Geocoding Survey Address Information

Geoffrey Yeomans
Landuse Manager
Transport Study Group NSW

Abstract:

Geographic aggregation of survey data above individual address level is generally necessary in order to render data analysis and modelling tractable. It is also necessary due to increasing concerns about survey respondent privacy

As part of its 1991/92 Survey programs, the NSW Transport Study Group (TSG) has developed PC-based software for automated geocoding of individual addresses.

This paper outlines the elements and performance of the TSG software and outlines a number of 'fuzzy logic' matching procedures which has been built into the software in order to accommodate:

inaccuracy, ambiguity and incomplete address reporting;
overlapping/'fuzzy' suburb boundaries;
multiple suburb names

The integral role of Geographic Information System software and related work by the Australian Bureau of Statistics and the Geographical Names Board is noted.

Contact Author:

Mr Geoffrey S Yeomans
Transport Study Group NSW
Locked Bag #1
HAYMARKET NSW 2000

Telephone: (02) 218 6637 Fax: (02) 218 6625

Overview

As noted in other papers contributed to this conference, the Transport Study Group NSW (TSG) is currently undertaking both a Home Interview Survey (HIS) of individual travel and a Commercial Vehicle Survey (CVS). Field work for these surveys is due for completion in December 1992.

These surveys update TSG's main travel databases and are to be related to both 1991 Census data and to the separate Journey to Work (JTW) data series processed from the Census.

As part of their data processing procedures, both the Australian Bureau of Statistics (ABS) and TSG had a need to geocode respondent address data. All data sets are large and beyond the scope of purely manual geocoding. As no suitable computer software was available at the time TSG's work was required and, as the base reference material for the complete 1991/92 Travel Survey had to be compiled anyway, TSG developed its own address geocoding software in-house.

This paper outlines the role Spatial Information System (SIS) technology played in the construction of suitable reference files and how limitations associated with this work were overcome using textural data from a variety of sources. Procedures developed in order to deal with both incomplete and inaccurately reported transport trip addresses are also summarised, along with measures used to control and report the quality of geocoding achieved.

Availability of this geocoding software opens new opportunities for adding a spatial dimension to address records which were not originally constructed with this in mind. Once geocoded, such databases can be accessed and analysed with an SIS platform, thus helping to realise some of the promise of SIS technology.

Spatial Information System Concepts

Application of "Geographic" or "Spatial Information System" (GIS or SIS) technology is increasing. Mystique still surrounds the concept of an SIS and the vendor promise is typically that all the databases in an organisation can be integrated by using this technology and hence, that massive savings are available.

In essence, an SIS is a powerful form of relational database which accesses digitised spatial information in the form of 'points', 'lines' and 'polygons'. Like other forms of relational database, it is unable to fulfil its promise if the relationships between different data sets are not correctly specified or if relevant key fields cannot be established. The elements which make an SIS different from a purely numeric or textural relational database system is its ability to operate with spatially keyed data and its ability to relate this to maps.

In general terms, address "geocoding" is the process of allocating coordinates or spatially-significant codes to specific addresses. Geocoding can be done at either a 'point' level whereby each individual address is allocated a unique pair of geographic coordinates such as latitude and longitude or at an 'area' or 'polygon' level whereby

specific collections of addresses are allocated a geographic code which is unique to the closed 'polygon' in which they are located. In an SIS, a separate table is held containing each different polygon code number and a collection of digitised points together with software-specific instructions to permit the polygon to be plotted accurately. In a relational database, separate tables contain common 'key' fields which can be used to link different tables. In an SIS, a *geographic* key field (or number of fields) is required. Thus, specific information relating to a polygon can be extracted from a suitably keyed database for indication on a map by shading or by some other form of marking.

SIS databases are ideally coded at a 'point' level since this maximises their flexibility as a base for enquiries at any higher level of geographic aggregation. In practice however, privacy considerations, non-availability of suitable reference materials and budget constraints usually prevent this level of geocoding for some or all of an organisation's databases. For each level of geographic aggregation which is imposed above this point level, the richness of a database is diminished. As a result, the full promise of an SIS being able to link all of an organisation's databases is correspondingly reduced.

Geographic 'layers' are a critical concept for an SIS. The schematic diagram of Fig 1 illustrates this concept of 'layers' or 'overlays'. Once digital information for each relevant layer is compiled and links to suitably structured data tables are established, multiple topological and other transformations may be performed and a resultant image or report produced. A companion constraint to reaching an SIS's full potential is therefore the degree to which different spatial layers are compatible. The digitising of individual SIS layers has typically been:

- undertaken by different organisations
- based on different map scales and,
- undertaken to conform to different standards.

