PROCESSING, ANALYSIS AND ARCHIVING
OF TRAVEL SURVEY DATA

Resource Paper for Workshop A5

7th International Conference on
Travel Survey Methods
Costa Rica, August 2004

23/06/2004

Gerd Sammer
Institute for Transport Studies
Department for Spatial, Landscape and Infrastructure Sciences
University of Natural Resources and for Applied Life Sciences
Peter Jordan Strasse 82, A-1190 Vienna, Austria
Tel.: +43-1-476545301 Fax: +43-1-476545344
e-mail: gerd.sammer@boku.ac.at
Abstract:

Processing, analysis and archiving of travel survey data seems to be rather routine compared with other aspects of travel surveys. However, this part of the process is very important and influences the quality of the data obtained, their continued accessibility and use. The goal of this resource paper is to analyse and identify those activities and procedures at the state-of-the-art level which are essential to ensure the quality of data processing and data analysis. The paper is based on an international literature review of recent publications. It aims to propose quality standards in terms of guidelines for procedures and content, as well as to discuss how to enable that the execution, so that the procedures of the survey, the results and the original data are adequately documented in a proper way.

A proposal for a harmonized and standardised terminology of the steps of a travel survey is provided, serving as a basis for appropriate data documentation and archiving. The section data processing deals with database building, questionnaire editing, data entry, standardisation for flexible coding, geocoding, data editing and cleaning as well as data weighting. Another part of the paper discusses the quality management of travel surveys during data analysis by validation procedures. A specific focus is given to the consideration of standards for archiving in order to ensure maximised use of the data today and in the future. Also the question is addressed as to how to proceed with the implementation of this task successfully with regard to the institutional and organisational framework in the international scientific community.

Content

1. Introduction ........................................................................................................................ 2
2. Data Processing .................................................................................................................. 2
   2.1. Database Building .......................................................................................................... 2
   2.2. Questionnaire Editing ..................................................................................................... 2
   2.3. Coding and Data Entry ................................................................................................... 2
   2.4. Data Editing / Cleaning .................................................................................................. 2
   2.5. Data Weighting .............................................................................................................. 2
3. Data Analysis ..................................................................................................................... 2
   3.1. Validation analysis ......................................................................................................... 2
   3.2. Data presentation ............................................................................................................ 2
4. Documentation and Preservation ....................................................................................... 2
   4.1. Status-quo Analysis ........................................................................................................ 2
   4.2. Objectives and Requirements on Documentation and Preservation .............................. 2
   4.3. Technology of Metadata and Data Archives on the Market .......................................... 2
   4.4. Future Needs in Data Archiving and Documentation .................................................... 2
5. Conclusions ........................................................................................................................ 2
6. References .......................................................................................................................... 2
7. Abbreviations ..................................................................................................................... 2
8. Annex .................................................................................................................................. 2
1. Introduction

Processing, analysis and archiving of travel survey data seems to be rather routine or boring activities than other aspects of travel surveys such as the design of the survey, sampling or the development of the survey instrument. However, this part of the process is also very important and influences the quality of the data obtained and its continued accessibility and use in the future. The goal of this paper is to analyse and identify those activities and procedures at the state-of-the-art level which are essential to ensure the data quality in this part of the survey execution. It tries to suggest quality standards in terms of guidelines, for procedures and content as well as to discuss how to ensure the execution, the procedures of the survey and the results and the original data are adequately documented. A specific focus is given to the consideration of standards for archiving in order to ensure a maximum of use of the data today and in the future.

No generally standardised definition of the individual steps and phases of processing, analysis and archiving for travel surveys exist. The analysis of the relevant literature shows no clear definitions and the terms are often used in a different way (Richardson et.al. 1995; Stopher; Jones 2003; Sammer et.al. 2000). A basic precondition for quality assurance is the use of a common terminology. In figure 1 a proposal for a harmonized terminology of the main seven steps (A to G) of the survey procedure is carried out, which represents a compromise between frequently used terms and a clear and logical structure. Because this paper is dealing with steps E to G the first four steps are not divided into further details. But it is evident, that this classification meets all needs of a sub-division into further steps, which is the task of the resource papers of the workshops A 1 to A 4 of the conferences. The structure of this paper follows the definitions of table 1. It has to be said that not all steps are treated with the same intensity, because this would be beyond the scope of this paper. The emphasis is put on those activities which are critical for the data quality.

Table 1: Proposal for a harmonized and standardized terminology of the steps of a travel survey procedure

<table>
<thead>
<tr>
<th>A. Survey Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. Sample Design</td>
</tr>
<tr>
<td>C. Survey Instrument Design</td>
</tr>
<tr>
<td>D. Survey Implementation (Execution)</td>
</tr>
<tr>
<td>E. Data Processing</td>
</tr>
<tr>
<td>E1. Database Building</td>
</tr>
<tr>
<td>E2. Questionnaire Editing</td>
</tr>
<tr>
<td>E3. Coding</td>
</tr>
<tr>
<td>E4. Data Editing / Cleaning</td>
</tr>
<tr>
<td>E5. Weighting / Correction</td>
</tr>
<tr>
<td>E6. Expansion</td>
</tr>
<tr>
<td>F. Data Analysis</td>
</tr>
<tr>
<td>F1. Validation</td>
</tr>
<tr>
<td>F2. Descriptive and Explanatory Data Analysis</td>
</tr>
<tr>
<td>F3. Presentation of Results</td>
</tr>
<tr>
<td>G. Data Documentation and Archiving</td>
</tr>
<tr>
<td>G1. Documentation</td>
</tr>
<tr>
<td>G2. Archiving</td>
</tr>
</tbody>
</table>
2. Data Processing

Data processing seems to be a rather less interesting issue, but it is an important part of the survey, which influences the quality of the data produced tremendously. The data processing can be structured into the following six steps: database building, questionnaire editing, coding, cleaning, correction, weighting and expansion.

2.1. Database Building

Database building includes the selection of the type of data structure, which is appropriate for the data collected and the software for the database system. At the beginning the requirements for the database and the software have to be analysed and defined. At this stage it is helpful to also consider the requirements for archiving and dissemination of the final survey data. Besides other needs those are the aspects to be considered: More complex database systems significantly increase the dissemination price because database enquiries are not royalty free. In addition complex database systems have high hardware requirements. Experience shows that relative simple and robust database systems have advantages if a frequent and unproblematic use of the data by different organisations and clients is desirable (DATELINE Consortium 1999).

There are two types of different data structure used for travel surveys (Richardson et.al. 1995):

- Flat-file database structure, which is mainly used in simple travel surveys with a small sample. This structure provides for each responding survey unit one record, where all information is stored in one record within one data file. This database structure is wasting storage units and is used rather rarely in nowadays, even when it is needed for most of the existing software packages for statistical analysis.

- Related database structures fulfil the needs of a hierarchical nesting of the data, which occurs often typically in travel surveys: A set of data describes the household characteristics, a subset of data is related to the household members, each household member is linked to a subset of trip data and each trip involves a data set of trip stages. The relational database is mostly used nowadays and fulfils the needs of modern data archiving systems.

The requirements for the quality of the database building include an intensive testing and a detailed documentation. Another issue must be addressed: If the survey design doesn’t use a computer-aided methodology often two separate databases are developed, one for the execution of the survey, which includes all information of the field work, and another database which consists of the collected data of the travel behaviour. In the future scientific analysis, especially for quality analysis and new weighting procedures information of the survey procedure which are not included in the final database, is often needed. Therefore it should be a minimum level of quality standard, that a version of the final database also includes the information collected during the fieldwork or that both databases are linked together in an appropriate way.

2.2. Questionnaire Editing

A very important activity to ensure the quality of the collected data is the editing of the completed questionnaires. Before the completed questionnaire of a personal interview survey goes to the coding procedure, the collected data have to be checked for their completeness, consistency and plausibility. This activity should be natural, but real life differs sometimes. The editing process consists of two stages (Richardson 1995): Interviewer and supervisor
editing. Both should be done as soon as possible after the interview was carried out. In principle each interview has to be edited by the interviewer him-/herself. CATI-technique enables this process online for the completeness, but the plausibility check, which cannot be made automatically, has to be done after the interview is carried out. If unclear information is found the interviewer can contact the interviewee again.

