

The Challenge of Geocoding Large-Scale Travel Surveys

Andrew J. Smith

Abstract

Understanding travel behaviour is key to designing transport systems that meet both the current and future needs of the population. Travel surveys are a vital tool in providing estimates of transport demand and network usage. At their simplest level, transport surveys consist of recording the origins and destinations of travellers, sometimes disaggregated by travel mode and by time of day. More complex travel surveys consist of self completion travel diaries, on-vehicle interviews or focus group discussions where other variables may be collected such interchanges, journey purposes, the frequency of the journey and trade-offs between cost and other factors.

Irrespective of the scale of the survey or the number of respondents, obtaining accurate location information on traveller origin, destination, and any interchanges is a key requirement of all travel surveys. This is particularly true when the survey data are to be used for transport planning purposes. Unfortunately, many surveys are carried out in less-than-ideal conditions (roadside, stations, trains, etc.) hence the data yielded by the survey can be inaccurate or incomplete. Therefore, a mechanism – **geocoding** – is required for post-survey “cleaning” of the location information to an appropriate level of accuracy.

MVA Consultancy has designed, built, and used a geocoding tool – **TARA Geocoding** – which has been applied in various forms on several very large-scale travel surveys including the London Area Travel Survey (LATS), the DfT’s National Rail Travel Survey, and the Countryside Agency’s England Day Visits Survey. It has also been employed on many smaller regional-level surveys. To date the TARA Geocoding software has cleaned more than 10 million survey addresses. It has been used both as a “back office” application and as Computer-Aided Telephone Interview (CATI) software.

TARA Geocoding utilises a three-stage process to “clean” the location information present in travel surveys, namely: **auto-cleaning**, **manual geocoding**, and **logic checks**. These processes are described in the paper together with the results from different surveys. These demonstrate how TARA has significantly improved the geocoding rate of survey data including the coding of ambiguous locations using TARA’s advanced “geosearch” algorithms.

The TARA Geocoding software, whilst very successful, is effectively a tool for fixing a data collection problem. Collection of precise address information at the roadside is a difficult task. The

challenge is to get the data at the appropriate level of accuracy when it is collected whilst minimising the impact upon the time taken to complete the survey. Using our experience from previous surveys, we have developed a lightweight “portable” version of the TARA Geocoding application – **TARA Mobile**, allowing deployment on a handheld device. This allows the surveyor to collect accurate location data electronically and precisely at source. This reduces the amount of post-survey processing of data required. We are currently trialling this software on a small travel-based survey and the results will be presented to the conference.

The paper will review the benefits and challenges of hand-held devices versus traditional survey techniques and discuss the implications for designing surveys where location data is sought.

Keywords

Travel survey; surveying; geocoding; address cleaning

1. An introduction to travel surveys

Travel surveys are a standard technique used by Transport Planners and Market Researchers as a method for assessing why, when, where and how people travel. Travel surveys at their simplest level are simple roadside counts of the number of vehicles travelling along a stretch of road at a particular time of day. From this simple count, planners are able to make an assessment of what capacity of the highway is being used. Travel surveys can also be more complex. For example, a travel diary recording a person’s day-to-day movements, journey purpose, and mode of transport can provide a wealth of detailed information about how people use the transport system.

Travel surveys are generally undertaken using a mixture of manual techniques and automated techniques. For example, traffic counts are, in 2007, frequently still undertaken by a surveyor standing on a stretch of highway with a “ticker” counter. This is a cheap, relatively reliable, and mobile solution. Where traffic counts are required on a regular basis (for example to phase traffic lights or provide congestion updates), assessments of traffic volume can be automated using under-road traffic counters or over-road cameras with Automated Number Plate Recognition (ANPR) systems. Camera-based solutions are expensive to set up and generally immobile, however once installed they provide around-the-clock reliable information.