Invariably therefore, none of the layers overlay correctly and considerable SIS effort is required in order to get them to do so. This is a generic problem which bedevils all SIS work at present and which greatly limits the promised contribution of SIS technology and substantially increases its cost. SIS vendor promises therefore need to be assessed critically against these practical limitations to application of the technology.

Rather than considering an SIS to be a single all-purpose tool, its full value is usually realised when it is combined with a suite of other compatible computer-based data collection and analytic tools. For the type of data maintained by the TSG, an SIS is a critical tool for capturing, structuring and reporting spatial information. Some data accessing and analysis functions, while being crudely possible on an SIS, are performed more effectively using other software tools. Thus the heart of TSG's data storage processes is currently comprised of ORACLE and GENAMAP software (for numeric, textural and geographic information). Its data analysis and modelling needs are satisfied by a wide range of spreadsheet, statistical and programming packages. Map reporting can be achieved by means of TSG's transport modelling software, (EMME2 or TRANSCAD) or by one of a number of SIS software packages (GENAMAP, MAPINFO or ARCINFO).

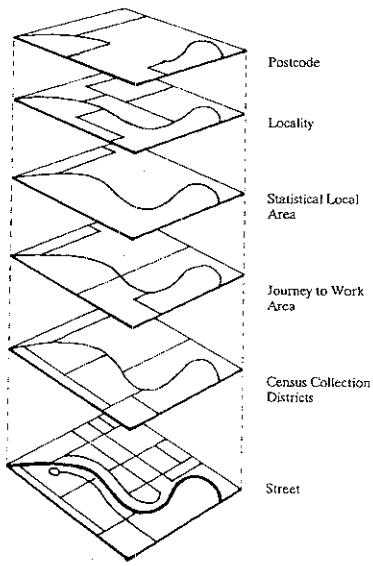


Fig 1.. SIS layers required in order to create address matching reference files

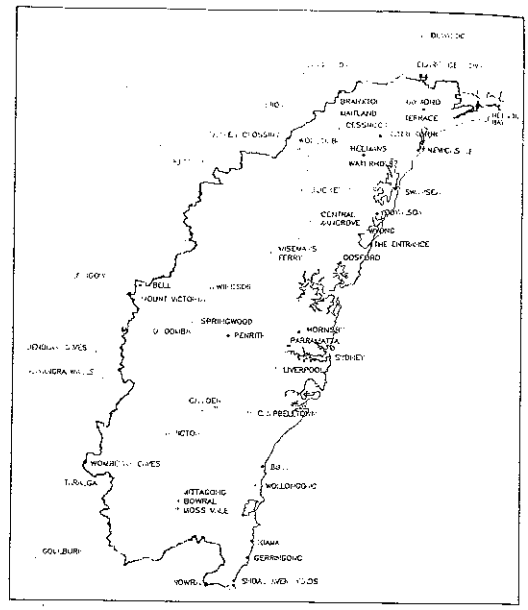


Fig 2.. 1991/92 Study Area

Traffic Zones: A Basis For Transport Modelling and Analysis

Transport is a spatial activity.. In economic and modelling terms, individual utility is a function of how the spatial separation of activities is overcome.. Much of the recent transport modelling debate has focused on the need to analyse the choices which individuals make rather than to analyse the responses of some notional 'average' set of individuals.. Thus, while 'individual choice models' are now often thought to be desirable (for which individual respondent records are necessary), real database management issues exist which tend to indicate a level of geographic aggregation in the data sets..

In order to ensure individual respondent confidentiality and in order to render database and modelling work tractable (especially for large study areas) databases have traditionally been structured in terms of spatial aggregates.. Similarly, the fact that statistical reliability can only be attached to aggregated data from surveys such as the HIS and CVS tends to stipulate databases which store information at greater than individual record level.. Spatially aggregated data sets also become necessary due to the non-availability of geocoding reference material which will support geocoding at the level of individual households.. The 'correct' level of aggregation is difficult to define unambiguously but, since different landuse is a key determinant of travel demand, a geographic system based on landuse is clearly desirable..

Definition of Traffic Zone areas (sometimes known as 'Destination Zones' in recognition of the fact that transport databases often contain data about the origins and destinations of individual trips) is a technique for capturing transport information based on homogeneous land uses and yet maintaining a reasonable degree of data variability. Zones represent a compromise between highly disaggregated areas such as Census Collection Districts (CD's) and more highly aggregated areas such as Statistical Local Areas (SLA's). These zones are defined so as to completely cover the space inside a specified study area and to be able to be mapped without overlap. Their design should conform to a range of criteria and should relate to other geographic boundary systems about which data is collected. In transport modelling terms, there is a need to relate transport survey data to both population census data (which is CD-based) and to economic data series (which are industry-based).

