

WORKSHOP ON DATA PRESENTATION

Resource Paper

Presenting and Preserving Travel Data

Kay W. Axhausen

*Institut für Straßenbau und Verkehrsplanung
Leopold-Franzens-Universität*

The investment in travel data is substantial in terms of money, time and effort, but very often that investment is not fully utilized or even protected. This workshop will look at how to maximize the use of the data through:

- dissemination via new media (i.e. WWW etc.)
- improved data presentation
- proper archiving and will suggest procedures and guidelines for these activities.

The workshop will also identify future research needs.

INTRODUCTION

The collection of travel-related data is costly, time-consuming and subject to a large number of possible errors and fallacies. While the reduction of these costs and times has been the subject of much research, as demonstrated by this conference, the issues revolving around how to present the results clearly, both in writing and with graphs, and how to protect the investment into the data have attracted essentially no sustained attention in transport planning and engineering. The results of this neglect are predictable, for example: important data sets get lost, as for example the London 1962 transport study data or the famous FHWA Baltimore 1977 dataset¹, but also more recent data sets; the documentation of many datasets is incomplete, especially with respect to their survey methodology; many graphs display the rough-and-ready aesthetics of the engineering drafting shop, which prizes comprehensiveness and precision, but neglects clarity and elegance.

It is idle to speculate, what the exact reasons behind each individual problem are, but three factors should play a role in each:

- The professional background of most transport practitioners and of a majority of the academics in the field is in civil engineering, which implies in general both the lack of awareness of a social scientist for the long-term value of source material and a lack of design training;
- The project culture, which focuses the attention on the requirements of the moment and which discounts the value of the data collected for later projects or problems;

¹. At least, the author has not been able to obtain the data set from a number of colleagues involved in the study, which he has approached over the last two years, but he would be happy to stand corrected.

The “do-it-yourself”/ “reinvent the wheel”-culture of transport planners with respect to data collection and presentation, but also the lack of appropriate and generally distributed tools, which provision would be the task of the relevant specialists in statistics and cartography.

While addressing each of the problems would be a valuable contribution to professional practise under normal circumstances, recent developments in the planning context make their solution urgent, and maybe even a precondition for the collection of data at all. The following keywords describe these developments: tighter budgets, stricter modelling requirements (see for example Garrett and Wachs, 1996 (1)), more intense citizen participation, stronger feelings of public ownership of public data (freedom-of-information or public records acts).

Especially in the US, but slowly also elsewhere, transport models have to fulfil tighter requirements with regards to their realism and validity. While such requirements can be fulfilled to some extent with the traditional data collected for the traditional task of transport modelling, i.e. the evaluation of the capacity expansion of the various networks, they cannot be fulfilled without new data and new models for most new policy measures being discussed, i.e. road tolling, congestion pricing, information provision, etc. The new data requirements increase the complexity of the respondent task substantially with the attendant worries about the resulting data quality and validity. This is especially worrisome, as there are already substantial doubts about the quality of the data collection for the current set of items. They also increase the costs, potentially beyond the currently accepted levels.

Accepting that public budgets are not necessarily set to reflect need, but history, politics and lobbying, transport planners have to expect still tighter public budgets for data collection, especially if there are doubts about the data validity. The logical result will be an increase in the average interval between successive data collection exercises. The recent American experience shows, for example, that local/regional politicians seem happy to accept 10 plus year intervals between substantial travel behavior data collection efforts. The reverse problem exists for traffic control data, which tends to be collected in huge amounts, but not stored or analysed beyond the immediate needs of signal control, variable message signing etc. Combining these first two trends, it becomes clear that the data collected has to be protected over time to keep it available and that it has to be shared between localities to make it affordable.

The protection over time implies both full documentation and adequate physical storage, as multiple generations of planners will have to work with the data. Documentation is especially important, as organisations cannot any more rely on informal knowledge stored in the organisation itself, especially as the increased outsourcing of planning tasks will scatter any such informal knowledge beyond reconstruction. Protection over time implies also continuous updating at various levels of aggregation of the observed behaviors².

The costs involved force organisations to share their data in a properly documented way. While national data, such as the various NTPS's, are an example of this sharing, the new policy initiatives will also require the sharing of more specialist data sets, such as stated-preference work, focus group results, interactive-interview protocols etc, as either the costs

² It is not the task of this workshop to pursue this specific topic, but there are a whole range of issues to be discussed here, e.g.:

- consistency checks between updated origin-destination matrices and the underlying travel patterns
- disaggregate models of activity pattern generation updated on current (traffic) measurement
- aggregate models of activity patterns updated on current (traffic) measurement

will have to be spread among multiple organisations or will only be funded from national/regional sources, if reuse is possible. The increasing intensity of citizen participation and the increasing sophistication of the interested public, i.e. pressure groups and lobbies, requires a much improved access to transport data and to the results of transport modelling. These groups want to be able to extract and tabulate data to suit their needs, which can never be fully satisfied in a single report to the project sponsor.

While printed material can still make a valuable contribution (2), the rapid spread of the World Wide Web has increased the expectations of both citizens and experts enormously. The profession is under pressure to increase the access to the data and to make at least simple “do-it-yourself”-analyses possible. Although this increased visibility enforces increased quality standards, which need to be matched by the practitioners, it also increases the public support for data collection. Data, which can be accessed, is useful and fundable data.

The increasing availability of central government collected data on-line and on the web will force planning agencies to follow. The existing examples (see Figure 1) will only be precursors to much improved and more sophisticated open access systems. The planning agencies should see this as an opportunity to make their concerns public and to find support for their activities, including proper data collection.

In summary, existing trends will force us to collect more and more difficult data, while at the same time funds for this become rarer. To build the support for the funding of data collection, the data has to be made available as widely as possible to both the professional and citizen public, with their differing requirements in type and style of access. To protect the investment into the data and to make it properly accessible, the standards of documentation and of storage will have to be raised. The thus intertwined issues of data preservation, access improvement and data presentation will be the subject of the next three sections. A guidelines section sets out immediate actions, which could be taken, while a final section suggests further work.

