B02.1

Sources of Data – a Transport Planning Case Study

Hugh Neffendorf, Katalysis and Ian Williams, ME&P

Abstract

Many disciplines require information from a variety of sources. Transport planning is a particularly heavy user of data. This includes data of several types:

- Economic
- Statistical
- Surveys
- Land use
- Operational
- Transport network

Most of these have geography as a key dimension.

Among the issues that face users of data sources are:

- Data content
- Data quality
- Responsibility
- Maintainability
- Accessibility
- Geographic detail
- Coverage
- National variation
- Standardisation & harmonisation
- Ability to forecast

This paper reports on a project undertaken for the Department for Transport (DfT) that investigated sources of data to support the transport planning process. This project is concerned with freight transport and the modelling methods used for forecasts of travel. The investigation is considering existing data sources, emerging sources and the needs of new planning methods. It has involved interviews with a large number of public and private sector providers of data and considers the issues involved in assembling a

comprehensive set of information for a key policy topic. The report also considers options for a repository of data, ranging from a metadata service to a full data bureau, operating either in the public or private sectors.

This project is likely to be of interest to organisations that need a range of information to perform their functions. It considers the issues that face those involved in data assembly and highlights some important features of new data sources. The paper will pay particular attention to the geographic information dimension.

Introduction

This paper describes aspects of work undertaken in a Review of Freight Modelling for DfT. Part of the project considered data that is currently available and data in current use for such modelling, as well as emerging sources of data that might be useful. This included a commentary on the issues associated with the data, and its strengths and weaknesses.

Many types of data have a geographic dimension and issues that are associated with this. However, there are some types of information, such as economic series or operating costs that are not spatial in nature. This paper concentrates on the data likely to be of interest to the geographic information community.

Types of Data

Data to support transport modelling can be considered under six broad headings:

1. Economic/financial data – e.g. operating costs, tariffs, wealth indicators
2. Statistics – e.g. trends, import/export, vehicle registrations
3. Surveys – e.g. interviews with operators, roadside interviews, traffic counts
4. Land use data – e.g. area usage categories, usage quantities, gazetteers
5. Network data – e.g. link/junction locations and types, speeds
6. Operational data – e.g. logistics relationships, transhipment locations

In addition, it is useful to consider data in terms of a categorisation of use:

- For building causal relationships
- For disaggregation of other data or results
- For validation
- For forecasting
- As controls

Organisations Consulted

During this review, project team members met with, or otherwise contacted or read material from, the following:

- DTLR (TSF and PLUS Divisions)
- Office for National Statistics (Neighbourhood Statistics & Business Data Divisions)
- Valuation Office Agency
- Information and Development Agency (IDeA) for local government

- Greater London Authority
- Transport for London
- Transport & planning consultants
- Oscar Faber
- WS Atkins
- MVA (LTS)
- Roger Tym
- Private data firms
- Intelligent Addressing
- Point X
- CACI
- Kingswood
- Geoplan
- JMP (TRICS)
- University researchers (UCL, CASA)

Descriptions of Data Sources

This section comprises the main part of the paper and contains brief descriptions of data sources that have a spatial dimension and highlights issues that arise. Particular attention is given to relatively new sources or those that are valuable but under-used.

A Surveys of Transport

CSRGT and Other Road Freight Surveys by DfT

In order to obtain information on the broad spatial pattern of road goods transport the DfT carries out three continuing freight surveys:

- CSRGT – the Continuing Survey of Road Goods Transport
- IRHS – the continuing International Road Haulage Survey
- RoRo – the quarterly Roll-on/Roll-off Enquiry

The CSRGT is the only national survey of road haulage activity, and is a statutory survey of heavy goods vehicles over 3.5 tonnes gross vehicle weight in the UK. The survey is based on a random sample of trips made by about 400 vehicles each week. At present the geographic resolution is at County level, and this is insufficient for many purposes.

