode boundary. Knowing the number of multiples would allow for a randomised approach to selecting a pair of X and Y coordinates among the multiples It should be clear that multiple occurrences of a cross-street which are in different boundaries should not really be considered as multiples.

The geocoding of cross-streets, as in geocoding of full addresses, is also done in successive stages with the next stage using a larger boundary than the previous. Once again, the probability of a correct geocode decreases as a larger boundary is used. For cross-streets, this is aggravated by the random process of selecting a cross-street from its set of multiples, if any.

Geocoding landmarks

In completing the questionnaire, the respondent may specify a landmark as a destination address. Examples of landmarks includes the name of a restaurant, a school, a bank, a government office, a shopping centre, a park, a beach, etc. To be effective as a valid address, a landmark has to be qualified to identify it uniquely from all others with

a similar name. A bank, for example, needs to have the branch (usually a suburb) appended to its name.

The geocoding of landmarks is done by searching a comprehensive database of landmark names with geocodes. The geocodes of each landmark are obtained by various means. If a landmark is one that exists in the landmarks file provided by ERSIS, then the geocode is obtained directly from it using MapInfo. Otherwise, an equivalent full street address or cross-street address is determined manually from printed sources like street directories and phone books and then the geocoding methods for full street addresses and cross-street addresses are used to generate the geocodes.

Finding an equivalent full street address or cross-street address of a landmark poses a problem in cases where one is not available and/or the area covered by the landmark is large (e.g. beaches and parks). For such large areas, area centroids may be used more appropriately as geocodes. Centroids of areas can be marked and geocoded in a MapInfo map, but this process proved to be laborious. An alternative geocoding method was, therefore, developed.

The alternative method involved the development of a computer program that generates a geocode given a map reference from a street directory. An example of a map reference is "A 4 15" where "A" and "4" are row and column references respectively while "15" is a map number in the street directory. A map reference may also be specified as a fraction for a more precise specification as in "B.2 6.3 48A", where "B.2" refers to a point which is 20% of the way between row B and C, "6.3" refers to a point which is 30% of the way between columns 6 and 7, and "48A" refers to map 48A.

The manual task of assigning map references to landmarks was made less taxing by having somebody who was knowledgeable of the study area do the work. In addition, some data entry personnel entered map references as part of the destination address on a number of occasions.

The street directory map reference method was used extensively in the SEQHTS survey to geocode full street addresses and cross-street addresses that failed to obtain a geocode in their respective methods, primarily because the street networks in that area were missing from the MapInfo electronic reference maps. This method works well where the address can be positively located in the street directory maps. However, even for a full street address, the task of identifying the exact location using the street number was sometimes difficult, especially when the street directory maps did not show street numbers.

The accuracy of the geocodes obtained using this method, however, depends greatly on the accuracy of the Refidex maps and the accuracy with which the maps had been digitised into the computer files. Accuracy may be verified by mapping, in MapInfo, a sample of geocodes obtained using this method for each street directory map. This is important to determine the real position of each geocoding method in the hierarchy of geocoding methods (Figure 3).

Geocoding by sampling

Addresses provided by respondents were not always complete. Some intentionally omitted street numbers or just indicated their suburb or locality - probably for privacy reasons. The approach that was used to geocode these cases in the SEQHTS survey was to sample a point along the length of the street, if a street name was given, or to sample a point within a suburb, if a suburb was all that was available.

A long or winding street in a MapInfo map is divided into short segments, usually at street intersections and when it changes direction. Sampling a point along a street therefore consisted of gathering all the segments belonging to the given street within the boundary file, then randomly selecting which segment to use (segments may be assigned relative weights based on their lengths), and then sampling a point along the selected street segment. Sampling a point within an area (suburb or postcode) also followed this procedure, with the added step of firstly randomly selecting a street among the streets within the area.

In addition, the selection of the side of the street was also randomised, and the sampled point was then offset transversely from the street by about 10 metres. This was felt to be necessary as the lines defining the streets on a MapInfo map represent the centre lines of the streets and thus an adjustment had to be made to account for the street width. This adjustment was required because CCD boundaries also follow the centre lines of streets, and this method minimised the incidence of locations falling on the boundaries between adjacent CCDs. The offset of 10 metres is consistent with the way in which MapInfo geocodes full street addresses. This practice, however, resulted in some geocodes "spilling out" of boundary files or onto water areas when the street segment was near a river bank or beach. These occurrences were corrected manually, after visually examining a plot of the geocoded points.