Supervisor editing is necessary for a sub-sample of the interviews collected and is as important as a quality check and supervision of the interviewer. It is very important to avoid slackness of the interviewer and therefore supervisor editing should be a parallel procedure during the whole fieldwork with two special intensive phases, one at the beginning and one at the end. A third editing procedure which is carried out for a small sample by a representative of the client should be discussed.

2.3. Coding and Data Entry
Coding and data entry is one of the steps of the surveying process where a tremendously fast development of methods can be observed. Today the most frequently used method of coding and data entry is based on interactive computer programs. Although there are some aspects related to coding which influence the quality and especially the comparability of the data with respect to data analysis, archiving and comparability of different surveys. Some of these aspects are addressed in the following paragraphs.

2.3.1 Coding Execution
If the use of a computer assists the process of interviewing (e.g. CATI and CAPI) the question arises whether to do all the coding at the same moment the information is collected. On the one hand the advantage to check the data consistency on-line is evident, on the other hand there are some experiences which indicate that the time delay in on-line geo-coding and the interaction needed to obtain a valid location can dominate the interview and confuse the respondent at least (Stenart 1997). In any case the coding procedure should be executed as soon as possible after the completion of the interview so that any problem can be cleared up while the interview is fresh in the minds of the interviews and the interviewee. There are some other aspects, which should be mentioned to avoid missing quality. Coding should be carried out by a limited number of coders, who are in a close contact to discuss questions arising and agree on it immediately. This way it is ensured that questions which arise can be cleared and the relevant information is disseminated immediately. It is regarded as a matter of course that coding and data entry is a process, which has to be monitored or supervised permanently as a quality control check. The coding information itself (e.g. coder, coding time) has to be recorded as well as the survey data collected. Double coding of each information seems to be unnecessary since interactive computer programs are in use, but doubts are raising and it should be looked into this question if this practice decreases the quality in an admissible way.

2.3.2 Missing Values and Use of Zero
There is a great requirement to implement some standardisation on how to treat missing values and the use of zero (see also 2.4.2 Missing Data). Until today a considerable variability still exists between different surveys but also within the same survey as to how missing data are recorded. It is very important that missing values are flagged in a proper way so that no confusion is possible especially with using a blank and a zero in a numeric field (Stopher, Jones 2003). At the 10th International Conference on Travel Behaviour Research in Lucerne, Switzerland, a paper was presented which discusses a proposal for some standardisations to overcome those problems (Stopher et.al. 2003):
No blanks standard; blanks are not legitimate code and all data fields should contain alphanumeric data; zero is used as a form of a nominal variable;

Missing data standard; missing data must be flagged by a specific code, of which the first digit has to be 9 for missing value independent of the cause. The second digit should be used to code the reason for missing as “don’t know”, legitimate skips or non-applicability of a question, etc.;

Correspondence between numeric values and the codes standard; if closed questions are offered in a specific order, the order should be reflected in the codes standard; if any question has the legitimate response zero (e.g. the number of trips of a person or the number of accompanying persons), the code for that response should be the number zero;

Coding standard for binary variables; in travel surveys a binary variable occurs often as a yes/no response or the gender male/female. To avoid errors it is recommended strongly to code the response “yes” as 1 and “no” as 0. For coding the gender use 1 for “male” and 2 for “female” while no response is 9.

2.3.3 Standardisation for Flexible Coding of Complex Variables
In most travel surveys the responses contain types of variables, which involve categories that may vary from survey to survey dependent of the specific goals of the surveys. Typical representative variables of that type are e.g. income, activity, employment status, education level, means of travel, travel purpose, etc. (Stopher et.al. 2003). It would be useful to define an international standard for the values, which enhance the comparability of surveys and avoid any ambiguity. Such a standardisation of coding would reduce errors because of the habitual effects on the one hand and would make comparative analysis easier on the other hand. For example, if the minimum set of standard categories for the travel purpose is 5, it has to be considered how to enable a possible way to extend the number of categories in a compatible way. An idea how this would work is shown in table 2. The main advantage of this method of standardised flexible coding is that an aggregation to the minimum coding categories can be made very easily by just dropping the last digit of the code.

Table 2: Proposal of a flexible coding-scheme for the variable “Travel Purpose”

<table>
<thead>
<tr>
<th>Minimum Categories</th>
<th>Minimum Coding</th>
<th>More Detailed Categories</th>
<th>More Detailed Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>01</td>
<td>Business</td>
<td>010</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- work place related</td>
<td>011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- non work place related</td>
<td></td>
</tr>
<tr>
<td>Commuting</td>
<td>02</td>
<td>Commuting</td>
<td>021</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- home related</td>
<td>022</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- non home related</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>03</td>
<td>Education</td>
<td>031</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- pupil</td>
<td>032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>student</td>
<td></td>
</tr>
<tr>
<td>Shopping</td>
<td>04</td>
<td>Shopping</td>
<td>041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- food</td>
<td>042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- non food</td>
<td></td>
</tr>
<tr>
<td>Leisure</td>
<td>05</td>
<td>Leisure</td>
<td>051</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- sport activity</td>
<td>052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- cultural activity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>Others</td>
<td>06</td>
<td>Others</td>
<td>060</td>
</tr>
<tr>
<td>Don’t know</td>
<td>98</td>
<td>Don’t know</td>
<td>988</td>
</tr>
<tr>
<td>Refused</td>
<td>99</td>
<td>Refused</td>
<td>999</td>
</tr>
</tbody>
</table>

C:\Documents and Settings\garym\Local Settings\Temporary Internet Files\OLK65A5 - Resource Sammer.doc
In connection with the “activity” and “travel purpose” variables it has to be stated, that in many of the travel surveys the mainly used categories of the mentioned variables should be extended, to enable a more appropriate handling for behavioural analysis and modelling. Especially the “leisure trip” category comprises about 25% of all trips but this categorisation doesn’t meet the modelling needs. The same needs can be identified for the variable of the big variety of leisure activities as sporting, cultural, visiting friends, religious activities etc. (Axhausen 2003). These different types of leisure activities have varying requirements for modelling exercises. Similar demand can be identified (Sammer 1998) for the characteristics “car driver” and “car passenger”, which should be recorded more in detail into “solo driver”, “driver with accompanying household member” and “driver with accompanying non household member” (car sharing mode). This specification of categories would meet the requirements of the young research field of behavioural analysis and modelling of measures of mobility management. In this content it has to be mentioned that the attribute “people with mobility restrictions” should also be recorded, when this has an impact on general mobility and this effects restrictions to the mobility of accompanying persons.

2.3.4 Geocoding
Today geocoding belongs to the state-of-art for coding addresses of origin and destination. But experiences of practical work show that there are some problems not yet solved satisfactorily, and this can lead to missing trip information. Different requirements have to be taken into account for local travel surveys and long-distance travel surveys (Richardson et.al. 1995; DATELINE Consortium 2003). For local travel surveys the following methods of geocoding are generally used: geocoding of full street address, of cross street address, of landmarks and geocoding by sampling. The latter one can be described as a geocoding imputation method which is used to compensate incomplete information of locations (see 2.4.2 Missing Data). Even then it is very important that the category of geocoding is also recorded to provide the information about the accuracy of geocoding. Local travel surveys require a very disaggregated form of geocoding, best suited is the coding of latitude and longitude. In practice experience results in an achieved quota of well-coded locations of about 80%. In some local travel surveys the completion of well-coded data can raise up to 95% when a good data base is available and if the respondents are motivated well to indicate their addresses accurately. There are mainly two solutions to increase this quota: Firstly to go back to the interview stage of the survey and to explore additional information from the interviewee. This is the better but a very costly way. The second solution is the imputation of the missing addresses which can be a useable way but is even not satisfying for each purpose.