A number of suggested improvements to the CSRGT have been made, notably:

- Collect land use data of origins and destinations, to enable the reason for which the goods were being moved to be identified (e.g. from factory, warehouse or modal terminal, etc.)
- Collect a finer geographic resolution of trips, at least to Local Authority District level
- Increase the sample size so that improved estimates of District to District freight movement can be derived.

DfT is in the process of conducting a quality review of its freight statistics, and has included these items, among others in the list of topics for review.

GPS Vehicle Tracking Surveys

There is an increasing tendency for large fleet operators to install vehicle tracking systems in lorries, generally using Global Positioning Satellite equipment. In addition, there are a number of examples overseas of vehicle activity surveys using GPS equipment that has been provided to drivers for the survey. The GPS device needs to be accompanied by some other data recording equipment to synchronise locations with information on the activity being conducted. The great value of a GPS reading in this context is to give an accurate location, which can be used on its own or to verify separately collected address information.

DfT is interested in the potential for GPS surveys, particularly now that the equipment price has made such surveys feasible on a substantial scale. We can expect to see GPS trials for freight movement surveys in future.

Roadside Surveys

Roadside interview surveys are commonplace as part of transportation studies and road scheme assessments. Typically, both cars and commercial vehicles are intercepted, and the latter are further distinguished by vehicle type. A normal sample ratio at a site will be 10%, although this will vary by traffic density and time of day. Interviews are necessarily brief, in order to avoid traffic problems at the site and to maintain the sample rate.

From a spatial analysis perspective, some problems with roadside interview data are:

- Land use of origin/destination is seldom collected due to limited time available
- It is often difficult to obtain good quality address detail from drivers
- Geoding of the addresses is difficult, since they are often poorly structured, incomplete and non-postal in nature

The latter issue, in particular, was covered in Neffendorf, Ramsey and Walker, Address Matching and Geocoding for Major Surveys, GIS 1999.

The DfT, in collaboration with the Highways Agency, has initiated the National Origin-Destination Transport Survey Project to assist organisations involved in the identification, selection and use of appropriate origin-destination transport data. The National Origin-Destination Databank now contains information about transport origin-destination surveys for:

- Roadside interviews
- Public Transport
- Home Interviews

The database contains information on the location and broad type of data available, including central contact details of the organisations that hold the data. Further details on the databank can be accessed at

http://www.roads.dtlr.gov.uk/roadnetwork/heta/datasource/databank/

which also enables the user to download and use the database application to identify suitable survey locations.

An important aspect to note about this database application is that it does not contain actual traffic data itself from the interviews, or any detailed information on the data definitions of the data that was collected. While the database does contain information relating to a wide body of data, the considerable differences in methodology and definitions between surveys implies that assembling them into a consistent source of integrated data covering a large area would be likely to be very difficult. There is a considerable need for standardisation in this area.

B Land Use Data

There are several categories of land use data that can be considered:

- Classifications of areas (polygons) of land, without further identification of the intensity of use or development – e.g. just that an area is agricultural, housing, industrial, etc.

- Quantities associated with land use, usually in the form of statistical values for zones – e.g. retail or industrial floorspace, number of employees, number of houses, tourist numbers, etc.

- Gazetteer data, that lists properties, addresses or items such that their position is known. Sometimes, additional information is available with gazetteer entries, for instance the type of use of a property at an address.

In general, it is the latter two types of land use data that are of most value for freight modelling. The remainder of this section summarises the main sources of land use data of these three types.

Population Census

The 2001 Census was taken on 29 April and is being processed currently. Small area statistics are due to become available in summer 2003. The Census Web link is:

http://www.statistics.gov.uk/census2001/

Census data is used extensively as a basis for person trip models, but less so for freight. The main elements of data that are relevant for freight studies are:

- Numbers of houses, which can be generators for some freight trips. These will be available in Output Areas, which are reporting units of about 40 Households on average. Additional data about the classification of households may also be used.

- Numbers of employees by SIC at their usual workplace, derived from the Journey to Work tables. Note, these are normally only available at Ward level, although special processing to smaller units by ONS may be possible. SIC is a useful indicator, but can be misleading as to the actual activity at a site, as described later.