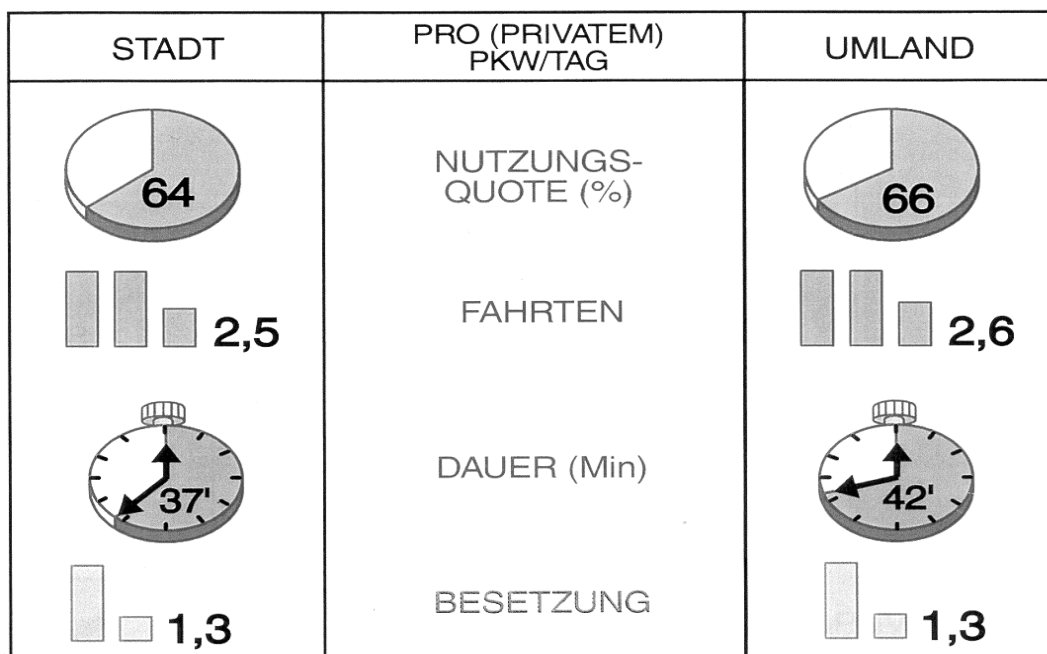


FIGURE 1 Examples of central government collected data.

PRESERVING THE DATA

Current Situation

Travel behavior and traffic data are expensive. Recent estimates of the costs for a completed household travel diary, including travel forms from all household members, range around 50-75 ECU or \$65-100, although of an unspecified quality. Unfortunately, the costs of the collection are only a small part of expense involved. This author would be willing to guess that the costs in terms of staff time for the preparation of the survey (selection of methods, design of survey, search for funding etc.), the supervision of the work and for the checking of results (further plausibility checks at the aggregate and disaggregate level) are one to one and a half times the collection costs, obviously with economies of scale for larger surveys. For large surveys this translates into large amounts of funds, which should require adequate protection. As mentioned above, these measures are rarely properly taken.

Consider for example, that the Land Oberösterreich, a federal state of Austria, is storing its massive travel behavior surveys on a very likely outdated IBM mainframe. Its most recent survey of 1992 had covered a third of the population of the state and achieved a response rate of 70% translating into records for about 350.000 persons. Commercially, this survey would have cost about 5-7 MEcu, more than any other European travel behavior survey known to the author, but for the London Area Travel surveys. At the time of writing only two small consultancies had been able to transfer the data from the IBM mainframe via an outdated cassette format tape, to an accessible platform. One does not wish to think about the effects on the data availability of a phasing out of the IBM mainframe, or the departure of the one staff member dealing with data in each of the consultancies. One should also note, that the analysis of this data set is far from complete. Nothing more than a series of tabulations has been published so far. Full scale "standard" modelling has not yet begun. The loss of this dataset would be a severe blow to the Land and its transport planning for the next ten-twenty years, but also to all other possible users.

Consider as a second example, that the German Ministry of Transport uses a medium- sized consultancy to store the original data files of the various KONTIV studies. The consultancy is highly regarded, but has an awkward legal structure as a voluntary association and an uncertain future should its head retire. In addition, the smooth operation depends to some extent on the good working relationship between the person responsible in the Ministry and the consultancy. One can easily imagine situations, in which this institutionally unstable situation could break down and the general access to the data becomes in turn slow and unreliable, as the current host loses interest. The loss of these data sets in their authoritative form would make the analysis of trends in German travel behavior impossible and would remove the temporal context from ongoing work in Germany.

It is clear, that archiving the data with an institution, which task it is to preserve the data and to provide access to it is the only way to avoid loss of the data and to obtain the maximum value from the data beyond the immediate short term project use. These institutions exist, but hardly any of the researchers and practitioners in transport planning are aware of their existence or of their services and many of those, who are aware of them, refuse to deposit their data. The issue is not even raised in the recent NCHRP Synthesis on household travel surveys (Stopher and Metcalf, 1996 (3)). What are the services and why are they not used?

The archives (see Table 1 for a list of the relevant institutions) are institutions funded by their host universities or relevant research councils to store data in electronic form and to provide access to it. They will check the data set and the documentation of the data for completeness, will publish the data availability in their newsletters and most importantly in their web-based catalogues, will provide the data on request to new users and will maintain the data over time in a secure fashion.

TABLE 1 List of Selected Data Archives

Country	Name

<i>Europe</i> (see http://www.nsd.uib.no/cessda/europe.html)	
A	Wiener Institut für Sozialwissenschaftliche Dokumentation und Methodik
B	Belgian Archive for the Social Sciences
CH	SIDOS
D	Zentralarchiv für Empirische Sozialforschung
DK	Danish Data Archive
F	Banque de Données Socio-Politique
H	TARKI
I	Archivio Dati e Programmi per le Scienze Sociali
NL	Steinmetz Archive
N	Norwegian Social Science Archive
S	Swedish Social Science Archive
UK	ESRC Data Archive at the University of Essex
 <i>North America</i> (see http://www.nsd.uib.no/cessda/namer.html)	
CA	San Diego Social Science Data Center
CA	Social Science Data Archive at UC Irvine
CT	The Roper Center at the University of Connecticut
MA	Harvard Data Center
MI	ICPSR at the University of Michigan
NJ	Princeton University Data Library
NY	Columbia University Electronic Data Services
VI	Social Science Data Center and GIS Laboratory at the University of Virginia
WI	DPLS at the University of Wisconsin - Madison
BC	University of British Columbia Data Centre
ONT	University of Western Ontario Social Science Computing Laboratory

The institution depositing the data can stipulate how the archive has to handle requests for data. The UK Data Archive, for example, has three modes of access:

1. Access at the discretion of the Director of the Archive subject to the undertakings discussed below

2. As A, but with information of the data depositor
3. As A, but subject to prior permission of the data depositor

The data depositor can therefore retain control over the access to the data.

The control over the data is further extended in the case of the UK through the undertakings mentioned above, which the institution applying for the data has to sign. They cover i.e:

- Restriction of the analysis to academic research and teaching
- Prohibition against further distribution of the data and the associated materials
- Requirement of proper acknowledgements of the data in reports
- Requirement to deposit copies of any reports or papers based on the data
- Potential imposition of a requirement, that any publication needs to be approved by the original data depositor
- Requirement to deposit any data set derived from the data set used.

