DEPARTMENT OF TRANSPORT AND COMMUNICATIONS
FEDERAL OFFICE OF ROAD SAFETY
DOCUMENT RETRIEVAL INFORMATION

Report No.	Date	Pages	ISBN	ISSN
CR68	August 1988	52	0-642-51123-3	OR = 0158-3077 CR = 0810-770X

Title and Subtitle

REVIEW OF THE AUSTRALIAN DAY-TO-DAY TRAVEL SURVEY DATA BASE 1985/86

Author(s)

Performing Organisation (Name and Address)
INSTAT Australia Pty Ltd

Sponsor (Name and Address)
Federal Office of Road Safety
GPO Box 594
CANBERRA ACT 2601

Available from (Name and Address)
Federal Office of Road Safety
GPO Box 594
CANBERRA ACT 2601

Abstract

This report evaluates the quality of the risk exposure data base derived from the Survey of Day-to-Day Travel in Australia 1985-86. Repunching a 1.6% sample of the survey forms and comparing with the original data sets revealed a less than 1% error rate. Car licence and car driver information had an acceptable error rate but was less accurate than the main body of data. Links between Files on the IP Sharp system had been implemented correctly. Three sets of weight are defined for the data base. The first adjusts for non-response within the sample, the second extrapolates to the sampled statistical areas and the third extrapolates to the target population of all Australia, benchmarked to 1981.

Keywords

NOTES:

(1) FORS Research reports are disseminated in the interests of information exchange.
(2) The views expressed are those of the author(s) and do not necessarily represent those of the Commonwealth Government.
(3) The Federal Office of Road Safety publishes two series of research report
    (a) reports generated as a result of research done within the FORS are published in the OR series;
    (b) reports of research conducted by other organisations on behalf of the FORS are published in the CR series.

STATISTICAL DATA ANALYSTS

Review of exposure data base from the Survey of Day-to-Day Travel in Australia 1985-86

INTSTAT Australia Pty Ltd

August 1987

Summary

This report evaluates the quality of the risk exposure data base derived from the Survey of Day-to-Day Travel in Australia 1985-86. Quality was assessed by repunching a 1.6% sample of the survey forms and comparing the repunched and original data sets (section 2), by checking the consistency of the links between the data files in the data base and by performing range and logical checks on the data fields in the data base (section 3). Weights for extrapolating the survey results to the whole of the Australian population were also derived (section 4). Lastly, the report provides additional system documentation for the data base (section 5)

Analysis of the repunched sample showed that keypunch errors in the data base are less than 1%, except for the car licence field in the person file and for the car driver field, the trip start time field, the trip duration field and the trip distance field in the trip file. These error rates are acceptable, but suggest that care will need to be taken when interpreting the trip times and distances. In addition, the car licence and car driver information is less accurately recorded than the data for other fields.

The data are stored as four hierarchical files on the IP Sharp system (Households-Persons-Trips-Trip modes). The link between each file and the files adjacent to it in the hierarchy was checked. Links were found to be implemented correctly and there were no records which did not link.

The values recorded for each variable were examined for out-of-range values. Other checks concerned the consistency of values between data fields, including those on different files. After correcting the documentation of the code frame, most values were found to be valid, although some inconsistencies were found. For example, trip day sometimes differed between the persons and households files.

Three sets of weights are defined for the data base. The first adjusts for non-response within the sample, the second extrapolates to the sampled statistical areas and the third extrapolates to the target population of all Australia. The sampled statistical areas includes all capital cities, but only selected Statistical Districts and LGAs from elsewhere.

For persons, the first set of weights adjusts for the lower response rate of non-travellers compared with travellers. The second set of weights benchmarks to the 1981 Census population, classified by sex and age group (9-14, 15-29, 30-49, 50-64 and 65+), in each of the sampled areas, while the third set of weights benchmarks to the 1981 population of Australia, classified by sex, age group, State or Territory and capital city/rest of State.

Most users will wish to use the third set of weights to estimate quantities from the survey.

Review of exposure data base from the
Survey of Day-to-Day Travel in Australia 1985-86

CONTENTS

	Summary	2	
1.	Objectives	4	
2.	Repunched data	6	
	2.1 The repunched sample		6
	2.2 Comparison with the data base		9
3.	Evaluation of the database	16	
	3.1 Difficulties in checking the data base		16
	3.2 Consistency of links		17
	3.3 Consistency checks for other fields		21
4.	Weighting	28	
	4.1 Introduction		28
	4.2 Weights within the sample		32
	4.3 Weighting to the sampled population		33
	4.4 Weighting to the target population		42
	4.5 Summary		44
5.	System documentation	45	
	5.1 Location of the data		45
	5.2 MABRA field descriptions		45
	5.3 Code frame changes		49
	5.4 Linking files		49
	5.5 Identifying the Victorian sample selected from SEC records		50
	5.6 Changes made to the data base		50
	5.7 Accessing the IP Sharp system		51
	5.8 Logging onto the IP Sharp system and accessing the data base		52

1. Objectives

The aim of this report is to evaluate the quality of the risk exposure data base derived from the Survey of Day-to-Day Travel in Australia 1985-86.

Accuracy of the data base

Ideally the data base would accurately record details of all sampled households, of all persons within these households and of all trips of these persons for their travel day. However, it is possible for errors to occur at each stage of information transfer. Firstly, respondents must accurately record the characteristics of their household, themselves and their trips on the survey form. Secondly, this information must be correctly coded before data entry, and, thirdly, the coded information must be correctly punched. Lastly, subsequent processing of the punched data, and its transfer between computer systems, must preserve the accuracy of the data.

Information transfer from respondents to survey forms. Information reported by respondents may be present and correct, present and incorrect or omitted. Examples of present but incorrect data include ages incorrectly given, the wrong box ticked, or trips reported but not actually undertaken. Causes of incorrect data are varied, and include inadvertent and intentional misreporting and misunderstanding of what was required. For example, trips within a rural property might be regarded by the respondent as a trip, but such trips are not relevant for calculating risk exposure on public roads. Omitted information is a more likely circumstance, with respondents likely to miss items on the form or to forget to record trips. Temporarily absent household members might not be recorded as being in the household. Non-responding households are an extreme example of omitted information.

In some cases omitted or inaccurate information can be detected from the returned form. Where this occurred, and where the responding household gave a telephone number, Socialdata checked this information with the household. Socialdata also conducted reinterviews of about 300 households (see Section 2.2.2 of their report).

The survey form was designed by Socialdata to minimise these problems, and was tested in the field. Specific attention was paid to encouraging response (for example, by motivating respondents with a letter from the Minister for Transport) and allowing them access to further details of the survey (for example, via an 008 number).

The work for this report has not specifically addressed the transfer of information from respondents to the survey forms, and the survey forms were accepted as the best information available on the sampled households, persons and trips. In particular, no survey respondent was recontacted in the course of this evaluation, as this would have been beyond the scope of the report.

Coding of information on the forms. Rigorous assessment of the coding was beyond the scope of this report. Since the survey form was designed for ease of use by the respondents, and this included not having coding or punching information which might confuse respondents, coding of the forms was essential for accurate data entry. However, in the course of analysing the repunch sample (Section 2.1), it was apparent that not all forms were coded before data entry. Therefore, some forms must have been coded while their data was being entered. Moreover, because the original coder annotates the survey forms, it is not possible to check the coding by recoding the forms without knowledge of the original coder's codes and this means that unbiased coding error rates cannot be determined.