Travel diaries are generally undertaken using a traditional paper-based approach. Respondents are given a paper diary and instructed to complete it for each trip that they make stating, for example, their origin, their destination, the time of day, their mode of transport, and their journey purpose (e.g. travel to work). Travel diaries are usually undertaken on a large scale – the complexities of the survey usually require a relatively large sample size in order to yield accurate results. Therefore, the paper diary provides a cheap solution and is usually self explanatory to complete.

Electronic diaries can also be employed for smaller focus groups where, for example, the purpose is to understand better the activity patterns of a household and the trade-offs between transport supply, allocation of time budgets and other factors including cost, convenience, work and family

commitments, etc. While a useful tool in travel behaviour research there use in determining aggregate travel movements is of less value and so these tend not to be widely employed.

2. What is geocoding and why is geocoding important?

On a typical travel survey, the data collection exercise generates a vast quantity of data, usually in a paper-based format. For example, even simple roadside surveys may yield ten or more sets of counts (on paper). Household travel diary surveys may yield tens of thousands of paper diaries. Clearly, the data require some post-survey processing in order to turn them into a format that can be used by transport planners to make decisions.

The traditional approach to post-survey processing is as follows. First the data are keyed in to an electronic computer system. This exercise is often called “coding” and is usually outsourced to specialist data entry organisations. These organisations have the infrastructure in place to enable this work to be undertaken cost-effectively. However, not all data errors are picked up at this stage. For example, “High Street, Clandon, Surrey” may be corrected to “High Street, Clandon, Surrey”, but in reality “High Street, Clandon, Surrey” may not exist.

Correcting addresses - beyond simple typographic errors - is part of the geocoding process. Geocoding is “...the process of assigning geographic identifiers (e.g. codes or geographic coordinates expressed as latitude-longitude) to map features and other data records, such as street addresses.”¹ Geocoding is a vital process if survey results are to be used for transport planning purposes. The level of precision of geocoding depends upon the purpose to which the survey results are to be put. For example, if the purpose of the survey is to ascertain pedestrian routes from car parks to a shopping centre, a higher level of precision is needed than for a survey of the distance travelled by shoppers to the shopping centre. For the former, a precision to less than one metre level would be acceptable. For the latter, a precision accurate to around 10 metres would be acceptable.

Geocoding a relatively small number of addresses is a simple process. Many of us in the UK will have used, for example, the Royal Mail’s on-line service to confirm a domestic postcode. When we book theatre tickets over the telephone, we are often asked only for our house number and postcode. This is effectively small-scale geocoding. Using simple address lookup software it is possible to geocode hundreds of address locations within one or two hours. However, faced with 100,000 addresses to geocode, an automated (or semi-automated approach) is required.

3. Geocoding Large-Scale Travel Surveys

One of the largest travel surveys in the UK is the London Area Transport Survey (LATS) conducted approximately every ten years by Transport for London (TfL). The 2001 survey yielded almost 7

¹ Wikipedia

million addresses, all of which required geocoding so that the data could be used by local authorities, developers, transport researchers and other interested parties. Achieving a high-level of precision on all address locations was a key client requirement so as to maximise the potential uses for the data both inside and outside of the London area. TfL commissioned MVA Consultancy to provide a specialist geocoding solution for LATS. Geocoding 7 million addresses using a manual lookup process would take an unacceptably long time. Assuming each address takes one minute, this would equate to around 16,000 person days. To automate the process an integrated piece of software called “The Transport Address Resolving Application” – or TARA – was developed to undertake the geocoding. TARA consists of three stages:

- An automatic **batch processing** of address lookup;
- A suite of semi-**manual geocoding tools**;
- A set of sophisticated **logic checks**.

The survey addresses were “denormalised”, or “flattened” into a standard format based around the UK Post Office Address File (PAF) – the *de facto* standard for UK addresses. The LATS data were then loaded into an Oracle database in the standardised format.

The three TARA process are described in more detail below.

Automated Batch Processing

Once imported, the TARA system then passed the denormalised survey addresses automatically to a sophisticated address lookup routine fully integrated into TARA. The automated address lookup routine comprises a series of gazetteer lookups and address lookups. The address lookup routine used a product called QuickAddress Batch² which is widely used in address lookup products. The QuickAddress product provided a high resolution geocode (one metre accuracy) when a large number of key address fields were supplied (for example house number, street name, post town). However, a key remit of the project was to provide a geocode for *every* survey record at the *highest possible level of accuracy* but without making unnecessary assumptions about the source data.