It is obvious for the case of the UK, that the depositor can retain near full control over the data and the results derived, while benefitting from the services of the archive in terms of publication of the data and physical care for the data and materials.

Returning than to the question of why these services are not used, a number of reasons come to mind:

- The existing archives have served and addressed social scientists in the past and have not encouraged “fringe” disciplines, such as transport, to deposit their data.
- Local and regional authorities are reluctant to publish data, which could become politically sensitive.
- Professionals might have doubts about the quality of their data.
- Professionals deem the data to be “too complex” to be made accessible to the public or other scientists.
- Too many additional data sets would have to be provided.
- Firms are reluctant to ask and authorities to pay for the effort required in the preparation of the data for archiving.
- Firms and researchers do not wish to make the data public, so as to build up data monopolies and to reduce scrutiny of their work.

It is impossible for the author to judge how far the first reason describes the realities in the various developed countries with data archives, but checking the electronic catalogues of three archives (Data Archive at the University of Essex, ICPSR at the University of Michigan and the Zentralarchiv at the Universität zu Köln) reveals hardly any transport related data sets other than the national passenger transport surveys in the US and the UK, while even those are missing in Germany. Certainly, the archives have not searched out transport related data sets in the past. One can only hope, that this will change.

In the past, authorities have been very reluctant to make any data accessible. This old presumption of secrecy and information control is certainly unacceptable today, where the publication of the data should be seen as an essential part of the citizens’ participation process, but also of better communication between the various branches of government. The recent wide spread development of administration-wide geographic information systems is a

positive sign and the integration of transport data into these systems should be an obvious task for the future.

Transport data, especially from travel behavior surveys are complex, but that should not be a reason to constrain access, but to make better tools available for the analysis of such data. Such tools would be a benefit for the public, but again also for other non-transport professionals, which want quick answers to small queries without large delays. See next section for further discussion.

One should also not forget, that many available surveys are equally complex, such as the various family expenditure surveys or census samples (e.g. the US PUMS), and that the complexity of nearly all transport data sets is trivial in comparison to the complexity of the panel data sets available from the archives, such as for example the US Panel of Social and Income Dynamics (PSID) or the UK British Household Panel Study (BHPS).

While errors are likely to be made by non-expert analysts, the benefits of access: additional secondary analysis, application of new statistical methodologies unavailable at the time of first analysis, and the improved scrutiny applied to the data before and potentially after publication should easily outweigh the problems possible created by those errors; errors, which should be easily identifiable by the experts. The exclusion of non-academic users from data access, as for example mentioned above in the case of the UK Data Archive, should reduce any such fears further.

Most social-science data sets can stand on their own, i.e. depositing them requires the data, a codebook and description of the methodology, including the weighting procedures. While transport data, travel behavior data in particular, can also be deposited in this form, they require in most cases additional material to be of full use in later analyses, in particular (electronic) maps of zones, locations and networks, logical descriptions of transport networks and services, descriptions of transport-related fares and fees, and aggregate count data for validation. The first group of items could be replaced by matrices of travel times, distances and costs for each mode and origin-destination pair. These additional items should not stop the process of depositing of as much of the data as possible, but they raise additional problems for which solutions need to be discussed.

The preparation of the data for the archive requires staff time and therefore contributes to the costs of the project. In a climate, where the data is perceived to have no value beyond the project, it is unlikely that those costs will be underwritten by the client. While it would be a change for the better, if transport planners would recognize the value of their data for current users from other areas and for future users, it is more effective to conceptualize the preparation of the data for the archive as a small marginal cost on the back of the essential data quality assurance procedures, which should be part of any serious data collection exercise, e.g. preparation of code books, documentation of the sampling procedures, documentation of the manipulation of the data (imputation, recoding, exclusion of outliers, calculation of derived variables etc.), documentation of weighting methods and of the data sets used to derive weights. Clients should in the future include the requirement to deposit the data with an archive as part of their quality assurance and firms should include it as a sign of their commitment to data quality and professional standards.

In the past, many clients, in particular in the public sector, did not have the resources to work with data themselves. It was therefore natural for them not to worry about the storage of the data and to be content to see the data stored with their consultants. In addition, as mentioned above most transport planners in government were not aware of data archives or not interested in using them. The consultants in turn were happy to accumulate data sets

and to obtain a monopoly over the data access with the associated follow-on work based on it. While this prospect might have enticed consultants to underprice the survey work, it is important to note, that the total cost for survey will have been borne in the end by the client, even if in stages through (overpriced) follow-up work.

This happy equilibrium is breaking down, as citizens demand better access to the data collected with their funds and as government authorities acquire more of their own data analysis capabilities. In this situation both sides should have an interest to employ data archives for the storage of the data, as the consultants cannot have an interest in providing the data storage for free, nor can the authorities have an interest in developing many archives of non-effective size of their own.

The improved access should also help to improve the quality of the data, as both sides, client and consultant, can be challenged by the public: the client for skimping on quality and the consultant for not providing it. Clearly, this is inconvenient for both sides, but it is in the public interest, if the data quality on which public decisions are based is to be improved and if both authorities and consultants give more thought to the amount and quality of data they need.

This author can see only two exceptions to this: first, the case of (multi-client) studies undertaken at the risk of a consultant or market research firm. Here the data are indeed the capital of the firm and it would be unreasonable to expect the firm to make the data available for free. The number of transport examples known to the author is small: the European Travel Monitor, Infratest's Mobility, the Victoria Activity and Travel Survey and the Web Travel Report/Survey³. Second, surveys undertaken by transport providers with regards to their unsubsidized services.

Supplementary Spatial Data

While the ASCII-format provides a simple, but universal standard for the storage of travel behavior data, no such standard exists for the storage of spatial data and networks. While ARCInfo-formats cover some of the area, they do not cover the important area of the logical network descriptions, as required for assignment or the description of timetables. Here a multitude of competing software formats exist. In addition to the many data format, a multitude of approaches exist to describe logical transport (service) networks and their attributes. Given the proprietary nature of the associated software, which is often changing quite quickly in response to market needs, current network files might not be extremely useful to later users without access to this software.

In the short run, as mentioned above, data archives should require depositors to make ASCII-matrices of the most important variables available by each origin-destination pair (costs, distances, travel times and their elements (access, waiting, driving/riding, egress)). Boundary files should also be made available in a transparent ASCII-format, which allows transformation to the available proprietary formats.