The ABS has established a comprehensive industry classification system composed of Australian Standard Industry Classification codes (ASIC). It also maintains an index of relationships between different levels of geography used for Census purposes known as the Australian Standard Geographical Classification (ASGC). TSG geographic systems have been designed to link to both the ASIC and ASGC systems.

The following criteria are employed by the ABS in defining CD boundaries:

1. In aggregate, CD's must cover the whole of Australia without omission or duplication.
2. The chosen CD boundary should, if possible, be readily identifiable on the ground, be defined in terms of permanent features, follow the centre of the road or river if these features are used. The use of major roads as CD boundaries in rural areas should be avoided where possible, i.e. to minimise splitting of identifiable rural localities.
3. CD's should conform where possible to existing/gazetted suburb boundaries. CD's must not cross SLA boundaries and, as a consequence, any other ASGC spatial unit boundary.
4. CD's should be consistent with both the collector's workload requirements and their role as a useful spatial unit capable of aggregation into broader level ASGC spatial units.
5. The area and population delimited by a CD boundary must be such that one collector can deliver and collect census forms within about ten days.
6. CD's should not be designed in such a way as to make them confidential for publication of data. Accordingly, a CD should contain, where possible, at least 100 persons.

Yeomans

Many of the same criteria, together with a number of different criteria, are employed by TSG when defining traffic zones. The main criteria include the following:

1. Zones should embody homogenous land uses as far as possible.
2. Topological features should be followed as far as possible.
3. Comparability with past zone boundaries should be achieved as far as possible.
4. Relatively constant population and area should be maintained within each zone.
5. Zones should be large enough to generate a substantial numbers of trips, in order that geocoded data is statistically valid. A rough rule is that each zone should generate about 10,000 trips daily.
6. Wherever possible, zones should be:
   - aggregates of Census Collection Districts
   - sub-aggregates of Statistical Local Areas
7. Recognition should be given to the possible location of future transport corridors so that no zone contains more than one freeway interchange or one railway station
8. Defined locations of committed future transportation facilities should be adopted as boundaries where possible.
9. Future planning schemes should be reviewed in order to determine the need for defining special zones for future major traffic generators.
10. Zones should be relatively compact in shape rather than irregular or elongated.

Travel databases are typically employed in a planning and forecasting context and hence current and future landuse is a key criteria for the definition of Traffic Zones. For definition of Census-related geography however, current residential density is a more significant criteria. It is unlikely that all CD boundaries in a study area will always uniquely map onto a Traffic Zone since their definition criteria and the weighting attached to any one criterion differ. Ultimately, construction of any zoning system embodies compromise decisions taken in relation to:

- Demands for 'strategic' versus 'detailed corridor or sub-area analysis'
- Data 'richness' (individual households) versus privacy and budget considerations (area-based aggregation)
- Computing and modelling capacity, speed etc (area-based and other forms of aggregation)
- Statistical significance of aggregated data (usually Statistical Local Area) versus travel data variability (Traffic Zone or smaller areas)

In recognition of the need to be able to connect databases which were established on different geographic bases however, TSG has expended considerable effort in establishing the relationships between past and current Census boundaries and the different zoning systems employed in modelling transport behaviour. Provided that this work is undertaken accurately, it is possible to construct past databases in terms of current geographic boundaries and, via an SIS, to map and analyse all data on a common geographic base. Often, time series analysis is weakened by not adjusting data for the different geography on which data sets have been based at different time periods.

Datasets Requiring Geocoding

Data from the 1991 Population Census is now being progressively released by ABS and work-related travel data geocoded to form the JTW series.

It is estimated that the ABS JTW geocoding task for NSW will involve some 1.6 million address records while TSG's total geocoding for survey and other purposes will involve upward of 4.0 million records. Clearly such tasks are best automated.

1991/92 Study Area

The area for which HIS, CVS and JTW databases are being compiled is outlined on the map in Fig 2. The extent of this area posed significant difficulties in compiling suitable locality and street address-level reference files. The extent of this difficulty can be envisaged by reflecting on the processes required in order to establish a reliable database of all addresses in this Study area.

For 1991/92 address geocoding purposes, some 1200 geocoding areas were defined and each of these areas was assigned a unique four digit numeric code. These geocoding areas are not Traffic Zones but are related to them. These areas are typically larger than 1991 Census CD boundaries and are linked to them. While whole CD's do not aggregate to geocoding areas in all cases, the geocoding areas always uniquely aggregate to Statistical Local Area (SLA). SIS 'overlays' were used to produce digital boundary files and to relate these boundaries to both a digital street centre-line base and to the formal Locality boundaries defined by the Geographical Names Board.