Geocoding for long distance travel surveys has different requirements, which are similar for out-of-region addresses of local travel survey. Here the spatial aggregation level needed is not so accurate. In general coding is required at the level of a traffic analysis zone. Geocoding should be carried out at the most disaggregated level of the collected information from the respondent. Generally this is the city, community or region, where the origin or destination of the region is located. The coded information should include again the latitude and longitude, which represents in this case the geographic centre of the city, community or region. Experiences with the DATELINE project (Consortium DATELINE 2003) show a satisfying quota of over 95% of well-coded locations.

2.4 Data Editing / Cleaning
No clear standardized term of this step of data processing exists (Figure 1). The following terms are used in this context: data editing, cleaning, correction, whereas the term data correction is also used for the weighting procedure. To promote a clear standardised survey
technology it is proposed to introduce data editing and cleaning for this step, which is dealing with checking of the range error, logical consistency and missing data in so far as the errors can be detected during editing and corrected by referring to the original questionnaire as well as recontacting the interviewer or interviewees.

2.4.1 Errors and their treatment
It is evident that five main types of bias can occur (Richardson et.al. 1995), which need a different treatment in order to ensure a high quality data set. These types arise from several sources as sample drawing, respondent, interviewer, data coder/typist (Table 3):
- Sample drawing bias which causes a deviation from the principle of random sampling and results in a biased coverage of the population. The treatment of this error is discussed in 2.5 Weighting / Correction.
- Range error is mainly a result of typing and recording where the code value is outside of the permissible code for that response. Mostly this error can be eliminated by referring to the original questionnaire but also double coding of the questionnaires helps to avoid some of these errors.
- Logical consistency errors occur if responses to different questions of one or several persons linked together give inconsistent results, e.g. if the last trip of the day is not to home or if joined trips of the household members do not fit together. To ensure high quality it is recommended to define survey specific rules how to deal with identified inconsistencies. A good example for such rules is described in a paper by Arentze et.al. (1999). This type of error can be corrected in many cases if the questionnaire is checked again. If the inconsistency cannot be eliminated in this way a new contact with the interviewee is very helpful as the experience of the New KONTIV-design, which includes such a procedure (DATELINE 2003, Brög 2000), has shown. This procedure is expensive but very successful.
- Missing data (unit and item) can be caused by coders and typists or by the respondent. An error of the first type of cause can be eliminated by double coding or checking the questionnaire, the second one is discussed in 2.4.2 Missing data.
- Coding error, which consists of a false value, which is not outside of the range of codes permissible for that response, can only be identified by double coding. Double coding requires a great effort but is the only way to correct this error. If such an error is identified, the correction needs a check of the original questionnaire. A minimum standard requirement should be fixed, so that in a pilot test a sub-sample of the interviews should be double coded, to assess the quality of the coders as well as for CATI or mail back surveys. The result of such a pilot test enables to assess the coding quality and to set additional training if necessary.

The editing/cleaning procedure at least needs a manageable definition of a useable and complete response. Such a definition fixes which key variables must exist in a defined quality to meet the minimum criteria of completion. It doesn’t seem to be meaningful to define general standards because the objectives and application of travel surveys are different. In the DATELINE survey, e.g. a useable household of person data set meets the following criteria of required person and behavioural data: age, gender, destination of journeys, duration of journeys, purpose of journeys. Some CATI-software used for travel surveys eliminate incomplete responses as soon the full sample has been obtained. This practice leads to a loss of important information for further analysis of several elements of the travel survey. Therefore it has to be stated, that incomplete responses, which are not useable for the sample, should be retained and documented in any case as a non-useable return. The definition of a useable response return is also influencing the response rate.
Quality management is based on information. Therefore the whole editing/cleaning procedure has to be documented in a qualitative and quantitative way: which data checks are made and a statistical analysis about the frequency of the identified types of errors.

Table 3: Type, sample and treatment of errors

<table>
<thead>
<tr>
<th>Source of Error</th>
<th>Sample Drawing Bias</th>
<th>Logical Consistency</th>
<th>Missing Data</th>
<th>Coding Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drawing and Coverage of Population</td>
<td>C</td>
<td>-</td>
<td>C</td>
<td>-</td>
</tr>
<tr>
<td>Respondent</td>
<td>B</td>
<td>B</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Interviewer</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Coder, Typist</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>
| Classes of Error Treatment:      | A: Errors which can be eliminated in most cases by the data editing and cleaning procedure. There seems to be no need to record the code of the corrected values from the quality point of view except for the error statistic.  
B: Errors which can be eliminated mostly by the data editing and cleaning procedure. The correction procedure should be documented for further investigation and analysis of the original raw data by setting a flag.  
C: Errors which have to be treated by the weighting and correction procedure (section 2.5).  

2.4.2 Missing data

Missing data caused by the respondent need a subtly differentiated treatment. It has to be distinguished between unit non-response and item non-response. Unit non-response occurs where a survey unit (a household or a person) fails to respond to the survey. Item non-response is defined that the respondent has provided responses to the most questions, but has failed to answer a specific question (e.g. the trip purpose or the whole trip). It has to be stated that some authors make a distinction between non-reported trips and the item non-response (Richardson, Loies 1997). Unit non-response is discussed in section 2.5 Weighting / Correction. It is appropriate to distinguish between three methods of how to treat the missing information of items:

- Imputation by re-contacting the respondent; one of the most promising procedures to get the information of missing data is the re-contact of the respondent. This method is part of the new KONTIV-design and eliminates the error totally but requires great effort and is expensively (Brög 2000). This type of treatment is to be preferred in comparison with the following two methods from the point of quality.

- Imputation of the missing information by logical consistency checks from other information; Stopher and Jones (2003, p. 30) call this the inference method. In order to ensure an objective and comprehensible inference procedure the rules have to be defined for each variable in detail. A very strong quality condition would be that only those imputation actions are accepted which are done by two independent persons and which lead to the same result.
- Imputation of missing data based on associative rules or other procedures (e.g. Monte-Carlo-random based, neural networks). Firstly it is necessary to define a standard of which variables are allowed to be imputed. The question is closely related to the minimum requirements for a usable return of incomplete interviews. This type of imputation should not be allowed for missing items which are part of the defined key variables of a usable household or person data set. It must be aware that at least any imputation of a key item or variable of the survey (e.g. missing person, trips, destination, trip mode, trip purpose) has the effect of reduction of the sample size related to the imputed variable besides other bias risks. Such a reduced sample size has to be taken into account for calculation of the sampling. The question arises if a limit of the rate of maximum number of imputations can be or should be defined (Stopher, Jones 2003, p. 31). In principle such a limitation shouldn’t be defined as a general standard, it has to be considered for each specific case.

An important question evident is if it is possible and desirable to define a minimum standard of imputation methods. The analysis of the existing methods shows a big variety with big differences in the consequences and quality achievable ranging from simple mean imputation to complex hot-deck or neural network imputation (Poduri 2000, Richardson, Loeis 1997 and 2001, Wilmot et al. 2003, Han et al. 2001). It wouldn’t make sense to fix the requirements for a specific method of imputation but to define a minimum standard of principles detached of a specific procedure, e.g.:
- the imputation method should be based on a missing item analysis, which takes into account the relation of the missing items to other variables in the current sample including the result of a validation survey analysis. The imputation should not only ensure that imputation is multiplying data that is not missing but also adding data that is missing;
- the number of imputed values should not dominate the number of reported values of the considered variables, being aware that imputation is always a risk and in reality reduces the sample in respect to the type of the imputed variable;
- key variables of a survey which are the content of the defined “usable return” as mentioned before should not be imputed;
- in any case of imputation the imputation process should be documented totally and the imputed values have to be flagged so that the analyst of the data is fully aware which information comes from the respondent and which not.

2.5. Data Weighting

Data weighting is necessary when the sample is stratified or any other disproportional sampling procedure is used, and if any other bias is expected in the cleaned and corrected data after the above described data editing, cleaning and correction process. From the quality point of view the following questions are of interest:
• Under which circumstances is weighting of the data required?
• Which method of weighting is appropriate for the biased data?