In the longer term two avenues are open:

1. The data archives could develop standards for network and map descriptions building on the on-going standards work for electronic maps (CEN and similar efforts in the

³ A March 1997 search with *AltaVista* has identified a range of firms collecting travel related data through the Internet for later sale to advertisers and firms, but given the demographics of the web none of them should be able to support public decision making at the moment.

US). This would require specialized efforts in areas outside the normal interests of the archives. It would also require efforts without having the necessary mandate to undertake them.

2. The data archives could also wait for the standardisation/oligopoly formation in the various software categories (GIS, assignment, public transport planning etc.) to end and to focus on tools for translation between the programs within these categories. While the process of oligopoly formation has stabilized in the popular software categories, such as word processors and spreadsheets, it has not even begun in most professional software categories. While ESRI's ArcInfo has a very strong position in geographic information systems, there are a number of competitors and a complete reversal is still conceivable, although becoming less likely as time goes by and firms and public administrations invest in their GIS-data bases. The problem for transport is, that ArcInfo does not provide standard ways of describing logical networks and transport services. This is the domain of specialized software, which might offer import facilities from ArcInfo or other GIS-systems, but rarely export facilities. The market for the most important types of specialized transport planning software (demand estimation; road and public transport assignment; cost-benefit analysis; timetabling and public transport run- and crew cutting; OD-estimation; off-line signal optimisation and real-time network control) are still national and in flux. National, as the software is mostly sold together with consultancy services, favouring local providers; national often also because the software has to reflect local practises and regulations; in flux, as the requirements are changing rapidly; e.g. the need for dynamic assignment opens up the market for assignment software; in flux also as the computing environment change; e.g. the change from DOS to Windows.

In this context, only a specialized archive, which has the facilities to maintain a range of these software products, can be of use. The other archives will have to wait until the GIS-market stabilizes around the likely three to five providers, which in turn either offer the specialized transport software themselves, or are adopted as base platforms by the specialized software developers. An analogous example is the market in tie-ins for AutoCAD.

Way Forward

The discussion has shown that the publication of transport data, just data from the social sciences, with professional data archives should improve professional practice and data quality. It would be in the public interest to store the data in a central facility, an archive, which is charged with documenting and maintaining the data, instead of dispersing the data sets between a large number of institutions, for which this is not the primary focus of their work.

Transport data ideally require a substantial supplement of spatial data to be of optimal later use. In short term, the archives should require substitute aggregate information in the form of matrices of travel costs, distances and times etc. by origin-destination pair, but in the medium term specialized spatial data archives should be built up, which have the capabilities of dealing with this specialized data. It seems unlikely that more than one or two such specialist archives each will be required or sustainable in Europe and North America. The ideal location would be at an existing archive, which can draw on prior local expertise in transport, planning and geography. Direct attachment to a government body, such as the

Bureau of Transportation Statistics, would create difficulties in terms of long range development, possible fears of centralized control or political influence.

IMPROVING ACCESS

Current Situation

Transport data is too expensive, too valuable and often too contentious to be analyzed only once or only by one group of researchers or analysts. It is also too multifaceted that a single report could shed light on all of them. It is also complex and reasonably full of traps for the unwary or first-time analyst. While it might have been acceptable in the past to conclude, that the best course of action was to keep the data relatively inaccessible and to publish standard sets of tables and one-off reports, today's technological possibilities and public and professional wishes make this no longer acceptable.

While the publication of the data with the data archives is an important element in this changed treatment of data, an even more important element is the provision of ready access to the prior results and to one-off-tabulations and analyses.

It is not necessary to discuss in detail the system of technological tools, which enable us to contemplate this switch, but the most important ones should be named: the proliferation of cheap and extremely powerful personal computers (portable and desktop) at the workplace and at home, the dramatic fall in communication costs in most parts of the world, the wild-fire-spread of the World-Wide-Web and its enabling technologies, the development of powerful front-end-machines for data bases and statistical tools (data warehousing) and the spread of GIS-systems into administrative practice with the attendant market for spatial data. There are even the first tools, which make data ready for web publication (see for example <http://bravo.berkeley.edu/csadocs>)

A surprising number of data providers has already responded to the availability of these technologies by developing interfaces to their data for their paying or non-paying customers:

1. The U.S. federal government has developed or is developing systems to give access, for example to its
 - Highway statistics (www.bts.gov/ntda/fhwa)
 - Fatal accident statistics (www.bts.gov/ntda/farsdb)
 - Airline on-time-arrivals records (www.bts.gov/ntda/oai)
 - 1995 NPTS ([www-cta/ornl.gov/npts](http://www-cta.ornl.gov/npts))
2. The San Francisco Bay area MPO provides access to tabulations, reports and data through its data mart (www.mtc.dst.ca.us/planning/data1.htm)
3. Team at the Wirtschaftsuniversität Wien provides access to a range of tourism surveys and statistics, including the European Travel Monitor (tourMIS.wu-wien.ac.at) (Wöber, forthcoming (4))
 - the origin-destination survey of the Montreal region (Chapleau, Trepanier, Lavigne and Allard, 1996 (5)),
 - and many offer real-time travel time and related information (e.g. Quiroga, Bullock and Schwehm, 1997 (6))

The Social Science Data Center at the University of Virginia provides access to interactive data resources (www.lib.virginia.edu/socsci/interactives.html), from among others

- County and city data books
- County business patterns
- Public use microdata samples (1990 Census)
- Regional economic information system
- World tables of economic and social indicators (1950-1992)
- Census data extraction system
- Panel study of income dynamics

It is clear that such improved access requires well-defined minimum standards in terms of data analysis and presentation to minimize the spread of erroneously tabulated data and of ill-founded analyses. The benefit of improved decision making by giving more people access to the results they want at the time they want it in the form they need it, should be on balance greater than the remaining errors.

This minimization of errors requires the specialists to codify their knowledge through these systems. While this reduces their income from standard work, it also relieves them from it and allows them to focus on more specialized, difficult and interesting work.

At the moment, the existing systems tested offer only tabulation and some rudimentary graphing. Only one system tested was planning to offer access to regression, but that tool was still under construction.

As a first draft the following requirements could be formulated for tabulation and graphing:

1. All results should be presented with proper weighting; tables for which proper weights cannot be calculated should be suppressed or marked with full “health warnings” about possible bias and errors.
2. All tables are printed with all relevant information:
3. Full title for the table:
 - a. Full descriptive labels for the variables
 - b. Full descriptive labels for the categories or classes
 - c. Dimensions of the variables, where relevant
 - d. Full description of the reference population:
 - Reference period
 - Sampling/reference area
 - Description of any subsetting/selection done from the full data set
 - e. Source of the data
 - f. Date of the table production, name of the tool used (maybe as copyright indication) and any formal disclaimers required.
4. Further details should be available via hyperlinks suitably indicated on the display and printout:
 - a. Explanation of the variables and of their categories, if not self-evident
 - b. Reference to any imputation process applied to the variables
 - c. Reference to any data cleaning applied to the raw data
 - d. Explanation of the weighting process, including a positive indication, if the data has not been weighted. The table should also indicate, if only a partial

weighting has been applied and the known relevant factors, for which no weighting has been applied, should be noted.