ABS and TSG Geocoding Approaches

Reported travel destination addresses inevitably contain a number of types of inaccuracy (i.e. missing address elements, incorrect address elements and incorrect spellings). Thus, even if perfect reference files could be compiled, judgement is ultimately required about whether a reported address matches a reference file address "well enough" to allocate a geocode in which data-users can feel confident. Diminishing returns are eventually encountered when trying to improve user confidence in the face of less than perfect input data and reference material.

The sheer volume of addresses requiring geocoding together with the labour costs and accuracy issues associated with any purely manual processing, effectively dictate the use of automated geocoding procedures.

For geocoding the 1991 NSW JTW data series, the ABS has adopted geocoding areas defined by the TSG together with street address and locality reference files also prepared by TSG and edited by ABS. ABS has developed a 'computer assisted' geocoding process for this purpose. Using the ABS procedures, coders sequentially key in individual address elements (suburb, street name, street number etc) until sufficient information is available to allow a geocode to be allocated without ambiguity. The address information is not retained electronically and, once the census forms are destroyed, verification of geocoding is impossible. Geocoding quality is established by measurement of the level of manual query resolution required and by performance benchmarks. 'Live' adjustment of the reference files is incorporated into this process thus implying an improvement in data quality as geocoding progresses.

TSG has adopted a different approach whereby trip identification keys and their associated addresses are electronically stored and passed to its geocoding software for batch processing. Actual geocoding runs can be undertaken outside of normal working hours and each input address is returned with a 'most likely' geocode and a number of individual parameter values which indicate both the nature of the 'decisions' taken in order to establish this code and any difficulties encountered by the matching procedures which had to be invoked in order to find a match. Geocoding quality was established by automatically processing addresses whose geocoding had been independently verified. On these verification runs, 99.93% of all input records are geocoded and 98.08% of these are geocoded correctly (ie 98.01% of all input records were correctly geocoded). Imperfections in the reference files rather than in the geocoding algorithms account for most of the small percentage of geocoding errors and failures.

Overview of TSG Geocoding Process

In addition to a capability to geocode on the basis of a 'fuzzy' match between an input address and a reference file address, TSG's software was designed to support matching on the basis of business and/or place names. In this paper however, discussion is confined to address matching.

Provided suitable reference files are available, TSG's geocoding software reads a file of input addresses and attempts various forms of exact and inexact match between each input address and addresses contained in reference files. If input addresses are

structured as a single text field, they are parsed into individual address elements (postcode, suburb name, street name, street type, street modifier [ie 'North', 'South' etc], corner street, street number and lot number) before a match is attempted.

The geocode areas associated with each suburb are critical to the address matching process as they provide a list of 'candidate' geographic areas in which street-level matches are attempted. If a match between the input street information and that contained in the reference files for the candidate areas is either impossible or of low quality, then the candidate area list is extended by adding all areas adjacent to those in the initial candidate list. Matching is subsequently attempted using this extended list. This approach was established empirically after determining that there was no significant improvement in either geocoding frequency or quality when the candidate area list was extended beyond the first 'ring' of adjacencies.

After establishing one or more geocodes which satisfy a range of match quality thresholds, the final candidate geocodes are ranked in order of a "measure of fit" (explained later in this paper) and the highest is selected as the most likely area in which the input address is located. Each input address is subsequently returned with a:

- "best fit" geocode number
- "return code" (explained later in this paper)
- "measure of fit" number

Each software run produces a summary report file which details:

- The list of parameter values adopted for the specific input file
- Count of the number of input addresses
- Frequency counts of:
  - The number of candidate geocodes from which the selected geocode was chosen
  - Cases where one or more candidate geocodes are found and the individual and composite return codes involved
  - Records returning zero, one or more than one return code
- Frequency counts and cumulative % of all records within each "measure of fit" group

Construction of TSG Reference Files

TSG does not hold detailed address information as part of its normal business and thus had to access third party records. Typically this information did not satisfy all TSG's criteria as it had been compiled for non-TSG purposes. This disparity was further complicated by the fact that no single organisation held comprehensive records for the complete 1991/92 study area. All files contained a range of omissions and inaccuracies and thus required considerable editing and re-structuring before being useable.

Introduction of SIS technology is not yet widespread and the organisations approached for address information were either unable to provide SIS files at all or were only able to provide incomplete files. As a general rule, SIS-based information was of

significantly higher quality than that sourced from organisations not yet using SIS. As a result, 'perfect' reference files were unable to be constructed.