Weighting is related to key variables of the data, which have to be defined. For travel surveys the following variables are mainly on focus, e.g.: household, person, number of journeys or of trips travelled, distances travelled, travel time. The travel-related variables are distinguished through different characteristics as purpose and mode. It follows from this that any bias has to be considered in relation to those variables.

2.5.1 Criteria of weighting

From the quality point of view it is more efficient to ask the question the other way around: Under what circumstances is no data weighting required? It is easier to formulate such
appropriate criteria: No weighting is required if statistical tests guarantee that no significant bias can be identified for the following proposed key variable:
- Distribution of household size and car ownership rate in relation to the number of cars;
- Distribution of person age, gender and occupation, nationality, etc.;
- Regional distribution and distribution of the set dates of the survey;
- Travel behaviour between respondents and non-respondents characterised by the number of journeys, trips, distance travelled, travel time budget, etc.;
- Item non-response.

The statistical test between the household and person variables, which are the first three above mentioned key variables, require appropriate secondary independent data resources. For the travel behaviour variables either a validation survey for non-reported trips/journeys, a non-response explorative survey is required or any other appropriate analysis can be used, e.g. the effect of number of reminders/contacts on questionnaire response rates (“Respond Speed Analysis”). A non-response explorative survey has the goal to explore some key variables of the travel behaviour for a sub-sample of the unit non-response. This result can be used to estimate the travel behaviour of the unit non-response. A validation survey of non-reported trips has the goal to survey a sub-sample of the respondents again to investigate if all trips are reported. As an example in figure 1 and table 4, the test result of the non-response bias for the DATELINE survey in Germany is documented. For this test, two different methods are applied, a non-response survey and a response speed analysis. The response speed analysis in figure 1 indicates that respondents who respond earlier are more mobile than persons who respond later in a mail-back survey. The extrapolation of the response speed-function is used for an estimation of the travel behaviour of the non-respondents. The response speed analysis in figure 1 indicates a significant bias, whereas the comparison of the main survey with the non-response explanatory survey shows no significant difference in the average number of journeys/households. One can see that the sample size of the non-response survey should not be too small to receive significant results.

Figure 1: Effect of the increasing response rate (response speed) on the average numbers of long distance journeys caused by several reminder activities, DATELINE Germany (Neumann 2003, p. 167)
Table 4: t-Test of average number of long distance journeys between respondents (main survey) and non-respondents (non-response exploratory survey); DATELINE Germany (Neumann 2003, p. 180); the result indicates that no significant difference exists between the mean values of the number of journeys in both surveys.

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Number of journeys / households and year</th>
<th>Standard deviation</th>
<th>Test Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Survey</td>
<td>7,922</td>
<td>3.60</td>
<td>7.47</td>
</tr>
<tr>
<td>Non-Response Exploratory Survey</td>
<td>119</td>
<td>3.40</td>
<td>5.43</td>
</tr>
</tbody>
</table>

\[ t_0 = 0.29 < 1.65 \ (\alpha = 5 \%) \]

2.5.2 Weighting method
In general principle all methods of weighting can be distinguished: factor and unit imputation weighting (Herry 1995). Whereas factor weighting is used mostly, unit imputation weighting has a minor important standing in practical survey work. Unit imputation weighting follows the idea of eliminating bias in a casual way by enriching the diversity of the distribution of the sample, in contrast with factor weighting which is based on the diversity of the existing distribution of a defined stratification of the sample. From the view point of quality, the unit imputation weighting procedure seems promising and requires further research.

Weighting stratification and sample size
Factor weighting is the most widely used method, but to ensure high quality in the results, some principles have to be adhered to. Factor weighting is carried out on an hierarchical aggregation level of the survey variables, e.g. household, person, journey, trip, trip stage. On each level, the variables are divided into appropriate multi-dimensional stratification characteristics. For example, on the person level, the multi-dimensional stratifications in age, gender and occupation classes are used. If the distribution of the population is available for the multi-dimensional stratification by means of secondary data, it is necessary to test whether the sample size of each class is big enough to ensure an unbiased weighting procedure. As a rule of thumb, the minimum sample size of a stratification should not fall below about 30 cases. This fact has to be considered for the definition of the sample size in relation to the selection of the weighting procedure. To avoid this problem, the weighting procedure often is carried out in a stepwise way. This means that the weighting procedure enables only an unbiased boundary distribution of each person characteristic but not of the multi-dimensional distribution of these person characteristics. Such a weighting procedure requires some quality principles:
- The stepwise weighting algorithm should be mathematically consistent (Sammer, Fallast 1996);
- It is recommended to use an algorithm that follows the principle of maximum entropy or minimum information gain (Wilson 1974, Snickers et al. 1977, Sammer et al. 1982). The goal if this algorithm is to provide an optimal solution for the multi-dimensional matrix of the weighting stratification characteristics by given a boundary distribution of the weighting characteristics which follows the defined optimum of the maximum entropy condition.

As is a matter of course, the weighting procedure must guarantee the original sample size in total by any appropriate standardisation algorithm.
Unit non-response weighting
Another issue should be addressed. Often the weighting of the unit non-response error is carried out by factor weighting of trips. This method ignores the evidence that unit non-response is mainly caused by the different willingness of respondents. Therefore the weighting procedure should be made on the level of persons or households, whether the non-respondent information arises from a non-response speed analysis or from a non-response explanatory survey. Figure 2 indicates the result of the response speed analysis, namely that the share of households with few long distance journeys increases with a rising response rate for mail back surveys and vice versa. Figure 3 shows the opposite result for a CATI-survey: The share of households with few long distance journeys decreases with a rising response rate.

Figure 2: Effect of increasing response rate (response speed) on the share of households with different long distance journeys (≥ 125 km as the crow flies), sample size = 861 households, DATELINE Austria (Neumann 2003, p. 171)
Figure 3: Effect of increasing response rate (response speed) of a CATI-survey on the share of persons with different long distance journeys (≥ 125 km as the crow flies), sample size = 7,363 persons, DATELINE France (Neumann 2003, p. 174)

Weighting statistics and documentation
When the weighting procedure is finished, it is important to document the weighting method including the analysis so that it is clear how it was carried out. That implies the following:
- Weighting documentation: aggregation weighting level, weighting steps and stratification, algorithm, target variables of weighting, results of the weighting analysis, critical comment on well solved and less solved distortions, etc.;
- Weighting statistics: distribution of weights, effect of the single weighting steps on the target variables, information on the convergence speed if iterative weighting algorithms are used.

In figures 4 to 6 some examples of weighting statistics of the DATELINE project are documented, which provide an insight into the quality of weighting of a travel survey.
Figure 4: Distribution of weights for Zone NUTS 1 of Austria, DATELINE Austria (DATELINE 2003, Weighting and Grossing up Report, p. 36); the form of the distribution of weights allows conclusions on the quality of the sample.

Figure 5: Effect of the single weighting steps on the average number of long distance journeys per person and year (≥ 125 km as the crow flies) as well as on the mean of distance and duration travelled, DATELINE Austria (Neumann 2003, p. 205)
2.5.3 Efficiency of Weighting
Data weighting is one of the key steps to achieve high quality data of travel surveys. Data weighting including the necessary analysis is a resources consuming effort. Therefore a balance should be achieved between the increase in the quality of results obtained in each step of the weighting procedure and the effort needed to perform it. That means more emphasis should be put on the cost-benefit ratio of data weighting. A admittedly not representative analysis of reports of travel surveys indicates, that for many travel surveys
- the weighting procedure and its efforts are not well documented,
- the improvement of the quality of the results of the individual weighting steps is not shown.

The cost-benefit ratio of the weighting procedure should be paid more attention in respect of saving resources.