5. Tables calculated from sample surveys should suppress statistics in cells based on insufficient number of cases (e.g. 30), when statistics, such as means, medians, standard errors are requested.

6. Tables calculated from sample surveys should, where appropriate automatically provide both point and interval estimates, e.g. means, standard errors and number of cases in the cell.

7. Tables derived from time series should be split into multiple tables, if substantial changes in the definitions or the data collection method happened during the requested time period. If corrected data are available for the whole time period, the type of correction applied should be described instead.

Comparable requirements could be formulated for the production of graphs.

The requirement for proper weighting is particularly onerous for sample surveys of travel behavior, where a full weighting has to consider:

- The socio-demographic sampling error.
- The seasonal/temporal error, if relevant.
- The various non-response mechanisms (self-selection, memory, social desirability etc.) at the various levels of analysis (household, person, daily activity programme, journey, trip or stage).
- Weights to reflect non-sampled populations, where relevant, as for example in the production of origin-destination matrices.

The standards for more complex types of analysis would have to reflect that increased complexity by guiding the user tightly to avoid the production of erroneous types of analysis. The open access to these types of analysis will have to be delayed until proper statistical expert systems are available to guide the user to a type of analysis appropriate for the data and the relationships under consideration.

It is impossible to formulate in general what capabilities such an open access tool should provide, but for the small number implied in simple graph and table production:

- Selection of data sets
- Selection of variables
- Selection of formatting/classification of the variables
- Selection of statistics to be produced (means, sums etc)
- Subsetting of the data with filter criteria

None of the systems tested by the author fulfils all the requirements listed above or offers all the capabilities mentioned. The in-house system for the Swedish National Travel Survey comes closest.

Way Forward

Transport planning can increase the value of its data by making it accessible to as wide a public as possible. Traditional planning data should be presented alongside the data collated for the available traveller information systems. The infrastructure should be used jointly.

The transport planning community should support the development of such open-access systems through the publication of the data and the development of comprehensive weighting and correction systems.

PRESENTATION OF THE DATA

Current Situation

The publication of the data in archives and the open access to the data via the web are two important contributions to justifying the expense of transport data, but an improvement in the way, in which the data are presented graphically, could make a further contribution. “A figure saves a thousand words” is an old engineering adage. Unfortunately, this only applies to well constructed and clear figures. The proliferation of desktop computing has in the main reduced the clarity and the quality of the construction of graphs by providing powerful tools, which make the construction of bad graphs easy and of good graphs difficult.

There is a substantial scholarly literature on how to construct good graphs, including guidelines from the various professional statistical societies, but this has had little impact on the practise in transport and even less so on producers of spreadsheet and presentation graphics programs and not even on many producers of statistical software.

It is impossible to discuss all of the forms of graphs used in detail or to discuss the errors made in detail, but one can state the following recommendations (see also Tufte, 1983 (7) or 1991 (8); Cleveland, 1994 (9); Schmid, 1983 (10)):

- Match graph, purpose and audience
- Maximise the “data ink”, but don’t overwhelm the reader
- Don’t mislead the reader
- If there are less than five (seven, ...) data items displayed, use a table instead

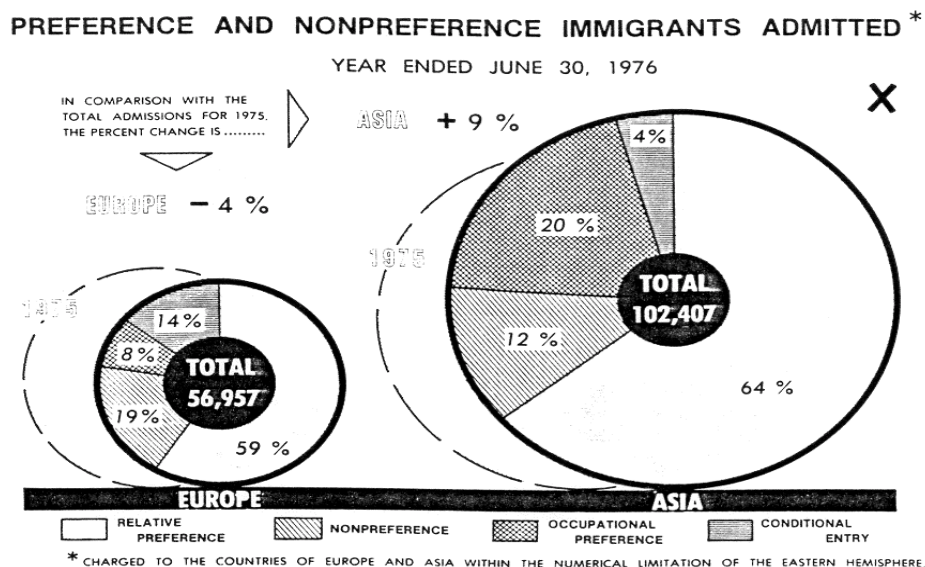
A graph in a report or presentation should not be a “Selbstzweck”, “l’art pour l’art”, but should contribute to argument made or record the data/provide a tool for later work. The style of the graph varies with each of these purposes. If the graph is there to support an argument or observation made, then it should focus on this argument or observation, for example a particular pattern or a striking temporal development. This implies that unnecessary information is suppressed by omission, deletion, smoothing or merging of categories. The other two types of graphs in contrast require that the graph records its information with maximum accuracy and that it provides the reader with the possibility to read the data with the intended accuracy from the graph. It should be said though, the graphs are not particularly suited for either of these two last purposes. Exact values are easier stored and retrieved from tables and nomograms have/should have long since been replaced by appropriate computer programs.

It is clear, that the style of the “argumentative” graph will vary with the audience for which it is intended. Graphs for the general public will be less complex in their construction and will have a lower threshold in terms of the minimum number of data points required. Graphs for professional audiences will be more complex and require more data points to be justified.