Transfer from coded survey forms to magnetic media. This was assessed by repunching a sample of survey forms. This is described in Section 2.1. The repunched data were compared with the corresponding data on the data base (Section 2.2).

Subsequent processing and transfer of information. Section 3 assesses the accuracy of data in the data base. The household and person records supplied by Socialdata have also been read and used by the consultant on another computer system, and the data bases on IP Sharp's and the consultant's computers gave identical results for all tables that were run on both systems.

Weighting

Socialdata have provided internal weights to adjust for differential response of households and persons. To enable the results of the survey to be extrapolated to a wider population, two sets of external weights are derived in Section 4. Both sets of weights use the 1981 Population Census. The first set benchmarks to only the sampled population, while the second set extrapolates to all Australia. These weights differ outside the capital cities.

2. Repunched data

2.1 The repunched sample

A sample of household survey forms was selected for repunching. **The aim of** repunching the sample was to assess error rates and to enable bounds for the accuracy of the data, as entered from the forms, to be calculated. It allowed the consultant to examine the quality of the survey forms in a structured way. Note that repunching of the data on the forms is not relevant to assessing either the quality of the data reported by respondents or the quality of the coding and checking of the data that preceded the data entry. **A secondary aim** was to determine whether a repunch of the whole survey would be feasible should the error rates prove unsatisfactory.

For a survey of this size, and for self-completion questionnaires, it is inevitable that there will be some errors in the punching of the data. **For** this survey, error rates less than 1% for the non-link fields have been considered as satisfactory error rates. **For the fields used for linking** between households, persons, trips and trip modes, complete consistency is essential.

The sample was chosen to be representative of the forms and processing in each State and Territory and throughout the survey. This was to allow monitoring of the quality of the survey data for each of the processing centres overall and over time. **Accordingly, the sample was selected by specifying an** approximate quota of households per State or Territory for each month of the survey. **For ease of sampling, and to enable checking that all records** pertaining to a given household were correctly entered, all persons within a household and all trips for these persons were also selected.

The number of households in the sample was chosen so that error rates in excess of 1% for individual fields could be detected for households in the whole sample and for persons and trips within the larger States. **The sample** size was about 1.6% of households in the survey.

The sample was drawn by Socialdata from the forms held for the whole survey.

	Number of households		
State or Territory	Specified resample (12 months)	Actual resample (12 months)	Total number in survey
ACT	12	12	482
New South Wales	60	55	4330
Victoria	60	61	4905
Queensland	36	36	2913
South Australia	36	36	2183
Western Australia	36	33	2358
Tasmania	24	27	583
Northern Territory	24	24	464
	288	284	18218

The consultant drew up a data entry specification for the forms, and the data were entered by an independent data processing bureau, South Side Data Processing Pty Ltd. Southside wrote its own data entry program. This program specified field sizes, that each field was to be numeric or blank, and that the household key would be repeated for the records within each household. The program did not use the number of persons or number of trips information (which would ensure that the correct number of persons and trips records were entered), and it did not include range checks for individual fields. The program did not check for consistency of the dates given on different forms, nor did it impute values for omitted data where this was sometimes feasible given the other data on the forms.

The data were entered once only. This would be expected to give much higher error rates than if the data were entered twice, as is common practice.

The repunched data consisted of records for 284 households, which had a total of 682 persons who made 2405 trips with 2551 trip components. Several consistency checks were performed on the repunched data:

Linking. Checks within the repunched data showed 3 errors in the household key (1.1% per field or 0.15% per punched column). There were also 2 omitted person records (0.29%), and 17 omitted trip records (0.71%). One reason for the high rate of omitted records was unclear coding of the forms.

Dates. Each missing or out of range date was checked against the original forms, and corrected where necessary. The correspondence of the travel dates between the household, person and trip forms was also checked. The birth year and age of each person were checked for consistency.

Inconsistent trip times. The start and finish times for each trip were checked for errors such as the start time following the finish time or preceding the finish time of the previous trip. These errors were resolved, where possible, from the original forms. A common problem was that respondents did not always specify whether their trip started or finished in the morning or afternoon.

Trips with a missing distance measure. Distance measures were commonly omitted by respondents, but could sometimes be rectified by reference to a trip in the reverse direction or the same trip for another member of the same household.

Persons with more than fifteen trips. Two cases arose from keying errors, while the other 5 were legitimate.

Trips lasting more than 2 hours. Of these trips, 20 were correct. The remainder resulted from errors in specifying the time of day (am/pm); most of these were coding errors on the original forms.

Out-of-range trip times. Trips with start times before 0400 or after 2359 were checked. Most of the early start times were correct, but some incorrectly resulted from incorrect specification of the time of day (am/pm).

Missing or out-of-range trip purposes. There were 2 trips with the trip purpose correctly specified as missing. The remaining 50 were keying errors.

Trip modes. There may have been problems with the coding of the trip mode. For example, 2 trips by 'utility', 6 trips by 'truck' and 2 trips by 'walk' had not been coded and so were not keyed. Some trip modes were not punched as '0' or '1'.

Extreme distances. Trips keyed as being less than 1 km or more than 100 km were checked. Most of the 190 short trips were correct; the remaining errors were keying errors, often because fractions were not keyed. Of the long trip distances, 21 were correct, while the remaining 40 were incorrect because the wrong trip measure was keyed or recorded on the form.

Households with more than 5 of any type of vehicle. This revealed 2 keying errors and 1 household correctly with 6 bicycles.

Out-of-range values on the person file. Two keying errors were corrected.

Errors detected in the repunched data were corrected before checking against the data base. The repunched data was also reformatted so that it resembled the original data set that Socialdata provided.

This preliminary examination of the repunched data showed:

(1) The time of day (am/pm) and trip distance measure fields were, as could have been anticipated, poorly completed. Unfortunately, these fields are critical for most uses of the data base.

(2) Coding of the forms is inadequate for accurate repunching of the data, and any recoding would be expensive and subject to bias. The adequacy of the coding for the custom-made data entry program is assessed in section 2.2.

(3) Any future repunch of the data should include verification, range checks and as many logical checks as possible. Certainly, a naive repunch (as was done for the repunch reported here) would not transfer data from the forms with adequate accuracy.

2.2 Comparison with the data base

2.2.1 Methods of checking the repunched sample with the data base

Two possible methods for checking the repunched sample data against the original were considered. These were:

(1) Directly comparing the raw data file produced by the repunching with the appropriate records in the raw data files produced by Socialdata, using micro-computer software.

This method was rejected as unwieldy, since selecting the appropriate records from the 50 diskettes of raw data produced by Socialdata would be exceedingly time-consuming and difficult, and downloading the selected cases from the IP Sharp system would require an investigation of software communications protocols, which was also considered too time-consuming.

(2) Loading the repunched data onto the IP Sharp system and compare the data sets using MABRA.

This method was chosen because it was simpler. It was also considered more desirable because the data being checked was a direct subset of that used for analysis.