3. Data Analysis

3.1 Validation analysis
A validation analysis indicates and verifies the accuracy of the surveyed data and is one of the most important steps of the survey. Before the cleaned, weighted and expanded dataset is ready for final use, it is recommended to carry out a validation check. The result of such a validation analysis can lead to a loop back to the previous steps of the survey if it doesn’t fulfil the defined quality level. This validation analysis can be made by two different organisations depending on the function of this validation exercise and the type of quality management:

(a) The organisation that is responsible for the survey project, if the validation analysis has the function of a final quality test: This validation check is conventional and should be part of the survey contract if type (b) is not chosen. It can be said that some survey studies have included such a validation analysis (DATELINE 2003), but it is not common state of the art.
(b) An independent organisation in the function of a referee or the client itself is responsible for the validation check: An independent referee organisation reviews whether the quality standards are fulfilled. Experience shows that it is very effective if the validation analysis is carried out not only at the end of the survey process but also an accompanying quality check is implemented to ensure the quality at each stage. This can be performed by an independent quality board consisting of independent experts. In any case it is important to define the quality standards at the beginning of the survey and include them in the tender.

The validation check must take into account the specific situation and framework of the travel survey, but some hints can be formulated to achieve high quality:

- **Validation survey, recollection of a sub sample (Jones and Stopher 2003, p.35, Stopher et.al. 2003)**
  A validation survey reports the data collection process with a sub sample of the previous by collected interviews. The validation survey is carried out by different interviewers and should use also a different way of data collection. As a rule this validation survey usually is focused on some key variables. As a minimum standard it should include the information, which is defined as “usable terms”. The validation statistic consists of a comparison of the original data and the data collected in the validation survey. The quality of the data is assessed by a comparative statistic, which identifies the discrepancies. Stopher et al. (2003) recommend the use of the root-mean-square-error, which provides a dimensionless measure of the average error or deviation of two values of the same variable:

  \[
  \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \frac{(V_{1,i} - V_{2,i})^2}{V_{2,i}}} \times 100\% \]

  **RMSE**: root-mean-square-error of variable V (e.g. number of trips)
  \(V_{1,i}\): collected value from the main survey
  \(V_{2,i}\): collected value from the validation survey
  \(n\): sub-sample size of the validation survey
  \(i\): index of interviews of the survey

  For the interpretation of the results it is necessary to define a limit of RMSE, which is acceptable from the quality point of view. It has to be stated that considerations of the appropriate size of the validation survey have to be taken into account.

- **Validation survey by passive tracking survey**
  New technology enables a complete recording of all person or vehicle movements over the reporting period of a survey. Experiences show (Wermuth et al. 2003, Murakami et al. 2003) that surveys based on GPS or mobile phone are an appropriate instrument to collect data of a sub sample of the main survey and to use this information for validating (and correcting) the data. It is recommended to use also the root-mean-square-error for the presentation of the results of the validation statistics.

- **Plausibility check with secondary data (Sammer 1997)**
  A plausibility check is the comparison of the result of target variables with appropriate results of independent secondary data sources. The type of the plausibility check is dependent on the available data sources, but this should be considered during the preparation phase of the survey. The following possibilities are noted:

  - A check of the total number of trips or journeys, which cross a defined cordon, deduced from the travel data and e.g. from independent traffic counts, is an appropriate assessment of the accuracy. In cities with a river, counts on bridges can be
made, or other breaks utilised such as railway lines with few streets that cross. The traffic movements counted at such network sections, differentiated according to modal split, should agree with the registered origin-destination travel behaviour. One should note that differences are not always due to survey errors. Other causes include: differing survey periods, different “populations” caused by tourist traffic etc. If one excludes these influences, experience shows that household surveys tend to under-report car and business trips;

- Comparison of the kilometres travelled for motorised transport as deduced from the household survey, with the average annual kilometres of the vehicles, determined from an independent car-ownership survey.

The result of the validation analysis is an essential part of the survey documentation. If the result doesn’t fulfil the defined quality standards, then a revision of the weighting/correction step may be appropriate.

3.2. Data presentation

The presentation of data seems to be a very simple issue using the diverse software on the market. But from the quality point of view, most of the presentations of surveys results neglect to inform the user of the results about the statistical accuracy and sampling error. It should be completely natural to provide the confidence interval for a presented key value or, for the comparison of two values information as to whether their difference is statistically significant or not. However, analysis of existing presentations of travel survey results indicates that information about the accuracy of descriptive statistics is rarely provided. This problem becomes more important when web-based data information systems for travel surveys are publicly available (see section 4.3). What could be the reason for that fact? It must be stated that the valid calculation of confidence intervals and other statistical tests requires the specific taking into account of the weighting and correction procedure (Hautzinger 1986, Sammer 1997, Brög et al. 1986), which is not included in most of the available software packages.

4. Documentation and Preservation

4.1. Status-quo Analysis

Documentation and preservation of transport surveys has become a frequently discussed topic at international conferences in the last decade and an increasing number of papers have been produced (Axhausen 2000, 2001; Axhausen, Wigan 2003; Pendyala 2003; Sharp 2003; Wigan 2001) because a survey should not end with the delivery of the data for the primary client once the collected data are valuable for the usage of other clients. While the need and value of high quality documentation and preservation is well recognised in the community of the transport research field many activities cannot be embraced to make progress in the normal professional practise especially with respect to the access to transport survey data. The question arises what are the reasons for this unsatisfactory situation? The implementation of institutional websites as a very suitable media for communication between experts enables a good access of information about travel surveys, but unfortunately they mostly provide the information in very differing quality and detail. The survey data are generally not available. No systematic analysis about this unpleasant situation was carried out but some cases are evident:

- The lack of awareness of public planning agencies and of the organisation which carries out the surveys about the importance of these data for the scientific community. The latter ones have mostly a commercial and no scientific background which doesn’t support one’s own initiative for data preservation.
- The missing willingness of planning agencies especially regional ones to finance the archiving of surveys for further use of the data outside of their area of responsibility and interest.
- The absence of an appropriate national organisation in many countries, which feels responsible for data preservation. The archiving of data by any internal travel organisation could raise legal questions of data protection.
- There are also scientific reasons to cite such as the absence of well understood and adapted patterns as well as standardisation as to how to document and archive such survey and data in the field of transport.

Summarising the causes, that public available documentation of transport data is not the rule, it can be categorised in two main classes, which needs handling differently:

1) Reasons which can be located in the institutions and their representatives that are engaged in transport surveys. It is very important to raise their awareness for this problem. How can this be initiated in a successful way? It would seem to be the thing to involve the existing relevant public organisations and the scientific networks and to start something like an information campaign and to keep this topic on the agenda of relevant conferences and journals.

2) Reasons which are of a scientific nature such as the absence of clear and easily useable guidelines which contain the minimum requirements of standardisation for transport surveys and data archiving. It has to be stated that the methodology of documentation and archiving and its technology is now well developed (e.g. NESSTAR, web-based data analysis software) even though it didn’t become state-of-the-art in the field of transport surveys.

4.2. Objectives and Requirements on Documentation and Preservation
The objectives and requirements have to be oriented to the involved organisations, persons and potential users:

- The sponsoring of financing organisations and data owners in general has a specific use of the data primarily in mind for their own purpose. They often don’t want to be hindered by the documentation archiving and maintaining efforts. It is very important that in future state-of-the-art documentation and archiving becomes a part of the whole survey task and consulting contracts.
- The organisation carrying out the survey must be responsible for the documentation and archiving of the data because only they have the necessary inside information. But the following framework conditions must be fulfilled: cost coverage of the documentation and archiving efforts, clarification where and by which competent organisation the documented data are stored and maintained, appropriate and standardised documentation guidelines and simple not lavish implementation procedure for the documentation and archiving.
- Data archives at the national and international level have specific requirements and goals: data structure of survey data to be archived, which can be implemented optimally into archive systems, long term sponsorship for the archiving costs to enable a long term storage, etc..
- Potential users want an easy access to archived data with low cost, standardised documentation, data character, data structure and coding to enable comparative analysis of different data sources etc. The users of data archives have to be specified in two main groups which have different needs:
  - the political interested users (public planners, policy makers);
  - the scientific analytical researcher (professional data user).