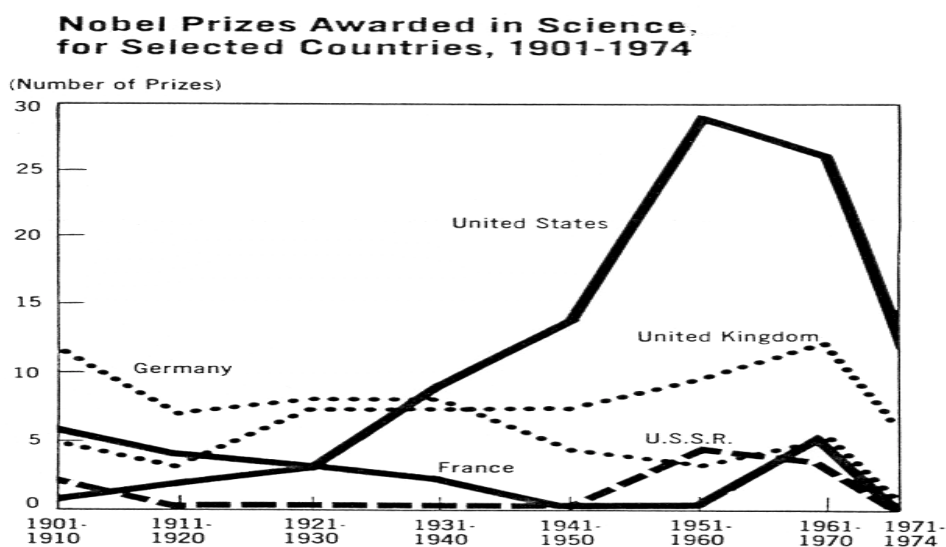
A graph should not contain elements, which do not add to the information provided, i.e. the “ink” used should be dedicated to conveying information and not to extraneous purposes. Typical examples of such waste are shadows, three-dimensional bars in one-

dimensional histograms, unnecessarily complex shadings and hatchings, superfluous reference lines etc. While the information content of the elements on the graph should be maximised, the graph itself should not convey more information than the reader can absorb. The graph should just tell the story, which the argument requires, but not more.

Designers can mislead the reader by distorting the elements of the graph. Examples are axes, which vary their scaling in unexpected ways, say from yearly to quarterly intervals, or axes, which do not begin at the natural reference point for the scale used. Clearly, this should be avoided (see figures 2 and 3).

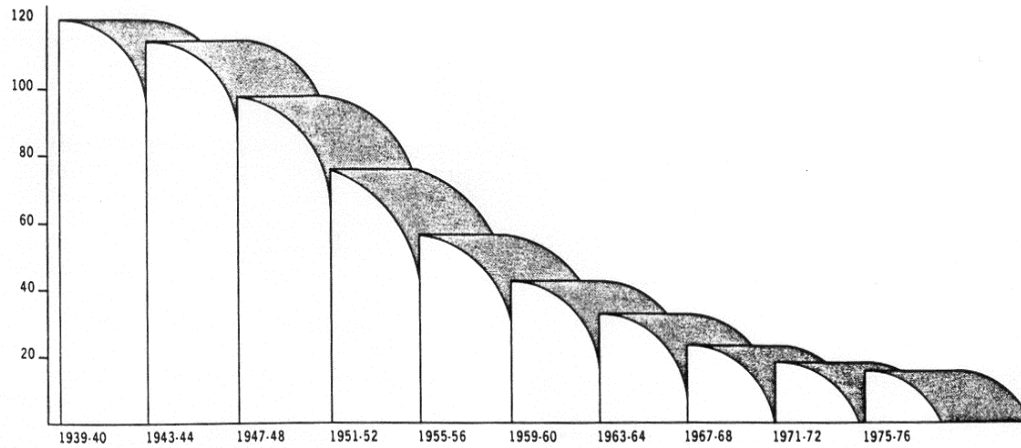


Note the wrong scaling of the pie (diameter instead of area) Source: Schmid (1983), 67.



Note the wrong scaling of the x-axis. Source: Tufte (1983), 60.

FIGURE 2 Examples of badly constructed graphs.



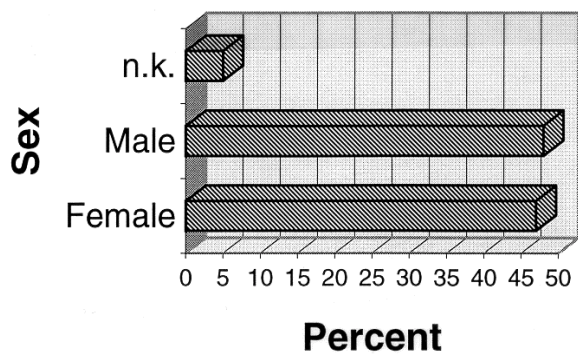
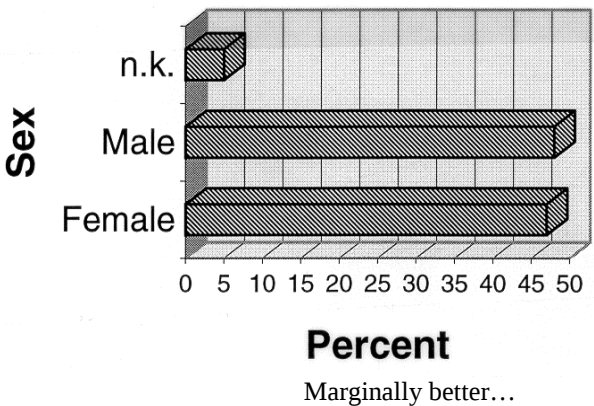
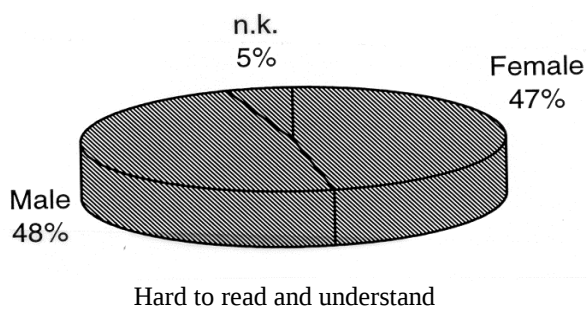
Note the confusing form of the bar and the superfluous 3D effect. Source: Tuft (1983), 97.



Moire effects (hatchings etc.) at their worst. Source: Schmid (1983), 129.

FIGURE 3 Further examples of badly constructed graphs.

The availability of easy computer graphing invites the production of graphs, which contain essentially no information in relation to the space they occupy: shares of males and females in a sample, for example shown with a 3-D two-colour pie chart with shadows (See figure 4). Although the exact limit is open to discussion, one should not produce such content-poor graphs, but use tables instead.



Share of respondent [%]

Female	45
Male	48
n.k.	5

a simple table would do

FIGURE 4 Example for the principle of “Maximised data ink”

While maps have a different function, they are used as the basis of many statistical graphs in transport; for example to show flows between zones, flows on links or densities. The same set of rules applies.