2.2.2 Loading the data onto the IP Sharp system

The repunched sample of 284 households was entered onto the IP Sharp system for comparison with the original data by Rino Ciaccia of IP Sharp. Four files which were identical in structure to the four files containing the original data had been created (see Section 2.1). These files were:

MABRA system	Description	Number of records	Checked in this section
60	Sample of household file	284	Yes
61	Sample of persons file	682	Yes
62	Sample of trip file	2405	Yes
63	Sample of Tripmode file	2551	No

Only the first three files were checked. The fourth file (Tripmode) was not checked since all information (except for mode duration and distance of multimodal trips) is identical to information in the Trip file, and can therefore be checked in the original survey file by comparison with the Trip survey file. Since the method used by Socialdata for imputing duration and distance for multimodal trips was too complex to replicate on the sample data, this information could not be checked. The files are currently available on the IP Sharp system for further analysis if required. A list of the MABRA fields for each survey file is attached. All checking was done using MABRA as it did not require the setting up of variables before analysis (cf XTABS). MABRA also allows more flexible handling of the data and has the ability to list records with ease.

2.2.3 Strategy for comparing fields

The comparison of the repunched sample with the original data was made by directly comparing fields. The strategy for each of the three survey files was as follows:

(1) Copy the fields from the original survey file to the matching repunched survey file (using the LINK SYSTEMS command) and save them. The copied fields were given the same names as the original fields, prefixed by the letter M, to indicate they came from the main file, for example, NOP became MNOP.

(2) Create fields which contained the difference between the field from the repunched sample file and the original data. These difference fields were given the same names as the original fields prefixed by the letter D, for example, DNOP=NOP-MNOP.

(3) List all records where the difference field was non-zero (for categorical data) or a specified tolerance limit (for continuous fields with large ranges). This was done for each field individually.

(4) Check the list of records with differences against the original questionnaires and record which of the original or repunched was in error. Listings are in the supporting documentation.

Technical notes:

(a) In linking the systems, the sample file was used as the host system and the original file as the linked system. The advantages of this were:

- working on a smaller survey file, so that it was faster and less likely to encounter the 'Workspace full' error message, and

- not creating survey files twice their original size (and therefore unworkable) and saving data that would eventually have to be deleted.

(b) Step (2) was necessary because it is not possible to directly compare two fields using a non-APL terminal, since the hydrant operator necessary (see p.118 of the MABRA manual) is not available for such a terminal. On an APL terminal, it would have been possible to omit step (2) and simply use a constraint on the LIST RECORDS command to select those cases where a difference occurred, for example, NOP NE MNOP.

2.2.4 Errors found

(1) Two households present in the repunched sample had been omitted from the original survey file. These were households #8013030 and #8019021, which accounted for 4 persons and 13 trips.

Household	Persons	Trips
8013030	2	9 (4+5)
8019021	2	4 (2+2)

These households are recorded as nonresponding households, with no data, in the original survey file. There did not appear to be any reason why these households were omitted. It is possible that they are in the data base but with an incorrect household number, but this cannot be checked.

(2) A summary of variables checked and their error rates appear in Table 2.2.1. The number of errors due to the omitted households are recorded separately from those due to keypunch errors. Errors in both the original (data base) and repunched data are shown.

Table 2.2.1 Summary of errors detected in the repunched sample and in the data base. All repunch errors are in keypunching, while the data base errors are classified as being due to either keypunching or omission of the complete record.

Households (284 records)

		----- Place where errors were detected -----						
		Total	Repunch (Keypunch)		Original data (Keypunch)		(Omitted)	
Field Description	MABRA Name	#	#	%	#	%	#	%
N of people	NOP	4	2	.7	0	.0	2	.7
N aged 5-8	NO5TO8	4	1	.4	1	.4	2	.7
N aged 9+	NO9O	2	0	.0	0	.0	2	.7
Telephone	TEL	5	1	.4	2	.7	2	.7
N of bicycles	NOBY	4	1	.4	1	.4	2	.7
N of Motorbks	NOMBY	3	1	.4	0	.0	2	.7
N of cars	NOCARS	6	4	1.4	0	.0	2	.7
N other vehicles	NOOTHER	6	4	1.4	0	.0	2	.7

Table 2.2.1 (continued)
Summary of errors detected in the repunched sample and in the data base. All repunch errors are in keypunching, while the data base errors are classified as being due to either keypunching or omission of the complete record.

Persons (682 records)

			----- Place where errors were detected -----					
		Total	Repunch (Keypunch)		Original data (Keypunch)		(Omitted)	
Field Description	MABRA Name	#	#	%	#	%	#	%
Year of birth	YOB	36	31	4.5	1	.1	4	.6
Age in years	HOWOLD	20	15	2.2	1	.1	4	.6
Sex	SEX	9	3	.4	2	.3	4	.6
Country of birth	COB	20	13	1.9	3	.4	4	.6
Education	EDUC	30	24*	3.5	3*	.4	4	.6
Home duties	EMP1	3	3	.4	0	.0	0	.0
Looking for work	EMP2	3	2	.3	1	.1	0	.0
Retired,age pens	EMP3	6	6	.9	0^	.0	0	.0
Other, not empl	EMP4	9	6	.9	3^	.4	0	.0
F/t study	EMP5	13	7	1.0	6	.9	0	.0
P/t study	EMP6	8	6	.9	2^	.3	0	.0
P/t,casual wrk	EMP7	10	9	1.3	1^	.1	0	.0
F/t work	EMP8	21	15	2.2	3^	.4	3	.4
Car licence	CARLIC	31	18	2.6	9	1.3	4	.6
Truck licence	TRUCKLIC	7	6	.9	0	.0	1	.1
Motor bike licnc	MBYLIC	3	2	.3	1	.1	0	.0
Number of trips	NOTRIPS	7	0	.0	3	.4	4	.6

* Error in both repunch and original, therefore columns add to more than total

^ Assuming 0, -1, blank are identical

Table 2.2.1 (continued)
Summary of errors detected in the repunched sample and in the data base. **All repunch errors are in keypunching, while the** data base errors are classified as being due to either keypunching or omission of the complete record.

Trips (2405 records)

Field Description	MABRA Name	Total #	Place where errors were detected: Repunch (Keypunch) #	%	Original data (Keypunch) #	%	Original data (Omitted) #	%
Walk	TRPM1	27	21	.9	6	.2	2	.1
Bicycle	TRPM2	21	20	.8	1	.0	0	.0
Bus	TRPM3	17	14	.6	1	.0	2	.1
Train	TRPM4	20	16	.7	4	.2	0	.0
Tram	TRPM5	13	6	.2	7	.3	0	.0
Taxi	TRPM6	9	4	.2	5	.2	0	.0
Ferry	TRPM7	2	2	.1	0	.0	0	.0
Motorbike	TRPM8	2	1	.0	1	.0	0	.0
Car driver	TRPM9	58	25	1.0	34	1.4	7	.3
Car passenger	TRPM10	40	25	1.0	11	.5	4	.2
Truck	TRPM11	6	0	.0	6	.2	0	.0
Trip start time	TRPST	127	45*	1.9	70*	2.9	13	.5
Trip duration +	TRPDUR	57	19	.8	26	1.1	13	.5
Trip purpose	TRPPUR	67	31*	1.3	15*	.6	13	.5
Trip distance -	TRPDIST	83	24*	1.0	51*	2.1	12	.5

+ Only differences in duration of over 10 minutes were examined. Of the 26 keypunch errors in the original data, 6 (0.2%) were due to confusion over AM/PM.

- Only differences in distances over 1 km were examined. Of the 51 errors in the original data, 30 (1.5%) were due to conversion errors for the distance measure.

* Error in both repunch and original, therefore columns add to more than total.

(3) Fields containing travel days were not checked as the date and not the day number was recorded on the questionnaire. However, records from the household file for non-matching travel day (TRDAYA) and QDAY (day returned) were listed. There were 5 differences, 2 due to omitted cases and the rest were missing from the repunch.