Neighbourhood Statistics

This is a major new series of statistics arising from the work of the Social Exclusion Unit. It will include Census data but, otherwise, currently is focussed only on person characteristics. It is not yet of value for freight modelling, but that could change as more information is brought in line. In due course (not before 2005), business-related data is expected to become part of Neighbourhood Statistics. The Scottish equivalent is independent and different. Developments in Neighbourhood Statistics should be monitored. The Web link is:

http://www.statistics.gov.uk/neighbourhood/home.asp

NLUD

The National Land Use Database is being developed by ODPM, IDeA and other partners, with the ultimate aim of providing a complete and maintained record of land use for England. It has a well-developed set of land use classifications, which can form the basis for other data gathering exercises. In principle, it appears to have potential value for transport modelling, but that value is restricted currently by two factors:

- At present, the only land use series that has been collected nationally is of Previously Developed Land, related to Government directives on use of brownfield sites. There is no firm plan as yet for collecting the other series, although research is ongoing.

- The current definition of land use relates only to area usage descriptions, and not to any quantification of usage. That may change in future.

Among the experiments being conducted under the NLUD banner, one in South Yorkshire, with Ordnance Survey, is particularly interesting. Ordnance Survey is developing its new digital map product, MasterMap, in which all map objects are polygons. Thus, individual building or property outlines, for example, can be identified as a unit, which was not possible previously. By associating the polygons with other available data (e.g. address gazetteers), there is a prospect of identifying more land use information automatically, and of allocating quantity or type attributes to the data. It is too early to say if this will be fruitful, or when, but early results show some promise, alongside some difficulties.

Thus, NLUD is currently of little practical use for freight modelling. As a general point, however, since NLUD is a formal national initiative, it is appropriate to respect its land use category codes when defining any land use application, such as a new survey. The NLUD Web link is:

http://www.nlud.org.uk

NLPG

The National Land and Property Gazetteer is an initiative of the IDeA, and has arisen as a data source for the National Land Information Service (NLIS), which supports property searches for conveyancing. The NLPG is intended as a gazetteer containing every address in the country with a unique property reference number (UPRN) and a grid coordinate reference. Most addresses are conventional properties that would receive post, but the NLPG is designed to also cover non-postal addresses.

The NLPG is being assembled by an amalgamation of individual local authority gazetteers, using a variety of address lists. At present, some 250 authorities (of 470) have signed up to the scheme. There are some questions about consistency and commitment to maintenance, but the NLPG has considerable official momentum and is being formally adopted or mandated for some applications (e.g. the Electoral Register address standard). The Scottish equivalent Master Address Database (MAD) is similar in definition, is endorsed by all Scottish authorities, but is less advanced in implementation. The NLPG Web link is:

http://www.nlpg.org.uk

The potential value of the NLPG for freight modelling lies in its aim to be the master address list for Britain, with potential to link to other data sources (e.g. classification details of all property, that could be used for trip generation estimates or for data disaggregation). It is not ready for this yet, but a data strategy would consider use of the UPRN as a linking mechanism.

PAF/ADDRESS-POINT

The Postcode Address File is Royal Mail's master list of addresses for UK postal operations. It has become the most widely used address list for many applications that are beyond its original purpose.

It is expected that PAF data will become a component of the NLPG, although this is not yet confirmed. Until the NLPG is complete, PAF can be used on its own or to fill gaps. It has a degree of property classification detail, but this is not refined.

Ordnance Survey, in partnership with Royal Mail, has created ADDRESS-POINT. This provides an accurate grid reference for each address in PAF, and makes the data more valuable for many purposes. A derived product, Code-Point, gives the mean grid reference of each postcode, which has value in geocoding transport surveys.

Valuation Office Agency data

The Valuation Office Agency (VOA) assesses the rateable value of all property in England and Wales for business rates and council tax. It assembles three main databases

- Non-domestic property
- Council tax (i.e. residential property)
- Property price (both residential and non-domestic).

Since there are important changes in the availability of VOA data, it is given considerable coverage here.