The improvement of the known types of graphs and maps is one task, the second is the design of new types of graphs and maps. Here the computer offers new possibilities to visualize the dynamics of travel and traffic over space, which are essentially impossible to capture on paper. Possible examples are:

- 3D-visualisations of the time-space paths of the respondents in a diary survey
- Dynamic display of network loads on a map
- Dynamic display of network speeds on a map
- Dynamic display of occupancy of a zone or facility
- Dynamic analysis displaying the origins and destinations of vehicles or persons using a particular link, service (train or bus) or zone (current values or time aggregates)
- 3D-visualisations of spatial correlations or interactions (over time)
- Hyperlinking databases and analyses

Way Forward

The proper presentation of the data collected using graphs is essential for the communication with the public. The availability of easy computer graphing has led to too many ill-constructed graphs. The level of skill in the area can only be improved by better training of the practitioners and researchers. The training should focus on the awareness of the factors guiding the design of graphs: the audience (general public, professionals or experts in the field), the purpose (support of an argument or storage/retrieval) and the maximisation and *data ink*.

Future web-based systems should emphasise the importance of this aspect by designing their graphing systems carefully. Equally, the review of reports and scientific papers could be tightened to improve this aspect of our work

RECOMMENDATIONS FOR THE GUIDELINES

The Guidelines to be prepared as a result of this conference should include the following recommendations based on the discussions above:

- *Publication of data:* All data sets, which involved more than six person months of effort, should be published with a recognised data archive within a year of the end of field work. The archive should accommodate partial deposits and successive versions of the data.
- The reporting requirements of this publication should form part of the general quality requirements of the data collection effort.
- The wish to publish the data should be indicated in either the call for tender or in the tender. Any limitation on data publication imposed by a private client should have a deadline of no more than five years. No limitations should apply to data fully funded publicly.
- All relevant supporting data sets, e.g. maps, networks and support information data sets, such as for weighting or travel costs, should be deposited as well in either the original form or as matrices by origin-destination pair.

- The relevant data protection, public record and freedom-of-information regulations should be observed.
- *Data access*: data sets, which form the basis of national or regional policy, should be made accessible through the WWW within six months of publication with a data archive. The services should initially be restricted to tabulations.
- Any static or dynamic exhibit (tabulation, graph, map etc.) made should be as complete as possible, in particular inform the reader of the relevant limitations which apply.

The presentation of data cannot be made the subject of these guidelines, but see Chapman (1986) (11) for an attempt to introduce good practice in the British civil service with respect to tables or Tufte (1983) (7) for graphs.

RECOMMENDATIONS FOR FURTHER WORK

The recommendations made for the guidelines also suggest areas of further work. If the transport planning profession starts to publish its data consistently with data archives, these will have to develop additional skills to deal with this type of data. In particular, the map and GIS-data sets will need to be maintained in a more active form than the straight data elements, as these data sets need not only to be maintained physically, but also logically, i.e. kept up to date with the moving standards for electronic mapping and GIS-files more generally.

It might be necessary to set up one or two specialist data archives for transport data sets, which would be charged both with the archiving tasks and the maintenance of the data. They might also be charged with the development of suitable data description standards.

The profession should raise the awareness of the practitioners for the benefits to be derived from data archives and their active use:

- Improved data quality to match the archiving standards.
- Reduced costs through concentration of data archiving with specialists.
- Additional free secondary analysis through researchers and other interested parties.
- Larger data base for decision making through data sharing between localities.
- Improved understanding of travel behavior through longer time frames and wider spatial reference.

The profession should also accept the publication of a data set with an archive, as a proper and fully valid academic publication.

The development of the web-based access to data is very rapid, with the U.S. Bureau of Transportation Statistics in the vanguard of the development. One can therefore assume, that the transportation planning profession does not need to take a special initiative with regards to the web-technologies required. It does need, though, to take a special initiative with regards to the proper weighting methods and to make sure that these are applied in web-based access systems, otherwise the proliferation of erroneous data is assured.

ACKNOWLEDGMENTS

The author is grateful for comments of Chuck Purvis and Phillipe Toint and of the anonymous referees. The author would also like to acknowledge the work and dedication of

the organizers of the conference, among many others in particular of Werner Brög and his team, who generated the opportunity to present these ideas. Any errors are my own.

The paper is partly based on work undertaken as part of the EU-funded 4th Framework Transport Research programme project “Technologies for European Surveys of Travel Behavior” involving the Institut für Straßenbau und Verkehrsplanung, Leopold-Franzens-Universität Innsbruck, Statistics Netherlands, Büro Herry (Wien), University of London Centre for Transport Studies, Imperial College (London), Deutsche Versuchsanstalt für Luft- und Raumfahrt (Köln), INRETS (Arceuil), Transportes, Inovação e Sistemas (Lisbon), TØI (Oslo), Transport Studies Group Facultés Universitaires Notre Dame de la Paix (Namur), Socialdata (München), Statistics Sweden, TU Delft. All conclusions drawn are solely those of the author(s) and not of the CEC or the Consortium.

REFERENCES

1. Garrett, M. and M. Wachs *Transportation Planning on Trial*. Sage, London, 1996.
2. Socialdata. Stadt statt Verkehr. *Mobilität in Innsbruck: Verhalten*, 1, Magistrat der Stadt Innsbruck, Innsbruck, 1994.
3. Stopher, P.R. and H.M.A. Metcalf. Methods for Household Travel Surveys. *Synthesis of Highway Practice*, 236, NCHRP, TRB, Washington, D.C., 1996.
4. Wöber, K.W. Marketing Information and Decision Support on the Internet: New Opportunities for National, Regional and City Tourist Offices. In A.M. Tjoa (ed.) *Information and Communication Technologies in Tourism 1997*, Springer, Wien, forthcoming.
5. Chapleau, R., M. Trepanier, P. Lavigne and B. Allard. Origin-destination Survey Data Dissemination in Metropolitan Context: A Multimedia Experience, *Transportation Research Record 1551*, TRB, National Research Council, Washington, D.C., 1996, pp. 26-35.
6. Quiroga, C.A., D. Bullock and C. Schwehm. Dissemination of Travel Time Information Using the World-Wide-Web. Paper presented at the 76th Transportation Research Board Meeting, Washington, D.C., January 1997.
7. Tufte, E.R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, 1983.
8. Tufte, E.R. *Envisioning Information*. Graphics Press, Cheshire, 1991.
9. Cleveland, W.S. *The Elements of Graphing Data*. AT&T Bell Laboratories, Murray Hill, 1994.
10. Schmid, C.E. *Statistical Graphics: Design Principles and Practices*. Wiley-Interscience, New York, 1983.
11. Chapman, M. *Plain Figures*. HMSO, London, 1986.