QuickAddress could not provide a geocode for addresses where, for example, only the place name was supplied. For this reason, the consultants built and integrated gazetteers into the automated address lookup routine. The gazetteers consisted of sets of locations and an associated geocode, e.g. “West Clandon, 503070, 152895”. These gazetteers were produced using Geographic Information Systems (primarily MapInfo Professional and ESRI ArcGIS). Additionally, gazetteer data from other sources including airports, ports and businesses were included in the system.

² QAS Limited

Using the combination of QuickAddress and the gazetteer lookups, the automated batch processing routine was balanced to ensure that a large percentage of the 7 million addresses could be matched automatically, but the routine would not make unnecessary assumptions about locations. For this reason, the raw addresses needed to be of a reasonably high precision and accuracy to be matched by the automated batch process. For example, “8 Park Avenue, Barking, IG11 8QU” was matched by the system to “8 Park Avenue, Barking, Essex, IG11 8QU, Easting=544636, Northing=184663”. Very few addresses at a lower quality were auto-matched by the system, thereby minimising the chances of incorrect assumptions being made about the location.

From the TARA user’s perspective, the automated batch processing was triggered by a single button.

Semi-Manual Geocoding Tools

The TARA automated batch processing routine successfully geocoded around 50% of the LATS survey addresses, thereby substantially reducing the task of manual geocoding.

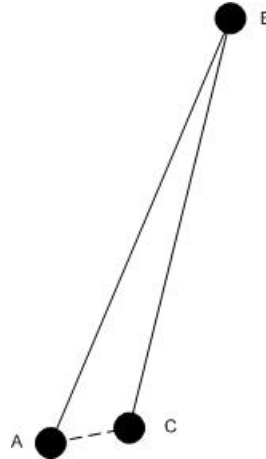
The next stage of processing consisted of **semi-manual geocoding** using a data entry form designed specifically for this system. The addresses which failed the automated batch processing were passed to a team of geocoders for semi-manual geocoding. The TARA system was installed on around 30 PCs at the consultant’s office. TARA then displayed each of the “failed” addresses to the team. The geocoders were presented with the raw survey address and were provided with a suite of tools to enable them to look up the address. The tools consisted of QuickAddress (in “manual” mode), the gazetteers, and a map display (GIS). In the majority of cases (over 80%), the geocoders were able to geocode addresses successfully using these tools. In a minority of cases, the geocoder had the option of passing the address in question to a “senior geocoder” for resolution.

For example, geocoders used the train station gazetteer to geocode “Porchester road opp royal oak underground w2” to “Royal Oak Station, London, W2 6ET, Easting=525771, Northing=181580” and the map display to geocode “Porchester road near library w2” to “W2 6HU, Easting=525784, Northing=181402”. In both cases, the geocoder managed to locate a suitable match using TARA and hence produce the desired level of output, i.e a 6-digit geo-referenced record.

The TARA system includes tools to manage the geocoding team such as monitoring geocoder performance (number of addresses geocoded per day), and allows a supervisor to select randomly a series of addresses to monitor the quality of the geocoders’ work. Additionally, the semi-manual geocoding screen presented to the geocoders limited the time that a geocoder could spend on a single address. This ensured that predictions could be made to the client on when the job would be complete.

Logic Checks

The final stage of TARA's geocoding process consisted of a series of **logic checks**. Many of the surveys undertaken as part of the LATS project consisted of multi-address responses. For example, travel diaries consist of a series of inter-related addresses such as "home", "work", "shops", "home". It would be possible, for each of the individual addresses to be geocoded correctly, but for the journey as a whole to be illogical. For example:



The diagram above illustrates a journey from A to B to C. In the absence of other information, the journey appears illogical as the distance from origin (A) to interchange (B) and interchange (B) to destination (C) is far greater than the overall distance from origin (A) to destination (C).