As in geocoding of full street addresses and cross-street addresses, progressively larger boundaries were used when the given street could not be found within the given suburb boundary.

The methods shown in Figure 3 belonging to this category of geocoding were all implemented outside of MapInfo using specially written program modules, but using the reference maps provided by ERSIS.

Locations in other parts of Queensland, in other States, and overseas were not geocoded, but were assigned a pseudo-SLA code to assist in identifying their location.

The methods described above gave rise to a range of geocoding methods, which were recorded with the geocoded location in the respective data files using the following codes:

Code	Geocoding Method
10	"full address, exact match on suburb"
11	"full address, exact match on postcode"
12	"full address, exact match on nearest 8 suburbs"
13	"full address, exact match on nearest 8 postcodes"
14	"interactive matching"
20	"cross-streets, exact match on suburb"
21	"cross-streets, multiple exact matches on suburb"
22	"cross-streets, exact match on postcode"
23	"cross-streets, multiple exact matches on postcode"
24	"cross-streets, exact match on nearest 8 suburbs"
25	"cross-streets, multiple exact matches on nearest 8 suburbs"
26	"cross-streets, exact match on nearest 8 postcodes"
27	"cross-streets, multiple exact matches on nearest 8 postcodes"
30	"landmark, exact match using MapInfo landmarks"
31	"landmark, with equivalent full address"
32	"landmark, with equivalent cross-streets"
33	"landmark, exact match using UBD Refidex landmarks"
40	"sampling along a street, within a suburb "
41	"sampling along a street, within a postcode "
42	"sampling along a street, within nearest 8 suburbs"
43	"sampling along a street, within nearest 8 postcodes "
50	"sampling of street, within suburb"
51	"sampling of street, within postcode"
60	"not geocoded, but pseudo-SLA coded"

4. GEOCODING RESULTS

This section of the paper will now consider the success of the geocoding process as a means of accurately locating points in space. The analysis of the geocoding results will be restricted to the Brisbane study area, because of the known deficiencies in the boundary file information outside the Brisbane area. As noted in section 3, there is a range of geocoding processes which can be used, depending on the quality and completeness of the locational information provided by the respondent. The quality of this information would be expected to vary depending on the type of location (i.e. home address, start-of-day location, and trip destinations). For example, one would expect that the information supplied about the addresses of the sampled households (from the SEQEB (Electricity Connections) datafiles) would be of the highest quality, whereas the addresses of the trip destinations supplied by the respondents would be of lower quality One could therefore expect that, within the range of available geocoding methods, different methods would be used for these types of location. The results of the geocoding process for the three major types of location are shown in Tables 1, 3 and 4.

Table 1 Geocoding Methods for Household Addresses

Geocoding Method	Percent
10 "full address, exact match on suburb"	72.2%
11 "full address, exact match on postcode"	10.5%
12 "full address, exact match on nearest 8 suburbs"	1.2%
13 "full address, exact match on nearest 8 postcodes"	1.0%
14 "interactive matching"	0%
20 "cross-streets, exact match on suburb"	0%
21 "cross-streets, multiple exact matches on suburb"	0%
22 "cross-streets, exact match on postcode"	0%
23 "cross-streets, multiple exact matches on postcode"	0%
24 "cross-streets, exact match on nearest 8 suburbs"	0%
25 "cross-streets, multiple exact matches on nearest 8 suburbs"	0%
26 "cross-streets, exact match on nearest 8 postcodes"	0%
27 "cross-streets, multiple exact matches on nearest 8 postcodes"	0%
30 "landmark, exact match using MapInfo landmarks"	0%
31 "landmark, with equivalent full address"	0%
32 "landmark, with equivalent cross-streets"	0%
33 "landmark, exact match using UBD Refidex landmarks"	6.0%
40 "sampling along a street, within a suburb"	6.4%
41 "sampling along a street, within a postcode"	1.7%
42 "sampling along a street, within nearest 8 suburbs"	0.2%
43 "sampling along a street, within nearest 8 postcodes"	0.8%
50 "sampling of street, within suburb"	0%
51 "sampling of street, within postcode"	0%
60 "not geocoded, but pseudo-SLA coded"	0%