The digital street centre-line and associated textural street name and number range information produced by Peripheral Systems Pty Ltd was a primary source of street-level information for the Sydney area. At the time that this work was undertaken, the Peripheral Systems coverage of the Blue Mountains, Wollongong, Central Coast and Newcastle areas was incomplete. Textural information from other organisations' customer databases was therefore a supplementary source of street level information for these areas. Access to this information was discussed in detail with the Privacy Committee and this access was judged to be appropriate since TSG did not access individual household occupant names from any of these databases. In addition, specifically drafted confidentiality and limited usage agreements were entered into by each individual staff member who accessed this data. One set of such agreements had to be tabled in Federal Parliament before access was granted. TSG policy is to treat data confidentiality seriously and a range of hardware and software security measures have been implemented for all TSG data sets. Files of street addresses contained within each geocode area were established by both database manipulation and SIS layer overlay procedures.

SIS-based information was amenable to detailed analysis whereby streets were broken into segments and geographically arranged in sequence in order to assist resolution of such boundary issues as the identification of the numbers on one side of a boundary street segment and those on another. Identification of street number ranges for streets which cross geocode boundaries was also greatly assisted wherever SIS information was available.

In addition to street-level information, definitive locality (= suburb) information was also required. Both the TSG and ABS processes begin with the suburb component of a reported address in order to narrow the search for relevant streets or street segments. Historically, locality names had "fuzzy" notional boundaries, which overlapped and were often identified by multiple names. It has been the Geographical Names Board's (GNB) task to establish legally binding unique names for all suburbs. At the time that TSG's work had to be undertaken however, the GNB had not completed its work for the entire study area. For the work which they had completed, unique legal names were available as well as digital locality boundaries. Peripheral Systems had also developed notional digital boundary files for many areas within the study area. These digital boundaries and names have a different status to those of the GNB as they reflect common usage and Peripheral Systems' judgement rather than definitive legal boundaries. If Peripheral Systems locality boundary information was available for areas not completed by GNB, they were incorporated into TSG's SIS locality layer. The remaining areas (ie those which were not digitised by either Peripheral Systems or GNB) were defined as unique aggregates of TSG's 1991/92 geocoding areas. Names for these areas were established from all textural sources accessed from other organisations and from the ABS National Localities Index. Files defining the geocode areas associated with each locality were subsequently established by SIS layer overlay procedures. Similarly, files of adjacencies and centroid coordinates for each geocode area were established by SIS procedures.

Like suburb boundaries, postcode boundaries have traditionally only been notionally defined and exact digital boundaries are unavailable. As postcodes have a role similar to that of suburbs in TSG's geocoding procedures, an appropriate index of

postcodes and TSG geocode areas was required. This index was established by database manipulation techniques rather than by SIS procedures and utilised topological tables which had previsouly been constructed at CD level.

As noted earlier, the fact that all files for the complete study area contained information from may different sources and information of substantially different quality, inevitably resulted in less than perfect reference files against which less than perfect input addresses were to be matched. A number of different 'fuzzy logic' procedures were therefore required in the address matching algorithms in order to resolve both input and reference file inadequacies wherever possible. Given these inaccuracies, the geocoding coverage and accuracy results achieved by TSG's software are judged to be extremely good.

TSG Address Matching

The aim of the matching process is to find the correct geocode for a given address. In less than ideal cases, multiple geocodes may be found with varying degrees of quality. In such cases, 'candidate' geocodes are sorted in descending order of their "measure of fit" (MOF) and the topmost geocode is selected as the most probable. Thus, the topmost geocode (5678) would be selected from the following list of final candidates.

Geocode	MOF	Return Code(s)
5678	0.742	x
5814	0.687	xN
5816	0.610	ZIN

Calculation of a "Measure of Fit" (MOF) is a technique whereby the "quality" of a given type of match between an input address and a range of reference options can be assessed indicatively. In general terms, a perfect MOF (probability = 1.0) is allocated wherever an address matches a reference address in every particular. The greater the degree of mismatch, the lower the "probability" that the reference address is the same as the input address. In the matching software therefore, all addresses are initially assigned a perfect MOF which is subsequently decremented according to specific rules as the matching invokes increasingly less accurate searches in order to find a suitable geocode. The MOF is also used as a value to assess against a range of threshold values at different stages of the matching in order to determine if a given candidate geocode is "good enough".

There are a number of parameters which can be controlled by the user in order to control the behaviour of any one address matching process. To ensure geocoding of comparable quality across different data sets however, these parameters are best not altered once software performance has been calibrated.

Parameters are identified by a number code and a unique value is associated with each parameter number. Broadly, there are three classes of parameter: those which set 'decision' threshold values, those which specify the type of matching to attempt and those used for calculation of "measure of fit" values

The "Return Codes" which are reported in conjunction with the chosen geocode and its calculated MOF give an indication of the type of difficulties encountered when assigning the selected geocode. Their meanings are summarised in Figure 3.

If input address elements are not recorded in separate fields, they must first be parsed as address matching procedures operate on the separate address elements differently.