The political oriented user wants fast and easy access to results of transport data, which are of public interest. He is strongly interested in on-line tabulation systems, which can
provide any results in a short time. He is also interested in good presentation of the required data results in the form of well understood graphs, tables and maps. For these type of cases it is very important that the results are presented in a way which avoids any misinterpretation which can occur e.g. by not statistically significant results. Therefore it is very important that any on-line tabulation software tool, which can be used by everybody, is provided with a minimum level on statistical validation (e.g. confidence interval, significance tests). NESSTAR light client software provides plenty of such tabulation software but it has the disadvantage that it doesn’t provide any statistical confidence information as it should by common standard (Sammer 1997).

The scientific analytical researcher has different requirements. Firstly, he wants to have access to metadata information on transport survey data. On-line tabulation is of secondary interest but it is also a valuable tool for pre-analysis. Great interest exists in a cross survey comparability of different data resources which can be facilitated only by an extensive standardisation of the data structure and coding. The dominant requirement lies in the access to the raw data of the surveys.

The outcome of the requirements defined above leads to the conclusion that our present predicament demands the implementation and use of metadata and archives to preserve the collected data and to enable an efficient use of it.

4.3. Technology of Metadata and Data Archives on the Market

There is no need to reinvent the wheel of metadata and data archives, they are well developed and used in other scientific fields as in the social and behavioural sciences. Metadata are besides the data storage the essential information of a data archive. Metadata have mainly the following function and objectives (Pendyala 2003):

- Assistance to facilitate the identification, evaluation and retrieval of data sets which are of interest for a specific reason;
- Facility to evaluate the relevance, efficiency and quality of the identified data sets;
- Possibility to retrieve the new data of the survey including the documentation and provision of all information which is necessary to use the data.

Mainly libraries and information system organisations have developed the formal structure of data description and specific languages were developed in the last decade, so that numerous metadata standards are available on the market (Pendyala 2003, Wigan et al. 2002). These standards DDI (the Data Documentation Initiatives – www.icpsr.umich.edu/DDI/index.html) have to be emphasised as one of the most advanced, flexible and user friendly metadata standard. DDI pursues the goal to create an international accepted methodology for the content, presentation, transfer and preservation of metadata about datasets in the behavioural and social field of sciences. DDI uses the XML (eXtensible Markup Language), which enables to make use of the advantages of the web function abilities.

The DDI is used by many national and international data archiving projects. One project must be emphasised because it seems to be very relevant and suitable for transport data and their archiving requirements. NESSTAR (Networked Social Science Tools and Resources – http://www.nesstar.com) is a web-based data analysis software which enables users and publishers of social data to exploit data and information via the web including survey data on-line tabulation and relevant other information.

NESSTAR offers four different archiving tools:
- NESSTAR Publisher is an advanced data management suite that enables to convert, enhance and manage data for publication to the NESSTAR server.
- NESSTAR Server allows the data provider to set up an interactive data publishing and dissemination service.
NESSTAR Explorer produces an integrated data discovery, browsing and retrieval platform, which runs on the user’s desktop and which interacts with the data on a server. NESSTAR light allows the user to operate most of the services of NESSTAR Explorer but within the confinement of a standard web browser. He can search for, locate, analyse and download a wide variety of statistical information as well as information relating to the NESSTAR server and its operation. There is no need for specialised software to be able to view the mobility data stored on the NESSTAR server.

It has to be stated again that the DDI codebook, which is used in NESSTAR, still needs a standardized classification for transport surveys which allows an automated processing and ensures a completely standardized documentation. In Annex 8.1 an improved proposal for the standardisation for the standardized classified documentation is presented for discussion, which is based on some recent papers (Axhausen 2001, Pendyala 2003, Richardson et al. 1995, Chalasani et al. 2002) and on the DATELINE-project (DATELINE Consortium 2003).

For the European research project DATELINE, which implemented and carried out a European long-distance travel survey in all 15 member states of the European Commission and Switzerland, a European Long-distance Mobility Information System (ELMIS) was developed and implemented, which is partly based on NESSTAR light (DATELINE 2003). It is available to the public since December 2003 on the web (http://chi.fg.uni-mb/elmis/). This example of data archiving and documentation is typical for a rather complex survey and allows both publisher and user to collect experiences with metadata software and archiving systems in practice and in concrete terms with NESSTAR. In Annex 8.3 a short description of ELMIS is attached.

4.4. Future Needs in Data Archiving and Documentation
Although tremendous progress in the archiving and documentation mainly in the social science and in the development of software tools and archiving technology was made, wide action needs for the travel survey community can be identified, which would contribute to raise the quality of documentation and to widen the possibilities and functionality of metadata and archives in connection with the web technology. These needs can be classified in two sections: organisational and institutional needs as well as standardisation needs.

4.4.1 Organisational and institutional needs
As discussed in section 4.1 it is very important to raise the awareness of the relevant organisations, ministries, agencies etc., who are engaged in transport surveys, that a good documentation, archiving and preservation of transport data is essential for the development and dissemination of knowledge. At least in the long term consideration expenses can be reduced for travel surveys if the international transport community would have the knowledge of and access to the relevant travel data at the national and international level. The possible way of doing this should be discussed at the 7th International Conference on Travel Survey Methods in August 2004 in Costa Rica. The formulation and dissemination of something called a “Charter for Transport Metadata” to all relevant organisations, which are dealing with travel surveys, would support this goal. The main important message of such a charter should be the integration of data archiving and documentation in the study contract.

4.4.2 Standardisation needs for documentation and archiving
It is very important to enhance the efforts of standardisation. This could enlarge the opportunities and functionalities of metadata analysis if the level of standardised information about travel survey would be improved. On this topic there is a large need of action for the transport scientific community. Even so the standardisation is a necessary task, and we should
stick to the facts that every standardisation should only define a minimum level, which is achievable with justifiable effort for the development and the use. This minimum level should be based on the identification of generic essential elements and characteristics that are common to all surveys and which are accepted by the transport community. Otherwise this exercise is doomed to fail.

- **Standardisation of a typology of travel surveys**
  Travel surveys can be classified in a large variety of types, which require different types of document definitions. Therefore archiving needs a hierarchical structure of survey classification, which enables to assign all types of surveys. Pendyala (2003) has made a first proposal which shows us how difficult it is to develop a consistent hierarchical structure. In Annex 8.2 a revised version of this proposal is documented. The following step consists of the development of a typology of the generic and specific characteristics and elements for travel surveys. Such a list of identified elements is meant as a base for developing it.

- **Development of vocabularies for metadata documentation of transport data (transport thesaurus)**
  Beside the formal structure of metadata the development of appropriate vocabularies for the field of transport data is necessary. These vocabularies need a broad acceptance of the transport community to gain the full benefit and to make use of the whole possible functionality from such metadata bases. Such a task has to be done in an interdisciplinary and co-operative way. Wigan (2001) suggests that e.g. the committees of IATBR, IATUR, AET and TRB together with libraries and Social Science Data Archives should carry out this task. One option to make a start would be the thesaurus of the OECD ITRD Transport document database, the development of which started over 30 years ago.

- **Development of a standardised flexible coding scheme**
  As discussed in section 2.3.3 the development of a standardisation of a flexible coding scheme for travel surveys would represent a milestone of the development. It should represent the minimum level of documentation which leaves to the user to use a farther detailed coding scheme without loosing the minimum level of standardisation and it would open a wide range of possibilities of analysis.

- **Enhancement of metadata standard to accommodate the spatial dimension of data** (Pendyala 2003)
  The spatial dimensions of transport data are evident. It seems that the currently available metadata standards do not take into account the spatial characteristics of travel data as they do include the information about time periods covered and time of the data collection. Therefore there is a need that the spatial dimensions represented by origin and destination will be incorporated in the metadata standard format.