The TARA system incorporated an automated check – performed by the user clicking a single button - that measured the angle (at B). If the angle was less than a predetermined threshold, the journey was flagged as potentially illogical. The potentially illogical journeys were then passed to the geocoding team for inspection. Some of the journeys were not in fact illogical – a natural boundary such as a river between A and C, or the absence of a road between A and C may force the respondent to travel from A to B to C rather than directly from A to C. However, a large percentage of records may be incorrect due to data capture errors (by the respondent), data entry errors, or geocoder errors. The routine therefore prompted further investigation and data cleaning to ensure that the survey data were of the highest possible quality and fit for purpose.

TARA incorporates many other logic checks to improve data quality such as.:

- **Speed against mode of transport check.** For surveys where the mode of transport and journey times are recorded, this check ensures that the speed of travel is appropriate for the mode of transport. For, instance, an average speed of 40 mph would be inappropriate where the mode of transport was cycle, but would be appropriate where the mode of transport was train.
- **Distance to rail station check.** For surveys which included rail stations, a threshold was set for travel to a rail station. For example, a journey by car of greater than 50 miles would be inappropriate.

- **Trip reversal check.** Occasionally data entry errors resulted in multi-trip surveys (e.g. travel diaries) being input in the reverse order (i.e. destination then origin). A check was built into TARA where the respondent's origin address was closer to the destination station than the origin station, or where the respondent's destination address was closer to the origin station than the destination station. When both origin and destination are flagged, a trip reversal error is implied and the survey is sent to the senior geocoder for investigation.

The logic checks, and subsequent data investigations by the senior geocoders, are the final stage in the geocoding process before the cleaned survey records are exported and presented to the client.

The fact that TARA encapsulates all of the above functionality into a single, integrated database application that is easily installed has made it a highly attractive tool for other large-scale travel surveys. Beyond the LATS surveys, TARA has been adapted for many clients' surveys including those undertaken by the Countryside Agency, Department for Transport, and Local Authorities.

These other travel surveys have required modifications to TARA. The key change made for one survey was to use TARA for both data entry and geocoding in one step. That is, the TARA user interface was used for keying in questionnaire data. As the data were keyed in, the geocoding inherently occurred at the same time. The combination of data entry and geocoding speeded up the process of going from survey to geocoded result considerably. Whilst it is impossible to compare directly the data quality from one survey to another, it was felt that data quality was also improved as the data passed through fewer steps.

Additional enhancements made to TARA were made to the Geographic Information System (GIS). For certain surveys, additional boundaries have been added providing more detailed information about respondents' travel behaviour. For example TARA has been adapted to display Areas Of Outstanding Natural Beauty and National Parks. Using the integrated GIS functions in TARA, it has been possible to ascertain very quickly respondents' travel behaviour and usage of National Parks. These levels of geocoding would be difficult, if not impossible, to undertake using simple address lookup algorithms.

4. Beyond The Transport Address Resolving Application (TARA)

Over the last 3-4 years, the TARA system has proved to be very successful in producing high quality results for large-scale travel surveys. The automation of many tasks that hitherto have been performed manually has reduced the time and effort involved in post-processing survey data, and through the application of techniques like geocoding and geo-logic checks the accuracy and completeness of the data has much improved.

Even so, there are several enhancements that could be made to survey data collection to improve the quality of the geocoded results. Elements of TARA could be used to enhance this survey process.

Whilst the TARA system undoubtedly adds value to the data through the sophisticated GIS and post-processing routines, its basic purpose is in fixing poor quality data. Collection of precise address information at the roadside, for example, is a difficult task. Inclement weather, the dangers of undertaking roadside (or indeed household surveys) and data collector training all contribute to data quality. The challenge is to get the data at the appropriate level of accuracy when it is collected whilst minimising the impact upon the time taken to complete the survey and the safety of the participants.