Before any matching is attempted, all input addresses are "standardised" (ie converted to capital letters, name modifiers are placed after the main name and expanded if necessary, illegal characters are eliminated and extremely long names are truncated). Thus an input suburb such as "*W Dewhy*" is standardised to "*DEWHY WEST* ". When attempting to match this suburb name, a Return Code "*b*" would be assigned to indicate that it could not be matched exactly with the reference file list since it is not spelt correctly. If the suburb name consists of more than one word and the last word is one of a predefined set of modifiers (ie "North", "South" etc), the modifier is ignored and a match attempted for the basic ("simplified") suburb name. In the "*DEWHY WEST*" case, "*WEST*" is ignored but again no exact match is found and a Return Code of "*c* " is recorded. This second Return Code supersedes the first as many matching procedures are undertaken sequentially and a failure at a later stage implies failure at preceding stages.

After attempting a simplified match, a "soundex" match is attempted. This involves a "soundex key" which represents the sounds of the letters in the name. "*DEWHY WEST*" would also fail a soundex match and a Return Code of "*d*" would replace "*c*".

A "skeleton" match on the full name (ie not simplified) is attempted next which, as the name implies, reduces the name to a prespecified list of minimal letters. The above example would also fail a skeleton match and thus attract a Return Code of "*e* ". Should a match have been found by either soundex or skeleton keys, the resultant candidate match would have been tested for how well it corresponded to the original input name. Relevant parameter threshold values would be invoked at this stage and candidates which did not meet the threshold value would be dropped.

After failing a skeleton match, all reference names of a similar length to the full input name are examined (ie scan matching). The results are also assessed in terms of their relationship to a threshold quality value. If they fail to achieve an adequate match with this procedure, a Return Code of "*f* " is assigned.

A final matching is attempted using soundex on the simplified name. If this fails then a Return Code of "*g*" is assigned. The process stops at this stage as suburbs are a first filter which produces a short list of possible geocode areas. If this final process is successful, a Return Code of "*w* " is assigned and the street matching processes are invoked using the geocode areas associated with the candidate suburb name as a 'seed'.

Input street address details are put through similar processes to those for suburb names. In addition, other variations between input and reference details are required. A street name, street type and street extension may match reference material information completely but the input street number may be different to the number range applicable to the most likely geocode area. Similarly, a street name might be correct while its type is not (ie "*JOHNSTON AVENUE*" rather than "*JOHNSTON STREET* ".) In these and related cases, decision rules are invoked in order to determine whether such a match is better than one which finds, say, "*JONESTOWN AVENUE*".

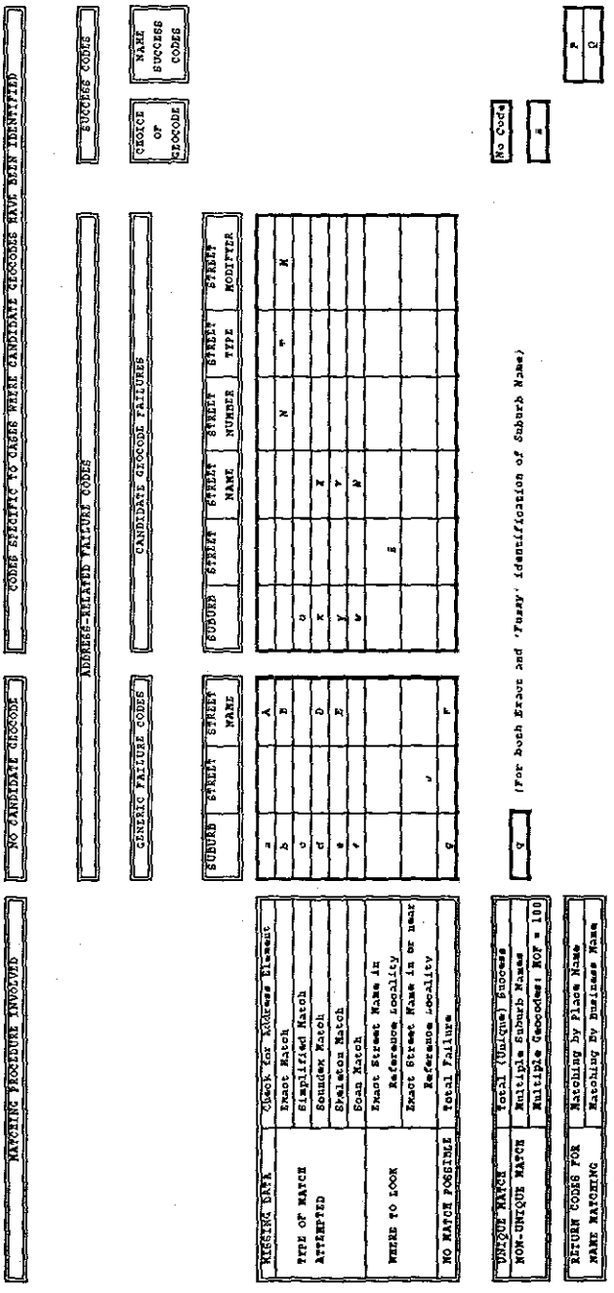


Fig 3. Summary of Return Code Definitions

Capital letter Return Codes are adopted to signal matching issues associated with street-level information. These are distinguishable from those relating to suburb names (which are indicated by lower case letters).