2.2.5 Sources of error

The following sources of error were identified:

(1) Poor quality coding. During the checking, a variety of coding errors leading to incorrect keypunching were found. These included:

- Failure to code multiple response codes (EMP1 to EMP8, TRPM1 to TRPM10) where boxes were ticked, so that data entry operators were forced to code the data while entering it. This lead to a greater number of errors at the end of these lists.

- Failure to code responses written in for the above fields (for example, failing to code 'Truck' as tripmode 11), so that operators either omitted these responses or guessed. For example, in one instance, 'Truck' was recorded as 'Car driver' by the data entry operator.

- Incorrect coding. For example, 'male' was coded as '2' in one instance.

(2) Coders modifying the responses in an unclear manner. This happened most often in the trip file. Coders would indicate (often in a faint pencil or illegibly in pen) that a particular trip should be recorded as two trips (without indicating which data belonged with which trip) or that two trips should be recorded as one, without indicating what should be entered. Sometimes the following trips would be renumbered by the coders, sometimes not. In some cases, the person number on the trip forms had been modified in an ambiguous way.

(3) Respondents marking the form in an unclear manner. Sometimes respondents marked the form in an unclear manner, for example one respondent wrote '?7' in the trip distance box. One data entry operator had entered '7', the other '77'. These ambiguities had not been resolved by the coder.

(4) AM/PM conversion and recording problems. Problems occurred in both the recording and conversion of AM and PM for trip start and duration. Sometimes respondents crossed out the box with the inappropriate time (for example, crossing out PM to indicate the time was AM), but the data entry operators entered the marked response.

(5) Distance measurement conversion. Kilometre conversion was the conversion most often carried out wrongly. The boxes were extremely small and it was sometimes not clear which box had been marked.

(6) Turning trip form sheets. In some cases, the trip forms had been turned in such a way that the first page of one person was entered with the second page of another. In some cases the trip forms were stapled together, leading to ambiguity about what should be entered.

2.2.6 Conclusion

Keypunch errors in the data base are less than 1%, except for the car licence field in the person file and for the car driver field, the trip start time field, the trip duration field and the trip distance field in the trip file. The reasons for the high error rate in these fields on the trip file have already been discussed in Sections 2.1 and 2.2.5. A possible reason for errors in the car licence field is that respondents without licences may be embarrassed to admit this, while those with licences may regard the question as not worth answering, especially if they have trips for which they are the car driver. Similarly, some car drivers might neglect to indicate this on later trips for the travel day.

Although it is disturbing that two households in the repunch sample were apparently omitted from the data base, it is encouraging that all households had the stated number of members and that all stated trips were present in the data base (see also Section 3).

The conclusions from Sections 2.1 and 2.2 are similar. These checks were done by different persons, using different computer equipment and methods. Note that the error rates for the repunch sample given in this section are after the extensive error checking described in Section 2.1.

These error rates are acceptable, but suggest that care will need to be taken when interpreting the trip times and distances. In addition, the car licence and car driver information is less accurately recorded than the data for other fields.

3. Evaluation of the data base

3.1 Difficulties in checking the data base

There were several aspects of the IP Sharp system which affected the checking process.

MABRA line limit of 80 characters. This meant that it was not possible to enter constraints which passed column 80. Since the constraint prompt takes up some of those eighty characters, it was not possible to enter complex constraints, and so the checking process was somewhat more long winded than it might have been. This constraint also affected the creation of pseudo variables, since a variable that was the sum of several variables had to be broken down into the creation of several preliminary variables so that all variables could be named and still fit on one line.

The use of a non-APL terminal. The IP Sharp system was accessed from a remote site using a modem and an MS-DOS micro-computer using the MITE communications package. This necessitated the use of the NOAPLCHARS command and so the special APL character set was unavailable. Therefore, there are some commands that cannot be used from a NOAPLCHARS terminal. The most important of these was the HYDRANT operator (MABRA manual p.118). Without this operator, it is not possible to directly compare two variable in a constraint (for example, the constraint 'QDAY GT TRDAYA' cannot be used). This means that a difference pseudo variable must be created to make this comparison - a much more time consuming process.

Difficulty in manipulating fields with large ranges and an unknown distribution. Minima, maxima and averages cannot be easily produced. No commands exist in MABRA or XTABS to perform these functions, although a combination of cumbersome file manipulations do make this technically feasible. It is also difficult to determine the ranges appropriate for frequency distributions on such fields, when the distribution, maximum, minimum and average are not known. This limitation made it difficult to work with continuous fields with large ranges (for example, duration, distance, travel day).

Lack of a batch file facility. With a data set of this size, frustrating time delays and errors in specifying commands are inevitable. A batch facility would have meant that commands could have been written and submitted, and the results examined later.

Re-creation of the data base. Because the data base was re-created on the 12th of May at the request of the consultant to distinguish between zeros and blanks (indicating missing values), the checking process had to carried out twice - once on the original data base and again on the re-created data base. Most of the checks documented here were carried out on the data base after it had been re-created. However, because of time constraints, some checks were carried out on only the first version of the data base. These are indicated on the listings. These checks will have been reliable, because all frequency tables from the second version of the data base were compared with those from the first, and the only differences were the breakdown of counts for zeros into zeros and blanks (transformed to -1 for the second version of the data and indicated on listings as "1).

3.2 Consistency of links

The data are stored as four hierarchical survey files on the IP Sharp system. Any two files can be linked by using the LINK SYSTEMS command (LS) and specifying the appropriate key field. The key fields were created by the programmers at IP Sharp. The keys consist of a combination of fields which uniquely identify a record within each particular file. Each file contains its own key and the keys of the files above it in the hierarchy. When linking two files, the appropriate key to use is the one belonging to the file which is higher in the hierarchy. These links have been appropriately defined for efficiently linking the files. Further details can be found in the system documentation.

MABRA system	Type of file	Key field name
40	Households	HKEY
41	Persons	PKEY
42	Trips	TKEY
43	Trip modes	

Links were checked between each file and the file immediately above it and below it in the hierarchy. It was not necessary to check all combinations of links since consistency between adjacent files ensures consistency across all files.

The MABRA LINK SYSTEMS command (LS) was used to link each pair of systems. The supporting documentation contains listings with the MABRA commands and results.

3.2.1 Households file (MABRA system 40)

(1) Linking to persons file (Linked system 41, Key field HKEY)

Household records with no matching person records. 11,135 household records had no matching persons records in the persons files. These are for non-responding households records and all were found to have the number of people equal to zero.

Comparison of number of persons specified in the households file with number of person records in the persons file for each household. A field containing the sum of the number of persons records was linked to the households file. The difference between this field and the 'number of persons 9 and over' field in the households file was found to be zero for all households. Therefore, the number of persons recorded for each household exactly matches the number of persons records in the persons file.

3.2.2 Persons file (MABRA system 41)

(1) Linking to households file (Linked system 40, Key field HKEY)

Persons records with no households. Each person record linked to a household in the households file. There were no person records without matching households.

Person numbers higher than number in household. The households file field for 'number in household' was linked to the persons file and compared with the person number for each person. None were found to be greater.

Comparison of reported travel days in the households and persons files. A field containing the reported travel day (TRADAYA) was copied to the persons file from the households file. 1,287 records were found to differ. Of these, 559 differed by more than one. These inconsistencies need to be resolved.