5. **Conclusions**

The definition of quality standards and some forms of guidance are also appropriate for the steps data processing, analysis, documentation and archiving of travel surveys. Analysis of the state of the art indicates that knowledge is well developed even though some questions need further research. But this knowledge is distributed among different sources and needs to be put together and standardised in an appropriate way. The definition of a minimum common denominator of standards for data processing and analysis would help to raise the data and output quality. For the definition of quality standards it is also very important to take into account the cost-benefit ratio between the effort needed to achieve a defined quality of results and the quality level itself. Although this question is well analysed for surveys in other fields, not much knowledge seems to be evident in the field of transport survey. The standardisation of elements of travel surveys, with regard to documentation and archiving, would enable
firstly utilisation of all the benefits of the web-based technology of metadata and data archives
on the market. But until that milestone is achieved, a great deal of internationally coordinated
work and effort has to be done, requiring a lot of institutional, organisational and funding
considerations in order to initiate a successful development. If such quality standards become
established in invitations to tender for travel surveys, it would effect a great leap of quality.

6. References

Arentze, Th., Hofmann, F., Kalfs, N., Timmermans, H. (1999); SYLVIA: A System for the
Logical Verification and Interference of Activity Diaries, Paper submitted for presentation at
the Transportation Research Board Conference, Washington D.C., January 1999, Contact
Address: E-Mail eirass@bwk.tue.nl.

Axhausen, K.W. (2000); Presenting and Preserving Travel Data, in P.M. Jones and P.R.
Stopher (Eds.) Transport Surveys: Raising the Standards, Transportation Research Circular,
E-C008, II-F-1 - II-F-19, Transportation Research Board, National Research Council,
Washington, D.C.

Axhausen, K.W. (2001); The public use of travel surveys: The metadata perspective.
Resource paper for the 2nd International Conference on Transport Survey Quality and
Information, Kruger National Park, South Africa, August 2002; Institut für Verkehrsplanung
und Transporttechnik, Straßen- und Eisenbahnbau (IVT),

Axhausen, K. (2003); Aktionsräume und die Suche nach Abwechslung: Aktuelle Ergebnisse
von Langzeitbefragungen; Presentation at the Institute for Transport Studies, University for
Natural Resources and Applied Life Sciences (Bodenkultur) Vienna on June 23th 2003;
Vienna, Austria.

Axhausen, K.W., Wigan, M.R. (2003); Public Use of Travel Surveys; The Metadata
Perspective, in P.R. Stopher and P.M. Jones (Eds.) Transport Survey Quality and Innovation,


Chalasani, V.S., Schonfelder, S.; Axhausen, K.W. (2002); Archiving Travel Data: The
Mobidrive Example, Arbeitsbericht Verkehrs- und Raumplanung, 129, Institut für
Verkehrsplanung und Transportsysteme (IVT) ETH Transport Data Archive, ETH, Zurich.

DATELINE Consortium (1999); Description of work of DATELINE, Technical Annex
(Design and Application of a Travel Survey for European Long-Distance Trips, based on an
International Network of Expertise), 5th Framework Programme, European Commission –DG
Tren.

DATELINE Consortium (2003); European Coding Book for Long Distance Travel Surveys;
Deliverable D6, DATELINE (Design and Application of a Travel Survey for European Long-
Distance Trips, based on an International Network of Expertise), 5th Framework Programme;
European Commission –DG Tren.
DATELINE Consortium (2003); *ELMIS - European Long-Distance Mobility Information System, User Manual*; Deliverable D9, DATELINE (Design and Application of a Travel Survey for European Long-Distance Trips), based on an International Network of Expertise, 5\textsuperscript{th} Framework Programme, European Commission –DG Tren.

DATELINE Consortium (2003); *European Long-Distance Mobility Information System (ELMIS)*; Deliverable D9, DATELINE (Design and Application of a Travel Survey for European Long-Distance Trips), based on an International Network of Expertise, 5\textsuperscript{th} Framework Programme, European Commission –DG Tren.

DATELINE Consortium (2003); *Final Report of DATELINE*; Deliverable D6, DATELINE (Design and Application of a Travel Survey for European Long-Distance Trips, based on an International Network of Expertise), 5\textsuperscript{th} Framework Programme, European Commission –DG Tren.

DATELINE Consortium (2003); *Weighting and Grossing Up Report*, DATELINE (Design and Application of a Travel Survey for European Long-Distance Trips, based on an International Network of Expertise), 5\textsuperscript{th} Framework Programme, European Commission –DG Tren.


Neumann, A. (2003); *Korrekturverfahren für Stichproben von Verkehrsverhaltenserhebungen der Personenfernverkehrs (Methods of data weighting for samples of long distance travel surveys)*, Dissertation am Institut für Verkehrswesen, Universität für Bodenkultur, Wien.

Pendyala, R.M. (2003); *Standards for Archived and Documentation of Travel Survey Data: Challenges and Opportunities*, Resource paper, Workshop 1-6, 10\textsuperscript{th} International Conference on Travel Behaviour Research, Lucerne, Switzerland, August 10\textsuperscript{th} – 15\textsuperscript{th}, 2003.


Sammer, G. (1997); *Problems and Solutions in Urban Travel Survey*; Urban Travel Survey Methods: Measuring the Presence, Simulating the Future; Editors: Bonnel, P.; Chaplean, R.; Lee-Gosselin, M.; Raux, Ch.; Les Chemins de la Recherche; Number 42; Programmes Rhone – Alpes; p. 145-159.


Stenurt, G.N. (1997); *Transportation Tomorrow Survey in Toronto: Special Features and DATA Dissemination Program*; Urban Travel Survey Methods: Measuring the Presence, Simulating the Future; Editors: Bonnel, P.; Chaplean, R.; Lee-Gosselin, M.; Raux, Ch.; Les Chemins de la Recherche; Number 42; Programmes Rhone – Alpes; pp. 93-108.


Stopher, P.R., Wilmot, Ch.G., Stecher, Ch., Alsnih, R. (2003); *Standards for Household Travel Surveys – Some Proposal Ideas*, Paper presented at the IATBR Conference “Moving
through net: The physical and social dimensions of travel”, 10th International Conference on Travel Behaviour Research; 10th till 15th of August 2003; Lucerne, Switzerland.


Wilson, A.G. (1974); Entropy in Urban and Regional Modelling, Pion, London.

http://www.icpsr.umich.edu/DDI/, 2004/03/06
http://cgi.uni-mb.si/elmis/, 2004/03/25
http://www.nesstar.com/, 2004/03/27

7. Abbreviations

AET: Association of European Transport
CAPI: Computer Aided Personal Interview
CATI: Computer Aided Telephone Interview
DATELINE: Design and Application of a Travel Survey for European Long-distance Trips Based on an International Network of Expertise
DDI: The Data Documentation Initiative
DIF: Data Interchange Format
ELMIS: European Long-Distance Mobility Information System
IATBR: International Association of Travel Behaviour Research
IATUR: International Association for Time Use Research
8. Annex


(1) Document Description
Title, abstract, author (of the survey), producers (of documentation), date (of documentation), copyright (of documentation), key words (topic classification), distributor

(2) Study Description

(2.1) Overview information
Title (of the study), objectives, definition of key terms of the survey (journey, trip, trip stage, mode, purpose, activities, etc.), abbreviations/glossary

(2.2) Administrative details of the survey
Funding agency (addresses), producers of survey and their responsibilities (survey design, data collection, data processing, data analysis), advisory committee, survey/study costs and budget, dates and duration of survey, project management, contact person(s) (addresses), confidentiality

(2.3) Scope of survey
Definition of target population, survey units, analysis units, time period covered, geographical coverage

(2.4) Survey design
Description of survey method and concept (reasons of selection, time frame), pilot-survey, pre-tests, validation and non-response survey, frequency of data collection (e.g. monthly, quarterly, yearly) critical feedback considerations concerning the experiences with the survey design

(2.5) Sample design
Sampling units, sampling frame (type, performance in terms of accuracy, completeness, duplication, up-to-dateness, representativity) sampling method, sampling size (mathematical definition), sampling stratification, data quality and sampling loss, conduct of sampling, problems of sampling, critical feedback, considerations regarding the sampling experiences