WORKSHOP SUMMARY

Chuck Purvis

Metropolitan Transportation Commission

WORKSHOP PARTICIPANTS

Kay Axhausen (resource paper author)
V. Thamizh Arasan
Wolfgang Arnold
Milena Bodmer
Robert Chapleau
Marie-Odile Gascon
Ursula Huettnner
Denise Lievesley (resource paper author)
Karen Lucas

Kerrick Macafee
Csaba Orosz
Charles L. Purvis (chair)
Johannes Reichel
Gerd Sammer
Peter Schlosser
Thomas Spiegel
Phillipe Toint (discussant)

OBJECTIVES OF THE WORKSHOP

The charge of the *workshop on data preservation and data presentation* was to provide baseline and highest standards and guidelines related to “tidying up” at the end of the transport survey process. The workshop focused on three major themes:

1. Data archiving;
2. Access to data tabulations and data dissemination;
3. Presentation of the data (reports, oral presentations, projections of the data).

Two resource papers were used as background material by the workshop participants: “Presenting and Preserving Travel Data,” by Kay Axhausen, and “The Value of Sharing and Preserving Electronic Data,” by Denise Lievesley (paper not published in this volume). Workshop participants Robert Chapleau and Peter Schlosser also gave presentations on transport survey results and the use of multimedia in presenting survey data.

Some of the challenges in the preservation and presentation of transport survey data are related to:

1. *Changing personalities.* It was stated that “expertise is transient,” and that we do not archive people (knowledge) for later access.
2. *Changing hardware.* Computer media change, they deteriorate and they improve. A challenge is to maintain transport survey data due to changes in technology (e.g., the move away from mainframes to microcomputers.)
3. *Changing software.* The ease in creating good as well as inferior tables and graphics is a challenge. Better training of existing and potential data users in software, statistical analysis and graphics presentation is critical.

RECOMMENDATIONS FOR BASELINE GUIDELINES AND STANDARDS, AND HIGHEST STANDARDS

The workshop discussed both baseline guidelines and standards, as well as the highest standards to aspire to, in terms of data preservation and data presentation.

Data Archiving

With respect to the first theme of the workshop, data archiving, the workshop recommended the following:

1. *Increase awareness of the archiving issue (outreach).* To accomplish this end, workshop participants suggested that professional archivists be invited to talk with transport professionals regarding the importance of proper methods for data archiving. In addition, archivists and transport professionals should write articles for various publications such as the Transportation Research Board's *TR News* and the European Council of Ministers of Transport's *ECMT Round Table*.

2. *Begin the survey process with the aim of archiving the data sets in mind.* Transport professionals should be specific in their call for tender (request for proposals) that the end product be full and that complete data file documentation and methodology documentation (of the full process, including sampling, survey execution, imputation, and weighting) be provided.

3. *Data protection regulations should be observed.* This is specific to each country's privacy, confidentiality, or disclosure-avoidance laws and procedures. Also at issue here is at what geographic scale is it appropriate to release microdata? At the latitude/longitude level? Or at the travel analysis zone level?

Other ideas regarding baseline recommendations for data archiving included peer validation of data documentation, providing checklists for developing transport survey data documentation, and "following the instructions" provided by professional archivists. Other issues are related to "version control" (what version of data is available and being used?); the credibility of small, unaudited data archives; the need for filtering poor or small surveys from a data archive; and "meta-data" development (data about data).

In terms of highest standards for data archiving, the workshop recommended that:

1. All data sets (data files and documentation) from major survey efforts (major survey meaning greater than 6 person-months of effort) should be published (or offered) to a recognized data archive within a year of the end of fieldwork.

2. All relevant supporting data sets [maps, GIS layers, network data (transit and highway travel times and costs), socioeconomic data files] should also be deposited in the recognized data archive.

3. The process should allow for staged deposits of auxiliary data, given limited time and financial resources.

Access to Data Tabulations

In discussing baseline standards for access to data tabulations, the workshop participants agreed that any data tables should be printed with all relevant information, including full

title, full descriptive labels, description of the reference universe, data source, and explanation of variables. As appropriate, include the number of sample observations and standard error for cells within a cross-tabulation matrix. Essential, supporting background material such as imputation or data cleaning processes, as well as the weighting process, is needed to keep associated with the data tabulations (e.g., reference citations to supporting background documentation.)

In discussing highest standards for access to data tabulations, the workshop participants were interested in the proposal that interactive data tables should be made available within a data archive through the World Wide Web (WWW), within six months of publication. (The baseline standards regarding posting number of observations and standard errors are critical for Internet-based tables.)

Data Presentation

Workshop recommendations for baseline standards and guidelines for data presentation included reference to basic handbooks and textbooks on graphic presentations, including the three books by Tufte (1-3), Gowers (4), and Chapman (5). These books were highly recommended reading. In addition, the Transportation Research Board's visual aids guidelines are also quite helpful when presenting survey data in oral presentations.

The workshop made no recommendations on highest standards for data presentation. It was felt that data presentation must account for style and taste variations within the context of the country and cultures that make up the transport survey community.

OTHER SUBJECTS RELATED TO DATA PRESERVATION AND DATA PRESENTATION

Participants discussed many other issues in detail during the workshop, but with no specific recommendations:

1. *Cost Recovery.* The needs for the data provider to recoup some of the costs of data collection and analysis were an issue. Various cost recovery models exist, depending on the culture of the data-owning institution.
2. *Freedom of Information/Public Records.* Related to cost recovery are the basic legal issues and rights to access data collected using public or quasi-public funds.
3. *Misinterpretation and Misuse of Data.* There were concerns expressed about these issues, and how a data owner/provider can use disclaimers or data-use agreements, as well as relying on full documentation to inform the data user on the use (and possible misuses) of the data. Misuse is the responsibility of the misuser. The job of the data provider is to provide the necessary information on the context and statistics of the database.
4. *Multimedia.* Workshop participants discussed the future of interactive maps and graphics to stimulate creative use of transport survey data.
5. *International Comparisons.* These types of comparisons will be useful, but are challenged by inconsistencies in variable definitions. It was suggested that EuroStat may play a role in harmonizing this process, at least in terms of European transport surveys.

REFERENCES

1. Tufte, Edward R. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, USA, 1983.
2. Tufte, Edward R. *Envisioning Information*. Graphics Press, Cheshire, Connecticut, USA, 1990.
3. Tufte, Edward R. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, Connecticut, USA, 1997.
4. Gowers, Sir Ernest. *The Complete Plain Words*, 3rd Edition. Penguin Books, London, England, UK, 1986.
5. Chapman, Myra. *Plain Figures*. HMSO, London, England, UK (out-of-print).