As might be expected, the majority of the household addresses obtained from SEQEB were able to be geocoded directly by MapInfo using the house number, street name, and suburb. Nonetheless, a disturbingly high 27% were not able to be geocoded in this way. Ten percent could not be found in the designated suburb, but were found in the same

postcode. Two percent were found in neighbouring suburbs or postcodes. Six percent could not be found in the ERSIS electronic maps and had to be located via the UBD Refidex street directory. Another 6.5% did not have a precise house number (usually a lot number) and had to be randomly assigned to a position on the street within the suburb. A further 2.7% had to be randomly assigned to streets within the postcode or in neighbouring suburbs or postcodes.

The variation in geocoding method by the location of the home address is shown in Table 2, where the percent of geocoding by each method used is shown for each home region (see Figure 1 for the location of regions, and Table 1 for the geocoding method numbers).

Table 2 Geocoding Methods by Home Location Region

	Geocoding Method Used								
Home Region	10	11	12	13	33	40	41	42	43
1 North Inner Suburbs	81%	8%	2%	--	1%	7%	--	--	--
2 South Inner Suburbs	86%	7%	2%	--	1%	4%	--	--	--
3 North Outer Suburbs	81%	5%	1%	--	3%	9%	1%	--	--
4 West Outer Suburbs	79%	11%	--	--	2%	6%	1%	--	--
5 South Outer Suburbs	81%	2%	1%	1%	4%	10%	1%	--	--
6 East Outer Suburbs	68%	16%	4%	--	2%	7%	2%	2%	--
7 Redcliffe	88%	1%	1%	2%	3%	5%	--	--	--
8 Caboolture	--	47%	--	5%	22%	--	17%	--	9%
9 Pine Rivers	71%	2%	--	--	15%	8%	3%	--	1%
10 Ipswich-Moreton	60%	18%	--	--	16%	3%	2%	--	--
11 Logan	63%	22%	1%	1%	7%	4%	1%	1%	--
12 Beaudesert	3%	13%	--	32%	37%	--	3%	--	13%
13 Albert	88%	--	1%	1%	4%	6%	1%	--	--
14 Redland	77%	--	--	--	5%	17%	--	--	2%
TOTAL BRISBANE	73%	10%	1%	1%	6%	7%	2%	--	1%

It can be seen from Table 10 that the major problems with geocoding of home addresses lay in the outer LGAs, particularly in Caboolture and Beaudesert. Reference to Figure 2 shows that these were the areas for which ERSIS did not supply suburb boundary files. Logan and Ipswich-Moreton also have a relatively low use of geocoding method 10, in line with the areas of non-coverage shown in Figure 2. However, in areas for which suburb boundary information was fully supplied, it appears that approximately 85% of households could be geocoded directly by MapInfo using the full street address, while the majority of the other households were geocoded either by using the postcode of the household or by sampling along the street within the suburb (usually because the household had a lot number for a postal address). Apart from this latter category of geocodes, it would therefore appear that household addresses can be accurately geocoded in areas for which suburb boundaries are provided. As the GIS industry matures and all areas are fully covered by locational files, it would appear that automatic geocoding of household locations will become the norm.

In addition to knowing where people live, the SEQHTS survey asked them where they started their day (i.e. where were they at 4 a.m. on the specified travel day). This location then became the origin of their first trip of the day. This location was geocoded by the methods shown in Table 3.