(2) Linking to trips file (Linked system 42, Key field PKEY)

Person records with no trips in the trips file. 7659 records in the persons file had no matching records in the trips file. These records were for persons who had not travelled that day, and all were found to have number of trips equal to zero.

Comparison of number of trips specified in the persons file with number of trip records for each person. A field containing the sum of the number of trip records was linked to the persons file. The difference between the 'number of trips' field and the number of trip records field was found to be zero in all cases. Therefore, the number of trips recorded for each person exactly matches the number of trip records in the trips file.

3.2.3 Trips file (MABRA system 42)

(1) **Linking to persons file** (Linked system 41, Key field PKEY)

Trip records with no persons. Each trip linked to a person in the persons file. There were no trip records without matching persons.

Trip numbers higher than the number of trips for a person. The persons file field for 'number of trips' was linked to the trips file and compared with the trip number for each trip. None was found to be greater.

(2) Linking to trip modes file (Linked system 43, Key field TKEY)

Trip records with no trip mode records. 148 trip records had no matching trip mode records. A list of these cases indicated that all trip modes for that trip were either blank or zero and therefore should have no trip mode records.

Comparison of number of trip modes recorded in the trips file with number of records in trip modes file for each trip. This check was carried out before the data base was re-created to distinguish between blanks (missing values) and zeros. An exact match was found. The number of trip modes in the trips file was calculated by summing the trip mode fields, since each mode field contained 1 (if the mode was used) or zero (if the mode was not used). It was not possible to rerun this test when the data base was re-created and blanks were converted to '-1'. Summing all trip mode variables would no longer give a count of the number of modes.

Comparison of duration recorded in trips file and total duration of trip mode records for each trip. There were 224 records where a difference occurred. A list of the 72 where the difference was greater than 1 km showed that all differences were due to the total field from the trip mode field containing zero. A second comparison was made, selecting those with a difference not equal to zero and number of records not equal to zero. Seventy seven records were found, all of which contained either blanks or negative trip durations. It therefore appears that, for some trips, information on trip duration from the trips file was not correctly carried across to the trip modes file.

Comparison of the distance record in the trips file and total distance of trip mode records for each trip. Only trips with at least one trip mode record were examined. There were 1625 cases where there was a difference. 1310 cases where the difference was 1 km or more, 654 cases where the difference was 5 km or more, 339 cases where the difference was 10 km or more, 60 cases where the difference was 40 km or more. A listing of a sample of these cases indicated that the trip distance total from the trip modes file was blank or zero. It appears that, for some trips, information on trip distance was not carried across to the trip modes file.

3.2.4 Trip modes file (MABRA system 44)

(1) Link to trips file (Linked system 43, Key field TKEY)

Trip modes with no matching trip records. Each trip mode record linked to a trip record. There were no trip mode records without matching trips.

Trip number in trip modes file matches trip number from trips file. The trip number field from the trips file was copied to the trip modes file. for all trips, the trip number matched.

There were several difficulties encountered when checking the trip modes file. Because of the difficulty of creating a 'number of modes' field in the trips file, it was not possible to check back from the trip modes file to the trips file in many instances. For example, a check of whether the distance and duration of the single mode trips matched across files could not be made because single mode trips could not be identified in the trips file.

The method used by Socialdata (an eight page PASCAL program) to allocate distance and duration to each mode of a multi-mode trip was too complex to check in detail. However, the algorithm frequently results in zero distances travelled by one or more of the modes. It is also somewhat surprising that there are so few multi-mode trips. For example, it would be expected that almost all train trips would be multi-mode trips.

3.2.5 Conclusions

The data are stored as four hierarchical files on the IP Sharp system (Households-Persons-Trips-Trip modes). The link between each file and the files adjacent to it in the hierarchy was checked. Links were found to be implemented correctly and there were no records which did not link.

Checks of the consistency of data between files were also made (for example, the number of trips recorded in the persons file with number of trip records in the trips file). For most checks, consistency was found between files. However, inconsistencies were found between the persons and households files (for trip day) and the trips and trip modes files (in the information about the trip distance and duration).

3.3 Consistency checks for other fields

3.3.1 Methods used

All consistency checks were carried out within MABRA using one or more of the following methods:

(1) **the LIST RECORDS command (LR) with specified constraint.** For example, LR with constraint (HOWOLD LE 16 AND CARLIC EQ 1) lists records for those aged 16 and under who have a car licence.

(2) **the FREQUENCY command (AN) with specified constraint** to examine the number of cases in a particular type of error. **This was** used if the number or range of individual errors might be too large to print.

(3) **the CROSSTAB command (AN) to print a cross-tabulation of two** variables which should match exactly, sometimes with a constraint specified.

(4) the ADD PSEUDO FIELD command (AP) to create a field containing the difference between two specified fields, then use either the LIST RECORD (LR) command with the constraint that the difference field did not equal zero (for categorical variables) or a specified tolerance range (for continuous variables with large ranges), or the AN command (frequency or crosstab analysis) to produce a table of the number of cases in each combination of categories (sometimes with constraints). These pseudo fields were not retained within the data base because of amount of storage that would be consumed.

The methods used, the MABRA commands and the results appear in the supporting documentation.

3.3.2 Consistency checks undertaken - Households file (MABRA system 40)

(1) Range checks on all categorical variables. No out-of-range codes were found. Several apparently out-of-range codes were present, but these were incorrectly documented valid values:

Comment code (COMM) 007 two different travel days reported within household, coders allocated same day

096,097,098 households in NSW whose forms received late because of mail strike; dropped for weighting purposes

Response status (RESP) 09 non-responding household (incorrectly documented as code 02)

The lack of out-of-range codes in a data base of this size indicates excellent editting by Socialdata.

(2) Total number of persons matches sum of number under 5, number 5 to 8 and number over 9. No errors were found using the constraint 'NOP = NOL5 + NO5TO8 + NO90'.

(3) Non-responding households with data. These checks were performed for respondents only (constraint 'RESP NE 1 AND ...'). The following fields were checked:

Check	Constraint	Cases found
Number of people	(... NOP GE 1)	0
Number of cars	(... NOCARS GE 1)	1
Number of bikes	(... NOBY GE 1)	1
Number of motorbikes	(... NOMBY GE 1)	0
Number of other vehicles	(...NOOTHER GE 1)	1
Telephone	(... TEL EQ 1)	0

These errors arose from one household (#410633) which reported 1 bicycle, 7 cars and 1 other vehicle, although the response status was 12 ('forms blank with refusal note') and the comment code was 5 ('forms blank with refusal note'). Presumably this household gave incomplete information. For ease of use of the data base, the vehicle fields should be changed to blanks for household #410633. Even without these changes, most analyses would automatically exclude this household because most analyses would select respondents only.

The numbers of people in each age group (NOL5,NO5TO8,NO90) were not checked explicitly because this test follows logically from tests already done.

(4) 'Responding' households with no data. The only field which must be present is the number of persons in the household, and this was checked using the constraint 'RESP EQ 1 AND NOP LT 1'. A total of 23 such households were found. For two, the comment code was 8 ('no household form returned'). All other fields were blank, except for the postcode, household weight and travel day originally assigned. The travel day reported and the day the questionnaire was returned was present for 17 of these households.

These households have too little data to be useful, and should be regarded as non-respondents. The response status should be changed to 9 ('non-responding household').