The VOA database for non-domestic property is the source of most relevance to freight modelling. It comprises two main data sets:

- the Rating Lists
- the Rating Support Application (RSA).

The key difference between these is that all non-domestic property is recorded in the Rating Lists (which do not contain floorspace data), whereas only around three-quarters of that property is also recorded in the RSA (which does contain floorspace data).

Individually rated properties are known as hereditaments and these form the basic unit of data. A large office or mixed-use commercial building will, if shared between several tenants or owners, consist of several hereditaments. These may occupy some floors, part of a floor or space in, adjacent to, or associated with the building. The only types of non-domestic property that are excluded from the Rating Lists are those which are exempt from rates: churches and other places of worship, and agricultural land and buildings.

The Rating Lists are legally defined and **public** documents that list all 1.7 million non-domestic hereditaments on which rates are paid. The Lists comprise a simple set of data items for each hereditament including:

- the address
- the rateable value
- two activity codes (the primary description, and the special category code
- and several codes for monitoring changes in hereditaments.

In contrast, the RSA data is **not public** for reasons of confidentiality, so access to data on individual properties is not made available outside the VOA. However, after a long break, the aggregate floorspace data from the RSA is once again being published at the local authority district level (DTLR, 2001). There used to be regular publication until 1986 but the only publication since then was of the 1994 data, in an incomplete form.

The RSA holds information for around 1.3 million hereditaments. These correspond to almost all of the hereditaments which are within what are termed the four "bulk classes" of shops, offices, factories and warehouses. As well as floorspace data the RSA contains two further activity codes (Building Use and Standard Industrial Classification) and some other information.

The VOA has, in recent years, become interested in the commercialisation of its data and has been in contact with the PLUS Division of ODPM to discuss the potential for linkage of the VOA data with other sources available to ODPM, which already has access to the Rating Lists data. This VOA interest in making data available for wider use represents an important step forward from the past, although the requirements for confidentiality of the RSA detailed data does introduce significant restrictions on its usage. However, through aggregation of data and the use of anonymised samples with coarse spatial definition, substantial benefit could still be obtained from the RSA.

The main potential usefulness of the non-domestic property databases for freight modelling lies in the ability to identify at a detailed spatial level the likely generators and attractors of freight shipments. They can complement data on employment from the IDBR (see below) that could play a similar role, since both data sources have different deficiencies for this purpose. Because of its usage for taxation purposes, the VOA data is necessarily kept up-to-date. This makes it of particular usefulness for modelling purposes, compared to one-off surveys that may rapidly become dated.

The Rating Lists data are available at the individual hereditament level including the full postcode so that each property can be located spatially. The Internet service at

http://open.voa.gov.uk/apps/rating/irlw_2000.main

enables any user to download the Rating List information freely for a user specified set of properties, though of course for modelling purposes a proper database of hereditaments would need to be obtained through official channels. The Primary Description code enables the type of each individual property to be distinguished, while the rateable value provides a coarse guide to the scale of the activity underway at this property.

Because the Rating List data are classified based on the use of the property, rather than on the SIC of those employed therein, they avoid one of the major problems associated with employment oriented databases. SIC based employment data has been shown to be misleading for many purpoes, since it reflects the overall nature of a business rather than the function of a premises within that business.

It may be reasonable to assume that, for a given Primary Description code in a specific district, the volume of such floorspace at each hereditament in that district is likely to be coarsely proportional to its rating value. This then provides an approximate means by which the floorspace data of the RSA could be disaggregated spatially to a zoning system that operates below district level. The use of the volume of floorspace by Primary Description code as an estimator of local freight volumes, is likely to be much more comparable across different regions than would be rateable values because of the pronounced regional variations in the levels of the latter.

The ideal course of action would be to use the NLPG UPRN or other address link to combine at the individual entry level the IDBR database on employment by site with the Rating List database on the premises in which this employment is located. The combined information on the size of workforce and the type and the value of the property would be of particular benefit. Creating and maintaining an updated linkage between such databases would be a major task, given the wide range of exceptional matches that would need to be analysed individually. Nonetheless such a database would be likely to have benefits for many other purposes across government beyond freight modelling.