Table 3 Geocoding Methods for Start-of-Day Locations

Geocoding Method	Percent
10 "full address, exact match on suburb"	68.8%
11 "full address, exact match on postcode"	10.3%
12 "full address, exact match on nearest 8 suburbs"	1.3%
13 "full address, exact match on nearest 8 postcodes"	1.1%
14 "interactive matching"	0%
20 "cross-streets, exact match on suburb"	0.1%
21 "cross-streets, multiple exact matches on suburb"	0%
22 "cross-streets, exact match on postcode"	0.1%
23 "cross-streets, multiple exact matches on postcode"	0%
24 "cross-streets, exact match on nearest 8 suburbs"	0%
25 "cross-streets, multiple exact matches on nearest 8 suburbs"	0%
26 "cross-streets, exact match on nearest 8 postcodes"	0%
27 "cross-streets, multiple exact matches on nearest 8 postcodes"	0%
30 "landmark, exact match using MapInfo landmarks"	0%
31 "landmark, with equivalent full address"	0%
32 "landmark, with equivalent cross-streets"	0%
33 "landmark, exact match using UBD Refidex landmarks"	6.9%
40 "sampling along a street, within a suburb"	6.4%
41 "sampling along a street, within a postcode"	1.8%
42 "sampling along a street, within nearest 8 suburbs"	0.2%
43 "sampling along a street, within nearest 8 postcodes "	0.8%
50 "sampling of street, within suburb"	0%
51 "sampling of street, within postcode"	0%
60 "not geocoded, but pseudo-SLA coded"	2.1%

Since 96% of all people started their day at home, it is not surprising that Table 3 should be very similar to Table 1. Note, however, that there is a start of a drift away from geocoding based on the full address in the correct suburb towards geocoding methods further down the hierarchy.

What starts as a drift away from full address geocoding in Table 3 has become a landslide in Table 4, which shows the geocoding methods used for trip destinations. Only one third of the trip destination locations could be geocoded using the full address in the designated suburb. Given that 30% of all destinations are at the respondent's own home, which can be geocoded by full address on 70% of occasions, this means that very few non-home locations (about 10%) can be geocoded automatically by MapInfo using a full address and suburb. The benefits of geocoding destination locations are not obtained without some considerable time and effort (at least this first time around, when all these lessons were being learnt on a large scale for the first time).

Table 4 Geocoding Methods for Trip Destination Locations

Geocoding Method	Percent
10 "full address, exact match on suburb"	32.5%
11 "full address, exact match on postcode"	5.2%
12 "full address, exact match on nearest 8 suburbs"	1.0%
13 "full address, exact match on nearest 8 postcodes"	0.6%
14 "interactive matching"	0%
20 "cross-streets, exact match on suburb"	9.6%
21 "cross-streets, multiple exact matches on suburb"	0.2%
22 "cross-streets, exact match on postcode"	2.4%
23 "cross-streets, multiple exact matches on postcode"	0.1%
24 "cross-streets, exact match on nearest 8 suburbs"	0.6%
25 "cross-streets, multiple exact matches on nearest 8 suburbs"	0%
26 "cross-streets, exact match on nearest 8 postcodes"	0.2%
27 "cross-streets, multiple exact matches on nearest 8 postcodes"	0%
30 "landmark, exact match using MapInfo landmarks"	5.5%
31 "landmark, with equivalent full address"	0%
32 "landmark, with equivalent cross-streets"	0.1%
33 "landmark, exact match using UBD Refidex landmarks"	23.7%
40 "sampling along a street, within a suburb"	10.9%
41 "sampling along a street, within a postcode"	2.4%
42 "sampling along a street, within nearest 8 suburbs"	0.7%
43 "sampling along a street, within nearest 8 postcodes"	0.8%
50 "sampling of street, within suburb"	3.0%
51 "sampling of street, within postcode"	0.7%
60 "not geocoded, but pseudo-SLA coded"	0.1%

As noted earlier, the final geocoding method depends on both the information provided by the respondent and the capability of the geocoding program and databases to utilise that information. For example, for car trips in the Brisbane study area, Table 5 shows the type of address information provided by the respondents and the geocoding method which was finally employed for the geocoding of those destination addresses. For example, of the total of 33035 car trips, 18024 (54.6%) of the destinations were described by respondents in terms of a full street address. Of these, 80.2% (or 43.8% of the total trips) were able to be geocoded using the full street address, while most of the remainder were geocoded using either a landmark (which may have been provided along with the full address) or sampling along the street (when the street number turned out to be not useful). Thus, we were not always able to use the full information given by the respondent. This cascading effect down the columns is also evident for the other types of destination information provided by respondents. For example, 26% of the destinations described by cross-streets had to eventually be geocoded by sampling along the first-named street, because the specified intersection of streets did not in fact occur. On the other hand, in a small proportion of cases we were able to geocode with higher level information than that provided by respondents. For example, 3% of destinations described by landmarks were geocoded by full addresses or cross-streets (obtained from Yellow Page information).