(5) Questionnaire returned before reported travel date. The constraint for this, 'QDAY LT TRDAY', detected 3 such households, with the difference in number of days 16, 3 or 1. That only 3 such households were detected indicates good quality keying and editting by Socialdata, and enhances confidence for other fields.

The return date is unlikely to be used in analyses, and so no action is required. If consistency of the data base is desired, the original forms could be checked, or the travel day could be assumed to be correct and the return date changed to blanks to indicate a missing value.

3.3.3 Consistency checks undertaken - Persons file (MABRA system 41)

(1) Range checks on all categorical variables. No out-of-range codes were found. This indicates excellent editting by Socialdata.

(2) Age out of range. There were 7 persons with age 8 or less detected by the constraint 'HOWOLD GE 1 AND HOWOLD LE 8'. One person (household #5502030, person 2) probably has an incorrect year of birth, since this person is looking for work (EMP2=1), is not a student (EMP5,EMP6=0), and has a car licence (CARLIC=1). The other persons appear to be children since all are full-time students (EMP5=1) and have no other work or vehicle licenses. They were all born in 1977, so the approximation inherent in the age calculation (based only on difference between year of birth and year of survey) may account for the recording of an age of 8 rather than 9 years.

This raises the possibility that other ages have been incorrectly assigned. Given that the algorithm is consistent, and cannot lead to errors of more than one year, the current ages in the data base should be retained unchanged. In analyses, the 8 year olds should be included with the 9 year olds. This change could be made in the data base.

The person form for person 2 in household #5502030 should be examined.

(3) Persons too young for a vehicle licence. There were 32 persons aged 16 or less with vehicle licences reported. These were detected using the constraints such as '(HOWOLD GE 1 AND HOWOLD LE 16) AND (CARLIC EQ 1)'. Of these, 26 reported car licences, none reported a truck licence and 7 reported a motorbike licence. One person reported both a car and a motorbike licence. Given the possible errors in the algorithm used to calculate the age of the respondent, and minimum ages for licences in different states, these 32 persons appear to have reasonable data.

(4) Too young for employment category recorded. These were detected using a cross-tabulation of age (for those aged 18 or younger) and each employment category.

Employment category	Number of persons
Home duties, 14 yrs and under	18
Looking for work, 14 and under	1
Retired or age pension, 18 and under	3
Other not employed, 14 and under	22
Part-time study, 14 and under	15
Full-time work, 14 and under	0

Although these young persons have unlikely employment categories, there are very few of them, and it is not necessary to alter the data base.

(5) Too young for level of education recorded. These were detected using a cross-tabulation of age (aged 18 or younger) with education. There were 18 persons who reported university, technical or business college qualifications even though they were aged 14 or younger, and another 18 persons aged 10 or younger who reported being at secondary school. Some of these are unlikely, even after taking into account that the education question asks about the highest level of education attended rather than completed.

It is not necessary to alter the data base for these few persons.

(6) Persons too young for the retirement or age pension. There were 253 persons aged 54 or younger who reported being in receipt of the retirement or age pension. However, these persons may be in receipt of another pension, and so should be retained in the data base.

(7) Incompatible employment categories. These constraints were:

Description	Constraint	Number of persons
Home duties and f/t work	(EMP1,EMP8=1)	141
Looking for work and f/t work	(EMP2,EMP8=1)	16
Other not empl and p/t work	(EMP4,EMP7=1)	64
Other not empl and f/t work	(EMP4,EMP8=1)	113
F/t student and f/t work	(EMP5,EMP8=1)	44
Retired/pension and p/t work (may be legitimate combination)	(EMP3,EMP7=1)	95
Retired/pension and f/t work	(EMP3,EMP8=1)	5

Some of these combinations are unlikely, and some are impossible. It is probably not possible to resolve these. This suggests that variables describing employment categories will need to be carefully constructed so that all persons are classified into the desired groups. No changes to the data base are required.

(8) Comparison of the reason for no trips with the number of trips. A reason for no trips (REASON GE 1) occurred only when the number of trips (NOTRIPS) was zero, and so no errors were detected. This, again, reflects good editting by Socialdata.

(9) Comparison of person's travel day with the reported household travel day. This test was carried out as part of the check of links between files in the Section 3.3. There were 1287 such persons. The person travel day should be used rather than the household travel day for all tabulations.

3.3.4 Consistency checks undertaken - Trips file (MABRA system 42)

(1) Range checks on all categorical variables. The only errors found were two out-of-range codes for the trip purpose field (TRPPUR). These values were 45 and 51, and could be recoded to blank, to indicate a missing trip purpose.

(2) Negative Trip duration. There were 15 trips with a negative trip duration detected by the constraint 'TRPDUR LT -1'. These durations ranged from -4 to -35 minutes, and occurred on the data as originally supplied as well as on the data base. There are few of these trips, and they could be ignored when tabulations are formed. One way to ensure this would be to recode these trip durations to blank, to indicate a missing value.

(3) Negative Trip Distance. The constraint 'TRPDIST LT -1' found no trips with a negative distance.

(4) Trip start time after the end of the day. The constraint 'TRPST GE 2400' found 887 trips, of which 707 started before 2600 and only 16 occurred after 3000. These trips should not be regarded as occurring on the travel day, and so should be discarded.

It is not clear how trips which include the times 0000 and 2400 have been treated. For consistency, all trips which began on the travel day should be included. Trips that continue after 2400 should have the whole of their duration recorded. Trips that began before the travel day should be excluded even if they included part or all of the travel day. Trips that began after the travel day are unequivocably out-of-scope.

3.3.5 Consistency checks undertaken - Tripmodes file (MABRA system 43)

(1) Range checks on all categorical variables. No out-of-range codes were found.

(2) Consistency of multimode trip components with the whole trip. The total distance and duration for one person (person 1 in household #1037024) with a 5-mode trip was checked and found consistent. This check was undertaken before the data was re-created on 12th May. Another trip, with two modes, was found to have zero distance and duration for one mode. Accordingly, this trip for this person shows that the algorithm used to split multimode trips into their components can give rise to extreme allocations.

This suggests that the data on trip component duration and distance be used with care. Given the low reported number of multimode trips, it is also likely that the number of modes for trips has been under-reported. No suggestions are made for a changed algorithm for calculating the duration and distance of trip components.

3.3.6 Conclusion

Range checks were performed on all **categorical variables.** Several apparently out-of-range codes were detected. **All but two were found to** be correct codes which had been incorrectly documented in Socialdata's report.

Other consistency checks were performed within each file, and a number of minor inconsistencies were detected.

In the households file, there were 23 'responding' households with no people, a non-responding household with 9 vehicles, and three households recorded as having returned their questionnaires several days before the travel day they had reported.

In the persons file, there were seven persons calculated to be age 8, 32 people aged 16 or under with motor vehicle licenses and a number of persons for whom their employment category or education level was inconsistent with

In the trips file, there were 15 trips with negative duration and 887 trips where the start time was later than 2400 hours.

In the trip modes file, there are trip modes with zero duration and distance, so care must be taken when using this file.

4. Weighting

4.1 Introduction

4.1.1 Aims of weighting

A sample survey is a survey of a sample of the population of interest. Weighting is often used so that the results from the survey responses can be extrapolated to the population of interest. This process of extrapolation can be regarded as comprising three stages (Figure 4.1).

Achieved sample	(1) ---->	Intended sample	(2) ---->	Sampled population	(3) ---->	Target population

Figure 4.1 Transitions in inference from the achieved sample in a sample survey to the target population.