Once the suburb and street-level processes needed in order to generate a list of "good enough" candidate geocodes (or a single geocode) are complete, a final ranking of all these candidates is required in order to select the one which best fits the input address (and is therefore most likely to be correct). This is achieved by a multiplication of all the MOF codes generated at the different stages of matching. Preset threshold values are employed and the final product MOF is used to rank all remaining candidates in order to apply the decision rule which selects the one with the greatest MOF. In cases where two (or more) top ranking geocodes have the same MOF, the selection is relatively arbitrary and the first in the list is selected.

Ultimately, the Return Codes and MOF associated with a specific address do not allow one to say that the chosen geocode is correct. In combination however, they do provide guidance about the likely source of error and give some indication of how confident one can be in the final answer. Indeed, the use of preset threshold values at various stages of the matching process as a tool for restricting the list of candidate geocodes, itself implies that a match that encounters too many or specific types of problems is not "good enough".

In order to establish if the geocoding is "good enough" for data users to feel confident in its output, benchmark program runs are required on addresses where the 'correct' geocode is known. This can only ever be established by adopting some other geocoding process (such as having the addresses independently geocoded by a competent coder who has detailed local knowledge of the area involved). As such a manual process may also result in 'incorrect' geocoding results, comparison of the results from each process ultimately yields only a 'balance of probabilities' conclusion. TSG has concluded that the reported geocoding results are highly credible and yield results about which data users should feel highly confident.

Software Performance and Geocoding Quality

As noted earlier in this paper, on verification software runs 99.93% of all input records were geocoded and 98.08% of these were geocoded correctly (ie 98.01% of all input records were correctly geocoded). These results are summarised in Tables 1 and 2 below.

Program execution time varies depending on the quality and nature of the input address information. In cases where addresses are input as a single text string, they must be automatically parsed into individual address elements (street number, street name, street type, corner street, suburb, and postcode) before matching is attempted. In these cases approximately 30 records are geocoded per minute. In cases where input address elements are pre-formatted into separate fields and the data is relatively 'clean', approximately 150 addresses are geocoded per minute. As accuracy was always preferred to processing speed during software development, these execution times are regarded as highly satisfactory. Development of an unsupervised batch processing capability effectively removed concerns about program execution times.

Table 1. Output Analysis by Measure of Fit and Geocoding Quality: Percentage of Total Input Addresses

	GEOCODING QUALITY		
M.O.F.	% ACCURATELY GEOCODED	% WRONGLY GEOCODED	% NOT GEOCODED
1.000	76.56	0.42	-
0.900-0.999	9.46	0.46	-
0.800-0.899	1.85	0.09	-
0.700-0.799	8.06	0.33	-
0.600-0.699	0.33	0.10	-
0.500-0.599	0.72	0.17	-
0.400-0.499	0.12	0.06	-
0.300-0.399	0.21	0.06	-
0.200-0.299	0.44	0.12	-
0.100-0.199	0.26	0.11	-
0.000-0.099	-	-	0.07
TOTAL	98.01	1.92	0.07

Table 2. Percentage of Total Records Returning Different Number of Candidate Geocodes and Input Address Error Indicators

	NUMBER OF CANDIDATE GEOCODES (% TOTAL)						
GEOCODING CONDITION	0	1	2	3	4+	TOTAL	2+
Missing part of street information	-	14.77	1.81	0.4	0.27	17.25	2.48
Streeet name spelling errors	-	4.27	0.25	0.06	0.04	4.62	0.35
No suburb given	0.07	-	-	-	-	0.07	0
Suburb name spelling errors	-	0.18	-	-	-	0.18	0
No errors	-	74.53	1.5	0.49	1.35	77.87	3.34
TOTAL	0.07	93.75	3.56	0.95	1.66	100	6.17