Table 5 **Destination Information and Geocoding Methods**

	Destination Information					
Geocoding Method	Full Address	Cross-Street	Landmark	Street & Suburb	Suburb Only	TOTAL
Full Address	43.8% *80.2%*	0.1% *0.5%*	0.1% *0.5%*	0 *0*	0 *0*	44% *14549*
Cross-Street	0.1% *0.2%*	11.8% *62.8%*	0.5% *2.5%*	0.2% *3.6%*	0 *0*	12.6% *4149*
Landmark	5.7% *10.4%*	0.6% *3.2%*	17.1% *85.1%*	0.2% *3.6%*	0.1% *6.7%*	23.7% *7827*
Sampling on Street	4.4% *8.1%*	4.9% *26.1%*	1.8% *9.6%*	4.5% *80.4%*	0 *0*	15.7% *5186*
Sampling in Suburb	0.5% *1.1%*	1.4% *7.4%*	0.6% *3.0%*	0.7% *12.5%*	0.9% *93.3%*	4.0% *1323*
TOTAL	54.6% *18024*	18.8% *6201*	20.1% *6648*	5.6% *1843*	0.9% *311*	100% *33035*

The type of geocoding method employed was also shown to depend on the length of the trip involved. For example, Table 6 shows the type of geocoding method employed for walk trips of various distances.

Table 6 **Trip Length and Geocoding Method (walk trips)**

	Straight-Line Trip Length		
Destination Geocoding Method	< 100m	100 -> 1000m	>1000m
Full Address	21%	7%	5%
Cross-Street	15%	19%	8%
Landmark	55%	46%	43%
Sampling on Street	8%	25%	33%
Sampling in Suburb	0%	3%	11%

It can be seen that for very short walk trips (less than 100 metres), we are more likely to use full address and cross-street geocoding, because respondents know locations close to their origin with more precision. As the trip length increases, respondents are less likely to know the detailed address information and hence the geocoding tends to use the more approximate methods involving sampling along streets. Care should be taken, however, in interpreting the right-hand column of Table 6. It is quite likely that the trip length appears to be long *because* we have used a geocoding method based on sampling, thereby introducing an error into the geocoded location of the destination.

The type of destination information supplied by the respondent also varies with the type of destination being visited, as shown in Table 7. Thus while a majority of workplaces, homes and holiday homes are described by their full address, public transport terminals, schools and universities are described only in terms of the landmark itself (eg. Roma Street Station, Wynnum High School). This has implications for the accuracy of geocoding of different types of land use activity

Table 7 Destination Information by Type of Destination

	Destination Information					
Type of Destination	Full Address	Cross-Street	Landmark	Street & Suburb	Suburb Only	TOTAL
Bus Stop	9%	44%	38%	8%	1%	5.9%
Train Station	0%	1%	98%	1%	0%	5.3%
Workplace	53%	21%	17%	7%	2%	10.7%
Other Workplace	38%	31%	16%	10%	4%	5.2%
Pre-School	26%	25%	38%	10%	1%	2.0%
School	8%	13%	76%	3%	0%	7.9%
University	7%	11%	78%	3%	1%	1.2%
Shop	12%	39%	39%	9%	1%	12.0%
Home	100%	0%	0%	0%	0%	29.6%
Other Home	67%	18%	3%	10%	2%	8.2%
Holiday	49%	18%	19%	8%	6%	0.2%
Service Station	18%	54%	13%	15%	1%	1.4%
Any Other	10%	24%	52%	11%	2%	0.9%
Personal Business	24%	36%	33%	7%	1%	3.0%
Social/Recreational	16%	33%	41%	8%	2%	4.8%
Social/Welfare	30%	27%	35%	8%	0%	0.2%
Medical	28%	32%	33%	7%	1%	1.1%
CarPark	10%	52%	28%	10%	1%	0.4%
TOTAL	48%	19%	27%	5%	1%	100%

The heavy reliance on sampling of locations, either from a UBD Refidex (method 33), randomly along a MapInfo street (methods 40 through 43), or from a suburb or postcode (methods 50 and 51) can be seen by the fact that, in Table 4, 42.2% of all destination location geocodes are obtained in this way. The question remains, however, as to the extent of any errors introduced into the geocoded locations by the use of these more approximate methods.