In an ideal survey, each transition reflects only random sampling. For example, in an ideal unstratified survey, each of these extrapolations would 'undo' the simple random sampling at each step. This would give a single weighting factor at each step, and for the combined transition from achieved sample to target population. An ideal stratified survey results in weighting factors that are constant within a stratum, but which typically differ between strata. Each weighting factor is the inverse of a random sampling proportion.

However, these transitions rarely involve only random selection, and this is reflected in differential sampling probabilities. Differential weights are often employed to counterbalance these differential sampling probabilities, even though their sizes may be difficult to ascertain.

Therefore, weights have two aims: the first to reflect random sampling probabilities and the second to adjust for differential response of different categories of survey units.

4.1.2 Description of the Day-to-Day Travel Survey

The Day-to-Day Travel Survey is more complex than most surveys because of the hierarchy of information that was collected. There are three types of survey units: households in Australia, persons in Australia and trips in Australia. Therefore, the survey can be regarded as consisting of three surveys, one of each of these types of units. Fortunately, the structure of the survey is such that there is a simple relationship between the persons survey and the households survey, and between the trips survey and the persons survey.

For example, conditional on the achieved sample of households, the sample of persons is a complete sample of persons in these households. Therefore, the persons survey can be regarded as a simple random sample (with probability of selection equal to 1.0). The relationship between the persons and trips survey is a little more complex. Conditional on the achieved sample of persons, the sample of trips is a clustered sample of trips of these persons. The sample of trips was a 1 in 365 sample, clustered by day of the year. The days of the year were fixed, and all persons in the same household had the same travel day.

For the household survey, the achieved sample is the set of fully completed household forms, and the intended sample is those forms that would have resulted if all households to whom forms were sent had responded. The sampled population in the household survey is the set of households in the statistical areas from which they were sampled. These households were those enumerated in a complete census of households in the given statistical areas some time before the survey. The target population is all households in Australia on their allocated travel day, had the survey been a complete census. Sections 4.2, 4.3.1 and 4.4.1 deal with appropriate weights for transitions (1), (2) and (3), respectively, in Figure 4.1.

For the persons survey, the achieved sample is the set of fully completed persons forms, while the intended sample is those forms that would have resulted if all persons in all households to whom forms were sent had responded. The sampled population is all persons in the enumerated households in the statistical areas that were sampled. This sampled population is taken on the travel day that would have been allocated had the survey been of all the enumerated households. The target population is all persons, including those not in households, in Australia on their allocated travel day, had the survey been a complete census.

For the trips survey, the achieved sample is the set of fully completed trips forms, and the intended sample is the forms that would have resulted from all trips of all persons of all households sampled. The sampled population is all trips of all persons in the enumerated households in the statistical area on the travel day that would have been allocated had the survey been of all the enumerated households. The target population is all trips in Australia in the year of the survey.

The reason for defining these samples and populations in some detail is that it will make it easier to identify some sources of survey error. In particular, these definitions will be used when discussing non-coverage and differential response.

4.1.3 Types of weights

Weights adjusting for non-response within the sample. These weights apply to transition (1), between the achieved and intended sample, in Figure 4.1. They have been derived by Socialdata, and are discussed in Section 4.2. The aim of these weights is to convert the achieved sample into a simple random sample of the intended sample. Within each stratum, each statistical area in this survey, the sum of these weights is constrained to be exactly equal to the total number of survey responses.

Weights for population estimates. These weights are derived in this report and implement the transitions from the survey responses weighted for non-response (as derived by Socialdata; see previous paragraph) to the intended sample, the sampled population and the target population.

These weights require choosing appropriate benchmark characteristics of the sampled and target populations. Depending on the availability of suitable benchmarks, this may result in a redefinition of these populations.

The first two benchmark characteristics for households were that the weighted number of households classified by season or day of the week for the actual travel day should be evenly distributed. The third benchmark characteristic was that the weighted daily number of households should equal the number of households recorded at the 1981 Population Census. This changes the target population to referring to the 1981 Population Census, and this differs in time and method of enumeration of households from the Day-to-Day Travel Survey. In addition, although this benchmark is straightforward to use for the sampled population, problems of non-coverage mean that the application of this benchmark to the target population is more controversial. These issues are discussed in Sections 4.3 and 4.4.

For persons, the first two benchmark characteristics relate to ensuring an equal representation of weighted persons by season and day of the week for their actual travel day. The third benchmark chosen was the age and sex distribution of persons at the 1981 Population Census. The age groups chosen were ages 9 to 14, 15 to 29, 30 to 49, 50 to 64 and 65 or older.

For trips, there are no suitable benchmarks beyond ensuring that the persons are appropriately weighted. Accordingly, only the person weights were used to weight the trip data adjusted for non-response.

The benchmark characteristics defined using the 1981 Population Census could be redefined using the 1986 Population Census, when these data become available. The age groups used could also be changed. However, the design of the survey ensures that the probability of selection of a person is independent of their age and sex. Therefore, it is expected that the weights for the age groups and sexes will be approximately equal, and that the exact age groups used should not affect any analyses greatly.

4.1.4 Weighting strategies

There are several schools of thought concerning the role of weights in survey analysis. One school holds that fixed weights can be calculated, and these should be applied to the stratum-specific estimates when deriving population estimates, while the other prefers to fit models to the survey data and then applies weights to the fitted values from the models to obtain population estimates.

The fixed weights approach. Estimation using the methods of the first school is usually more straightforward, and the usual fixed weights derived for a sample survey and used as weights in most statistical packages follow this method. However, some problems can arise.

With multiple benchmarks, the usual method of determining the weights in each stratum is to categorise both the achieved sample and the benchmark population according to all of the benchmark characteristics at the same time. The weights are the ratios of the numbers in the benchmark population to the achieved sample for each combination of the benchmark characteristics in each stratum.

However, this method fails when there are no members of the achieved sample in a cell that has members of the benchmark population. One solution to this problem of sparseness is to use fewer benchmark characteristics. Another is to combine 'similar' cells, although this tends to be both subjective and time-consuming. Lastly, it is computationally possible, but cumbersome, to benchmark on each characteristic independently. This last method also can be used when the full cross-classification of the benchmark population is not available. A common approach is to perform only the first iteration of the computational algorithm and assume that the result is sufficiently close to the correct weights.

Sparseness is a problem with the Day-to-Day Travel Survey. For example, the combination of season, day of the week, age group and sex defines 280 cells, but there are only 62 responding households in the Northern Territory outside Darwin.

A further problem is missing data. With fixed weights defined assuming that all of the achieved sample has complete data, any item non-response will cause totals to be underestimated. If this is rectified, it is usually done by assuming random non-response and rescaling all total estimates in a table by a constant factor.

The definition of the strata can also cause controversy. Post-stratification by variables not explicitly stratified by in the design is common practice. In the Day-to-Day Travel Survey, age group and sex can be regarded as post-stratification variables.

The modelling approach. In the modelling approach, models are fitted to the data in the achieved sample without using any weights. It is important that these models take account of the design of the survey, including the benchmark variables, which may be regarded as post-stratification variables. The fitted values from these models, which may be expected counts of units or estimates of totals in the achieved sample, are then weighted.

A typical analysis of a mean would involve the fitting of two models, one of the number of responses used to determine the mean and the other of totals for the variable being analysed. The weighting factors used to transform the first model to the benchmark population are then applied to the second model. Finally, the table of estimated means is produced.