There are some caveats that need to be taken into account with respect to the VOA data:

- The data in Scotland is not collected by the VOA and so would require separate study.
- The data for utilities and onetime government owned enterprises are complex and in a process of transition. For example all Railtrack properties are still treated as a single hereditament with no further spatial definition – this means that rail freight depots are not distinguished. Some but not all of the other utilities have already been through the process of being subdivided into normal localised hereditaments. Since the year 2000, government owned properties have been included in the Rating Lists.

IDBR

The Inter Departmental Business Register is managed by the Office for National Statistics (Business Data Unit). It is a comprehensive national database of all businesses that are registered for PAYE or VAT. It contains information on company employees, turnover, business type and addresses. It is used principally for selecting samples for surveys of businesses, to produce analyses of business activity and to produce lists of businesses. It can also be used for some administrative purposes.

The IDBR combines the former Central Statistical Office (CSO) VAT based business register and the former Employment Department (ED) employment statistics system. It complies with European Union regulation 2186/93 on harmonisation of business registers for statistical purposes. Both the CSO and ED previously collected key economic data from businesses for use in compiling the national accounts. The use of two different registers resulted in inconsistencies, including different estimates of employment, because of differences in classification and coverage. The IDBR leads to a single and more reliable set of employment estimates, improving the consistency of the national accounts and the quality of the productivity and unit wage cost estimates.

The IDBR covers all parts of the economy, but misses some very small businesses (self-employed and those without employees and low turnover) and some non-profit making organisations. There are around 3.7 million businesses in the UK of which 2 million are on the IDBR. The IDBR provides more than 99% coverage of economic activity.

It is updated using data from 3 main sources:

- Value Added Tax (VAT)
  Detail of businesses registered for VAT.
  Covers 1.7 million traders.
  Provided by HM Customs & Excise daily.
- Pay As You Earn (PAYE)
  Details of employers with employees in PAYE schemes.
  Covers 1.1 million employers.
  Provided by Inland Revenue quarterly.
- NS Surveys
  1,224,642 inquiry forms sent to 274,324 businesses in 1999

The information held for individual business units is:

- *Name*
- *Address*
- *Classification (industrial/economic activity)*
- *Employment*
- *Employees*
- *Turnover*
- *Legal Status (company, sole proprietor, partnership, public corporation/nationalised body, local authority or non-profit body)*
- *Enterprise Group links*
- *Country of ownership*
- *Enterprise Zone markers*
- *Company Number*
- *Value of goods traded with EU member states from Intrastat*

All the data on the IDBR are treated as RESTRICTED COMMERCIAL (i.e. confidential), although the IDBR can be made available to public sector organisations, subject to signed undertakings of purpose for which it will be used.

The IDBR Web link is:

http://www.statistics.gov.uk/themes/commerce/services/idbr.asp

As mentioned earlier, a valuable data source for freight modelling could be established by linking the IDBR with the VOA Non-domestic Property and Rating lists. This would combine business type, name and turnover with floorspace and an indication of property (rateable) value. There will be confidentiality issues associated with such data combination, but the potential should be investigated.

Central Business Directory

The CBD is being created by the ONS Business Data Unit, alongside the IDBR. The concept is to enlarge on the IDBR, by identifying businesses that might otherwise be missed, and to overcome most of the issues of confidentiality. This is being achieved by using publicly available sources, such as Yellow Pages and private sector data suppliers.

The CBD is in trial mode, and not yet ready for use. It has promise as a data source for freight modelling, and should be monitored.

Local Authority Data

Individual local authorities have been the source of much land use data that is used in local transport planning. This is unlikely to continue, being hampered by inconsistency between authorities and, increasingly, difficulty in obtaining access or resources.

Discussion with the Greater London Authority, for example, revealed that it does not hold a register of land use, but only a database of property planning applications. It was also noted that individual London Boroughs were inconsistent in their interest in land use data, and that pressures on cost were reducing such activity. The consultant that is preparing new land use data and projections for the London Transportation Studies model confirmed that this is the case, and it is necessary to rely increasingly on (improving) national data sources.