The accuracy of geocoding methods

The question of the accuracy of the various geocoding methods can be inferred from the effects of the geocoding method on the spatial characteristics of the resultant trips. One way of doing this is to calculate the straight-line distance between the geocoded origin and destination of a trip and then, by comparing this distance with the reported travel time for the trip, estimate the average straight-line speed for the trip. Some errors will be introduced due to errors in the estimation of the travel time by respondents, and by the assumption of straight-line distance, but these errors should even themselves out over the various geocoding methods.

Errors in geocoding will show up primarily by means of very fast trips or very slow trips. That is, the geocoding method will mistakenly position one, or both, of the trip ends in the wrong place, leading to what appears to be a very high, or very low, speed trip in the available time. In practice, it is less likely for geocoding errors to result in very slow trips because this would require the misplaced trip-end to be located nearer the other trip end than it is in reality. It is also more difficult to detect slow trips, because while being slow they are still possible. By examining the geocoding methods used to generate high speed trips, one can obtain an idea of the extent of error being introduced by the use of approximate geocoding methods.

The criteria for a very fast trip is dependent on the mode of travel used. Speeds greater than, say, 8kph would be considered a very fast trip for the walk mode, while for car modes the limit would be very much higher. Furthermore, for some modes, the high speed criteria would be expected to be dependent on the distance of the trip. For example, the average speed of the car modes would generally be higher for longer distance trips, than for short distance trips, as the relative effect of parking and unparking is minimised and as the trip becomes more likely to be made on freeways and rural roads. The relationship between straight-line trip distance and trip speed for car trips in the Brisbane study area is shown in Figure 4, confirming the above hypothesis.

As can be imagined, it is difficult to be precise in setting limits for very high speed car trips. In this analysis, the very fast car trips were identified by grouping the car trips based on distance ranges (ten groups with 10% of the trips in each group), and then identifying the upper and lower quartiles of the straight-line speed distribution for each group. The interquartile distance is called the "box-length" in a box-plot diagram (Norusis, 1990, p110). Extreme values are defined as values which are greater than three box-lengths from the quartile points. Using this definition, very fast car trips are defined as being extreme values on the high side. For walk trips, it was somewhat easier to identify unreasonably fast trips, because of the physical limitation on walking speed imposed on most people. A value of 8 kph was used as the maximum reasonable walking speed. This value was not used unilaterally, however, because in examining the speed distribution within each of the distance ranges, it was clear that the rounding off of travel times by respondents to the nearest five minutes was creating bimodal distributions of trip speed. The upper limit of reasonable walking speed was therefore raised to the tail of that part of the distribution containing the 8kph value.