Fig 4. Summary of Geocoding Quality by Measure of Fit

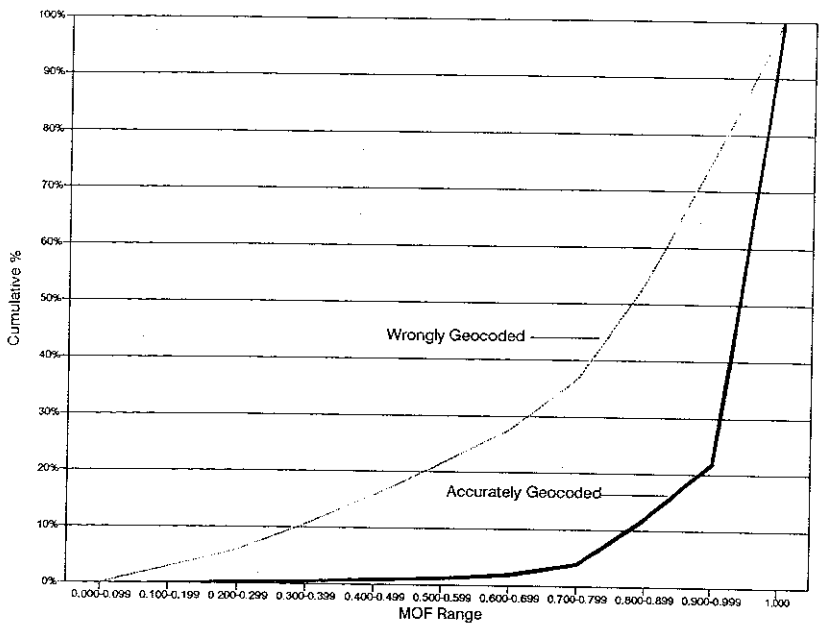
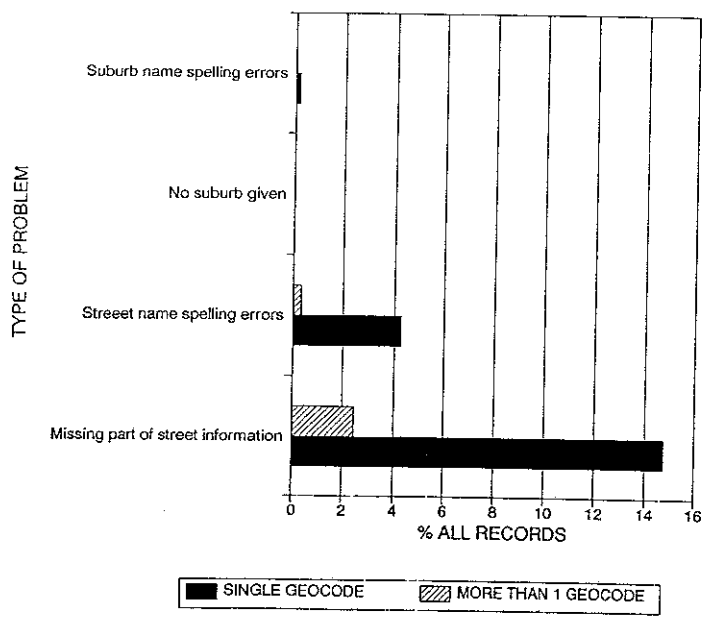


Fig 5. Summary of Geocoding Difficulties



Conclusions

TSG's geocoding task involved numerous pragmatic decisions and application of both well tried and newly emerging technologies. It was therefore far from an exact science and the very ambiguities associated with both reported addresses and street level reference material necessitated judgement about whether the resultant geocoding was "good enough". There are many practical file handling issues which severely curtail the potential benefits of SIS technology and vendor promises therefore need to be assessed against these limitations

Street reference information is relatively stable over time and administrative procedures could be put in place to update the relevant files for new developments. Both the geocoding software and the bulk of the reference material is therefore expected to have a relatively long operational life. Although geographic boundary systems change periodically, development of reference material is considerably easier when utilising a platform such as the files developed for 1991/92 geocoding purposes.

As noted in the introduction to this paper, the existence of this geocoding tool opens new opportunities to add a spatial dimension to any databases which contain address records, thus assisting their integration with data sets held by other organisations.

Glossary of Abreviations

Organisations:

ABS	Australian Bureau of Statistics
GNB	Geographical Names Board
TSG	Transport Study Group

Surveys/Data Series:

CVS	Commercial Vehicle Survey
HIS	Household Interview Survey
JTW	Journey to Work

Classification Systems:

ASGC	Australian Standard Geographical Classification
ASIC	Australian Standard Industrial Classification

Computer-based Geographical Analysis Tools:

SIS	Spatial Information System
GIS	Geographical Information System

Spatial Units:

CD	Census Collection District
SLA	Statistical Local Area

Geocoding Quality Indicator:

MOF	Measure of Fit

References

Cameron, R.J. (1987) *Australian Standard Industrial Classification*, Australian Bureau of Statistics, Catalogue No. 1202.0

Castles, I. (1988) *Australian Standard Geographical Classification (ASGC) Manual*, Australian Bureau of Statistics, Catalogue No. 1216.0