(2.6) Survey instrument design
Questions content, format and order, physical nature of survey instrument, documentation of forms and questionnaire, reasons for selected questions, critical feedback analysis and considerations regarding the experiences with the survey instrument

(2.7) Survey implementation and execution
Survey procedure, interviewer selection, training and supervision, non-response treatment, documentation of field-work, protocol of the survey (number of contacts, contact characteristics (date, time, interviewer ID), contact success,
proxy reporting, critical reflection of the survey execution, problems of the survey execution, software used for the survey execution

(2.8) Data processing
Database building, questionnaire editing, coding methods (geocoding), code documentation, data cleaning, correction (imputation), definition of a “usable return”, weighting and expansion procedure, software used in data processing, weighting statistics, critical feedback analysis of the data processing, problems of data processing

(2.9) Data analysis
Data quality statistics (with standardised indicators (Stopher et.al. 2003, Stopher and Jones 2003) e.g. missing values, proxy reporting rate, data cleaning statistics, data validation statistics, sampling error, coverage error, response rate), statistical overview of results about the main key indicators (rate of non-mobile households and persons, average activity and trip rate per person, average distance travelled per person, modal split of trips, purpose split of trips, etc.)

(3) Data File Description
File name, file structure and dimension, data format, software used to produce the file, unit of variables, number of cases, number of variables, record length

(4) Variable Description
For each variable of the data set the following information is requested: name and definition of variable, label, type, code, range of data values, question from survey instrument. The variables are ordered in a hierarchical way: household, persons, journey, trip, trip stage information

(5) Other Documentation
References, questionnaires, other notes


<table>
<thead>
<tr>
<th>Travel Survey</th>
<th>Trip Survey</th>
<th>Trip diary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Long-distance Survey</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Activity Survey</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Time Use Survey</td>
</tr>
<tr>
<td>Stated Response Survey</td>
<td>Stated Preference</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stated Tolerance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stated Adoption</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stated Prospect</td>
</tr>
<tr>
<td>Mode/vehicle based Surveys</td>
<td>On-board Transit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>On-road Travel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>On-airport</td>
</tr>
<tr>
<td>Passive Tracking Survey</td>
<td>GPS-based</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mobile phone based</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Licence plate based</td>
</tr>
<tr>
<td>Traffic Surveys</td>
<td>Traffic counts</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Parking Survey</td>
<td></td>
</tr>
<tr>
<td>Qualitative Surveys</td>
<td>Focus Groups</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attitude Surveys</td>
<td></td>
</tr>
</tbody>
</table>

8.3. **European Long-distance Mobility Information System (ELMIS), an Example of an Advanced Travel Data Documentation and Archiving System.**

ELMIS is a web-based metadata system for the European long distance travel behaviour data over a stand alone Windows application, which includes the use of NESSTAR light (DATELINE Consortium 2003, http://cgi.fg.uni-mb.si/elmis/). The advantages of this system are:

- A large group of users are able to instantly access and make use of the stored information. Any person who has an Internet connection can enter the site and use the application to browse and analyse results.
- The alternative production and distribution of CD ROMs would take time and create additional costs.
- Users are able to access the stored data from a less powerful computer and are not obliged to install and manage the application. This avoids or at least reduces technical problems left to the user (e.g. installation, environment settings, storage space, conflicting programmes etc.). Another operating system or a different hardware configuration is prevented.
- The testing of the application is much easier as is the detection and solving of technical problems. Interference stemming from the use of another operating system or a different hardware configuration is prevented.
- The use of a fully functional system already in existence, even more important, a system that has been developed within a project supported by the European Commission. If the DATELINE Consortium had developed a separate system, it would have been counter productive to the goals of the EC to use standardised metadata systems.
- Even though the NESSTAR licence is currently limited to three years, the system’s architecture provides for a transfer of the database to another NESSTAR server once the licence expires. NESSTAR is already developing an important provider that will allow disseminating the survey results.
- The future of a web-based system is more certain (stability). A stand-alone application is dependent on developments in the computer industry, which tends to introduce short-term innovations and changes.
- The format of the data supported by NESSTAR is based on the metadata description format (DDI), which follows international standards. This makes the database more interesting to the user community, because it allows for comparisons with other databases for many years to come.

The hierarchy presented in Figure 7 shows the contents and structure of the ELMIS web site organised by web pages. The contents are organised in three levels of detail, all of which are integrated as main parts of the home page. ELMIS is a web site supported by applications that deliver the DATELINE survey results in a highly interactive way. To allow ELMIS users to browse through the database and perform statistical analyses, the system integrates the NESSTAR server application, which contains a statistical engine. When using ELMIS, the user interacts with the NESSTAR statistical engine through the NESSTAR Light Client.
(NCL). For the purposes of ELMIS, the client has been adjusted slightly, so that it differs in part from the standard client normally used.

Figure 7: Structure of the ELMIS web site

![ELMIS web site structure diagram](image)

One of the key advantages of the NESSTAR system is the combination of metadata and the data browser. In NESSTAR, the user can access the data itself and perform statistical analyses, but at the same time the user can also reach crucial supplementary metadata such as text descriptions of variables and tables. The DDI scheme is laid out in a hierarchy and contains the following sections:

- **Study Description**, which contains information about the study and the collection of data (contents, collection methods, processing, sources, access conditions etc).
- **File Description**, which describes each file of the data collection (formats, dimensions, processing information, information about missing data etc.)
- **Variable Description**, which describes each variable in a data file (format, variable and value labels, definitions etc., see Figure 8)
- **Other Study-Related Material**, which can include references of reports and publications, other machine-readable documentation that is relevant to users of the study etc.
Figure 8: Example of a variable description

Online tabulations and analysis
Four kinds of statistical analysis may be carried out using the database. These are: tabulation, scatter plot, descriptive statistics and regression. To add variables to an analysis, one must click on the variable in the list of variables and press the ADD button in the left-hand panel. Results of the analysis presented in a table can also be displayed in chart form (Figure 9).

In NESSTAR, it is not necessary to run an analysis with the complete database. Sometimes it is interesting to analyse only parts of the dataset. For example any national analysis for the DATELINE project can be performed with data from only one country or a selection of just a few countries. In this case, the SUBSET folder in the right-hand panel must be used, which loads into the same panel a set of options that help define the subset (Figure 10). With this tool, one can select variables, operators and categories of variables to construct conditions that need to be satisfied before adding particular cases to the subset. It is possible to define several conditions linked with the two logical operators AND and OR.

Download
The data can be exported in several formats compatible with common statistical analysis tools (Figure 11). Through the NESSTAR Light Client, ELMIS offers the following formats for export: SPSS system file, SPSS portable file, NSDstat, Statistica, Stata, Data Interchange Format (DIF) (suitable for use in Excel), Dbase 3 and SAS. It is also worth knowing that ELMIS provides the complete DATELINE database in ASCII format for downloading. The following files are available: households, household state, persons, journeys, participants, trips, excursions and commuting.
Figure 9: Example of results displayed in a graph

![Graph Example](image1.png)

Figure 10: Defining rules for a subset definition of the database

![Subset Definition](image2.png)
O-D matrices

Since the *DATELINE* project is concerned with long-distance mobility, one of the key project reports is about O-D matrices. For the purpose of the project, it was decided that for all 15 EU countries, the O-D matrices should be constructed following the regional differentiation system of NUTS, i.e., NUTS1. The user is able to explore journeys undertaken by plain, train, car or by some other mode of transport. Naturally, matrices showing all modes of transport are also available. Through ELMIS, the user can view the O-D matrices in tabular or graphical form. In both cases, origins by regions and destinations can be selected from respective lists offered through the matrix page. Data from the matrix can be retrieved by clicking either the TABLE or the MAP button. The table can be saved as a comma delimited ASCII file or as an MS Excel file. The graphical presentation of the matrix includes the drawing of desire lines on a map.

Figure 11: Data download page