To address this requirement a scaled-down version of TARA, called TARA Mobile, has been developed that can be used in a Computer Aided Personal Interview situation (CAPI), effectively replacing paper and pen. With this system, a questionnaire is built on a single desktop computer. The questionnaire designer sets which questions will elicit text-based responses (Yes/No, Comment, Name, *etc.*), which questions will elicit address-based responses, and which questions require the use of a GIS. The questionnaire is then “deployed” (sent to) one or more handheld PocketPCs or PDAs. The handheld devices then allow interviewers to collect questionnaire responses. The handheld(s) are then “docked” back at base, and the questionnaire responses are transferred to the desktop computer or server for further analysis (data can also be transferred via wireless connection where this is set-up). TARA Mobile can be run on a tablet PC or handheld PocketPC/Smartphone. In a personal interview situation, interviewees can then, for example, point to a map to show their location rather than having to describe their location in words.

A key challenge facing transport or land-use planners is assessing the level of use of large open areas such as National Parks or Areas Of Outstanding Natural Beauty where there are few addressable locations. A National Park, for example, may contain only a few locations recognised by Royal Mail (*e.g.* Public Houses, Farms, *etc.*). Assessing usage of a National Park therefore requires a different survey technique. It would be possible to monitor through a traditional travel survey (personal interview or travel diary) which car park the respondent used or where they left the road. However, describing their journey *through* the National Park would be impossible using addressable locations.

A possible solution would be to use Global Positioning System (GPS) devices which are carried by respondents. GPS devices are relatively inexpensive (a consumer device typically costing between £50 and £100). These GPS devices can store a date-stamped “breadcrumb” trail (geocodes) of the respondent’s location to an internal memory card. When back at base, the data can be transferred from the GPS device to a central database (TARA, for example). The “breadcrumb trail” can then be electronically re-created in a GIS system (or a modified version of TARA) and the “route” plotted for further analysis. A further benefit of using GPS units is that they could be used in conjunction with personal travel diaries. If respondents were asked to keep travel diaries, a GPS device could be used as a “calibration” of respondents’ travel diaries. In other words, the travel diary could be “corrected” based on the results of, for example, one day’s use of the GPS unit.

As mobile devices such as PocketPCs, Smartphones, tablet PCs, and GPS devices become cheaper, more “ruggedized” and easier to program, it is highly likely that they will become increasingly prevalent in travel surveys. Whilst there is a larger up-front cost to the survey in procuring and configuring these devices when compared to a traditional paper-based survey, the benefit is a reduced

downstream cost in post-processing data. Additionally, and perhaps more crucially, these devices do not simply automate an existing survey process. Rather, they enable the collection of a far greater range and depth of information about people's travel behaviour that is not possible with traditional survey techniques.

About the Author

Andrew Smith is a Managing Consultant at MVA Consultancy. MVA provides advice on transport and other policy areas to central, regional and local government, agencies, developers, operators and financiers. Andrew has a degree in IT and is a PRINCE2 qualified project manager and Chartered IT Professional. Throughout his career, he has worked on many software development projects as a programmer, systems analyst, consultant, or project manager. Andrew has extensive practical experience of a wide range of technologies, including client/server and web-based development tools, relational databases, address processing, and Geographic Information Systems. Andrew is a technical expert in the Oracle relational database management system, applied in particular to solving transport-based problems and has written and delivered training courses in this field.

Andrew has developed MVA Consultancy's address geocoding business, having been responsible for the delivery of several large systems for clients including Transport for London (TfL), the Department for Transport, and the Countryside Agency. He project-managed the delivery of a system to provide statistics on the effect of the Congesting Charging scheme, and also the migration of a transport network modelling system to an Oracle database for a rail authority in Hong Kong.

Andrew has developed several systems for public authorities including London Buses, West Yorkshire PTE, and Hertfordshire County Council. He has also worked in the area of finance, developing and delivering a treasury management system.

Andrew can be contacted at: MVA Consultancy, Dukes Court, Duke Street, Woking, Surrey, GU21 5BH; Tel. 01483 742 870; e-mail asmith@mvaconsultancy.com .