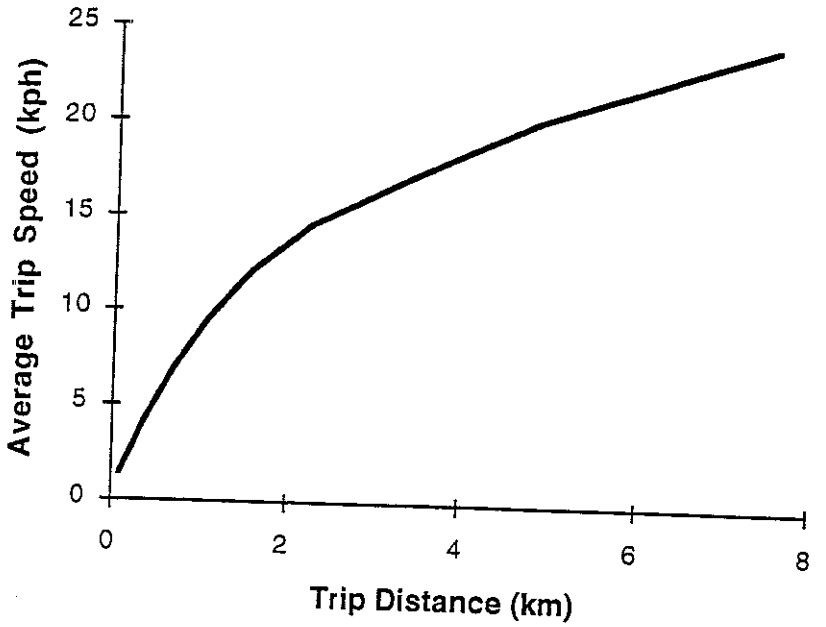


Figure 4 Trip Speed as a function of Trip Distance (straight-line)

Having identified very fast car and walk trips for each distance range, these extreme values for each distance range were then grouped into overall groupings of very fast trips for each mode.

For the modes of walk and car driver, the geocoding methods employed for the origin and destination of the trip were then examined for the very fast trips to assess the relative frequencies of the various geocoding methods giving rise to apparent errors in geocoding. Since geocoding errors at either end of the trip could result in a very fast trip, each trip was characterised by the more approximate method of geocoding used for the origin or destination (i.e. the higher numbered geocoding method from Table 1).

The results of this analysis for walk and car driver trips are shown in Table 8, where the percent of very fast trips for each geocoding method is shown. For example, for car trips where the worst geocoding method used for coding of the origin or destination involved one based on full address matching (methods 10 through 14), 2.6% of trip speeds were classified as being very fast, and hence possibly in error. When the worst geocoding method involved sampling within a suburb, the percent of very fast trips rose to 5.9%. A similar pattern was noted for walk trips, but the percent of fast trips was greater for each geocoding method and deteriorated more rapidly as one moved to the use of the more approximate geocoding methods. This may be due partly to the different methods used to define very fast trips for each mode, but is also related to the

size of the geocoding error (when using street or suburb sampling) compared to the length of the trip (which is much shorter for walk trips).

Table 8 Percentage of Very Fast Trips by Geocoding Method

	Mode of Transport	
Geocoding Method	Car Trips	Walk Trips
Full Address	2.6%	4.0%
Cross-Street	4.5%	3.6%
Landmark	5.1%	8.8%
Sampling on Street	5.8%	12.7%
Sampling in Suburb	5.9%	26.9%

5. CONCLUSION

This paper has described the geocoding methods used in the 1992 South East Queensland Household Travel Survey. Taking into account the location information supplied by respondents, and the capabilities of the MapInfo GIS program and associated databases, the paper outlines a hierarchical process of geocoding based on full address matching, cross-street matching, landmark matching, sampling along a street and sampling within a suburb.

The paper then describes the results of the geocoding process in terms of the types of geocoding method used for different types of locational information and for different geographical regions. It examines the type of locational information supplied for trip destinations by respondents and the subsequent use of geocoding methods. It examines the variation in geocoding method with the length of walk trips, and the differences in destination information supplied for different types of trip destination. The paper then examines the extent to which the more approximate geocoding methods give rise to incorrect positioning of origins and/or destinations as reflected in the incidence of very fast trips. It is shown that as one uses the more approximate geocoding methods, the probability of generating a very fast trip increases, especially for the shorter, slower types of trips.

However, the effects of geocoding errors should not be too over-dramatised. Of the 70,000+ trips recorded in SEQHTS in the Brisbane study area, about 95% of these trips have speeds which would be considered as reasonable for the mode involved. It therefore appears that most trips have been geocoded reasonably accurately.

References

Norusis, M.J. (1990). *SPSS Base System User's Guide*. SPSS Inc.: Chicago.

Richardson, A.J. and Ampt, E.S. (1993). Non-response effects in mail-back travel surveys. *Papers of the 18th Australasian Transport Research Forum*, Gold Coast.

Transport Research Centre (1993). *1992 South East Queensland Household Travel Surveys. Final Report 1 - Brisbane*. Report to the Queensland Department of Transport, June.

Yeomans, G. (1992). Geocoding survey address information. *Papers of the 17th Australasian Transport Research Forum*, Canberra, Part 1, p 115-133.