Private Data Firms

There is a growing availability of land use and demographic data from private firms. Prominent companies that publish catalogues of such data include CACI, Geoplan, Kingswood and MapInfo. Among the types of relevant small area or grid referenced data are:

- Enhanced versions of standard data, such as Census
- Neighbourhood classifications (e.g. ACORN, MOSAIC, PRIZM)
- Population estimates and projections
- Income estimates
- Lists of companies by location and type (e.g. petrol stations, pubs)

Such data is often only available in convenient form from private suppliers, and the range of data is growing. However, some customers of data (e.g. mailing list users) are less concerned about completeness or data quality than transport modellers would be. Thus, it is necessary to understand the quality of any data source before using it.

Nevertheless, this is potentially a rich, and underused, source of data for modelling. It is recommended that a separate detailed review of the suitability of such data is needed, involving discussion with suppliers and experimentation with data to gauge its quality and utility.

C Network Data

Traditionally, transport modelling projects have assembled transport network data in a relatively crude manner, often just using paper or digital maps to extract link lengths and then associating characteristics, such as speed and capacity, with the links by reference to link type. With a growing attention to congestion, junction types have also been obtained, again, usually from maps, but sometimes involving field survey or requests to local authorities. Identifying constraints, such as banned turns, one way streets or height/weight restrictions has typically involved local knowledge. A notable case is the London Transportation Studies, which had the benefit of a comprehensive inventory of the network for 1991. However, it has become increasingly out of date, raising issues for representing the current situation.

Two relatively recent developments offer improved sources of data for network creation and updating. These are:

- Commercially available aerial photographs as a GIS layer, which can be used for identifying current lane layouts and junction characteristics;
- Detailed network data files from suppliers of in-car navigation data, such as Navtech. These are now available commercially, and include speed assumptions, junction descriptions and movement constraints.

These sources are not exclusively of interest to freight modelling, but to transport planners in general, particularly in the urban context.

Main Findings

The conclusions that have been derived from the review of the data potentially available for use in developing freight models in GB can be summarised as follows.

There is a substantial amount of data that is potentially already available. However, it is not always feasible to make best use of this data, due to inconsistencies in definitions, methods of collection, incompatible geographic resolution, difficulties in access, etc. The problems tend to be greatest when matching data from different agencies, since each such dataset is typically collected to meet the specific needs of that agency and so it may not necessarily adopt the definitions that are most relevant to the needs of freight modelling. Key data resources tend to be ignored for precisely these reasons.

Recent changes due to the modernising government agenda are likely to improve the availability and integration of a number of land use data sources, but it may take some time yet before the full benefits of these activities are widely available. The attitude of many agencies that collect data is now more positive towards making that data available to outside users. An increasing amount of aggregate data is now freely available on the Internet.

The recent availability of the VOA non-domestic property data represents an important source of information that could be used to improve the spatial detail of origins and destinations of freight movements. However, substantial further investigation and development work would first be required to make best use of this data source, particularly if it is to be combined with IDBR data.

There is great variety in the quality and completeness of data, and in the readiness of new initiatives. It is recommended that all such new data, if it is to be used in transport modelling, needs detailed review and understanding to assess its suitability and consider its reliability.

There are many studies of transport (and other topics) that make use of the types of data described in this paper. Among these projects, there is substantial duplication of effort and evidence of differing interpretations of data along with substantial inconsistencies in approach and quality. A case can be argued for a centralised initiative to assist with the harmonisation, availability and appropriate use of important data.

Such an initiative could operate at various levels. Ideally, there would be a maintained repository, either in the public or private sectors (the latter perhaps publicly supported). At a simpler level, a metadata and advisory service could be established. This project has not explored the options in detail, but recommends that it should be considered seriously.

Authors

Hugh Neffendorf, Katalysis hneffendorf@katalysis.com

Ian Williams, ME&P inw@meap.co.uk