The modelling approach automatically allows for missing data and for independent benchmarking on several characteristics. It is inherently a more flexible technique than the fixed weights approach, although the results from both methods would typically not differ greatly. The main drawback of the modelling approach is that it is a more complex method, and can only be implemented using special software.

Ratio estimation methods can be used with both approaches. Because the modelling approach is not readily available using MABRA on the IP Sharp system, whereas fixed weights can be readily incorporated into MABRA analyses, fixed weights have been derived for this survey. These fixed weights independently benchmark on season, day of the week, age group and sex, although the method of deriving the weights guarantees agreement between the weighted achieved sample and the benchmark population in each geographical stratum for only age group and sex.

4.2 Weights within the sample

These weights were defined by Socialdata. For households, the weights ensure an equal distribution of the weighted households by season of the year. For persons, the weights adjust for higher response rate of persons with trips compared with those without trips, while for trips the weights adjust for the higher average number of trips for respondents than for non-respondents. The person and trip weighting factors result from calculations which infer the number of trips for non-respondents from the number of trips reported by persons depending on how quickly they responded to the survey.

These factors are not greatly different from 1, and their total for any geographical stratum should be equal to the total number of achieved units in that stratum. However, because the weighting factors were recorded to only three decimal places on the data file, these totals did not always agree. This discrepancy has been corrected in the calculations of the weights to the sampled and target populations.

The numbers used in the sample calculations and those in tables summarising the achieved sample differ slightly, but this would result in only minor changes to the weights within the sample. Accordingly, these weights were not recalculated. In any event, the fixed weights calculated in section 4.3 and 4.4 benchmark on an independent characteristic, and any relationships in the original weights are not precisely conserved.

4.3 Weighting to the sampled population

4.3.1 Households

The sampled population for households was the enumerated set of households in each statistical area of the survey. This enumeration was a complete census taken shortly before the survey began. However, the benchmark sampled population is the households in these statistical areas at the 1981 Population Census. Non-coverage errors will arise because the number of households will differ between the censuses; arguably a better benchmark would be from the 1986 Population Census, but these data are not yet available.

Except for three areas (see below), the statistical areas used were areas defined by the Australian Bureau of Statistics (ABS). These were Statistical Divisions for the capital cities and Statistical Districts or Local Government Areas for the other areas. ABS devised a sampling of Collector's Districts (CDs) within all but three (see below) of these areas. Within each area, this sampling was proportional to size, as measured by the number of dwellings at the 1981 Population Census, and so was a self-weighting sample within each area. Note, however, that different areas had different sampling fractions. For example, 1 in 32 households was sampled in the Dubbo LGA while only 1 in 446 households was sampled in the Newcastle Statistical District. Therefore, the total sample was not a self-weighting sample for all the areas sampled, nor for Australia as a whole, and separate weights are required for each area.

Accordingly, the appropriate fixed weights are the ratio of the number of occupied private dwellings at the 1981 Population Census in the given statistical area to the weight to adjust for the observed difference in the achieved sample of household travel days in each season. The latter weights are corrected for the limited precision of these weights on the data file. The sum of the fixed weights for all households in the achieved sample in each statistical area is necessarily equal to the number of occupied private dwellings in the statistical area at the 1981 Population Census.

In the sampling for the survey, the LGA of Malvern, in the Statistical Division of Melbourne, was incorrectly selected as part of the Statistical District of Geelong, and CDs from Malvern were included in the survey. This error was not noticed until the field work had been completed. The data from the LGA of Malvern has been retained in the data set, but is considered as part of Melbourne, not Geelong.

The remaining three areas were rural Victoria, Kalgoorlie and Brewarrina.

For rural Victoria, a random sample of households with electricity accounts was selected. This sample excluded households with addresses in the Statistical Division of Melbourne, the Statistical Districts of Geelong and Bendigo and the LGA of Warrnambool, and therefore covered the rest of Victoria after deleting the areas sampled using the ABS sampling scheme. Given that almost all households buy electricity, this sampling scheme should sample households with equal probability. The household weights for this area were derived as for the statistical areas sampled using the ABS sampling scheme, and the total of the household weights for rural Victoria is the number of occupied dwellings at the 1981 Population Census in Victoria less the Melbourne Statistical Division, the Geelong and Bendigo Statistical Districts and the LGA of Warrnambool.

For Kalgoorlie and Brewarrina, the electoral roll was sampled for persons whose address was in either of these towns. This is a sampling of persons, not households, and the persons are aged 18 or older. As a sampling of households, it selects households with weight according to the number of persons aged 18 or older. It is also debatable that such a process would correctly identify all voters, let alone all residents, of the LGAs of Kalgoorlie and Brewarrina. It is also possible that some voters living outside these LGAs have addresses within the LGA and so would be incorrectly included in the sample. Nevertheless, despite these biases, the household weights were derived like those for the other areas. For the achieved sample of Kalgoorlie residents, the total of the household weights is the number of occupied dwellings at the 1981 Population Census for the LGA of Kalgoorlie. Similarly, for the achieved sample of Brewarrina residents, the total of the household weights is the number of occupied dwellings at the 1981 Population Census for the LGA of Brewarrina.

Note that an occupied private dwelling is assumed here to be equivalent to a household, and any difference in their definitions would result in lack of correspondence between the achieved sample and the sampled population. These weightings are uncontroversial for the sample selected by ABS from the given ABS statistical areas, and are reasonable for rural Victoria, but less so for Kalgoorlie and Brewarrina. However, in the interests of simplicity, the same basic method for deriving the weightings has been followed.

The numbers of occupied private dwellings in each statistical area at the 1981 Population Census are given in Table 4.1. The corresponding weights are called HWEIGHT1 on the household added weights file. Use of these weights assumes that daily statistics are being analysed. For some purposes, the weights will need to be multiplied by 365 (for annual statistics), or some other factor.

Table 4.1 Number of households in the achieved sample and the benchmark population for the Day-to-Day Travel Survey. **The benchmark** populations are occupied private dwellings at the 1981 Population Census. For the 'Rest of state' stratum, the benchmark population relates to the state or territory less the capital city; otherwise the benchmark population relates only to the given statistical area.

Statistical area	Number of households: Achieved sample	Number of households: Benchmark population
Australian Capital Territory		
Canberra Statistical District (ACT part)	482	68351
Whole of ACT	482	68591
New South Wales		
Newcastle Statistical District	283	126242
Dubbo (C) LGA	259	8363
Shoalhaven (C) LGA	207	15771
Brewarrina (S) LGA	48	660
Rest of state	797	597679
Sydney Statistical Division	3535	1065079
Victoria		
Geelong Statistical District	219	43843
Bendigo Statistical District	303	18997
Warrnambool LGA	250	6711
Other non-Melbourne Victoria	622	277347
Rest of state	1394	346898
Melbourne Statistical Division	3515	892047
Queensland		
Townsville (C) LGA + Thuringowa (S) LGA	254	28405
Charters Towers C LGA + Dalrymple S LGA	145	2839
Gold Coast (C) LGA + Ayr (S) LGA	2236	45558
Rest of state	2635	366317
Brisbane Statistical Division	2279	331915

Table 4.1 (Continued)

Statistical area	Number of households: Achieved sample	Number of households: Benchmark population
South Australia		
Whyalla (C) LGA	208	9347
Central Yorke Peninsula (DC) LGA	53	1302
Rest of state	261	111971
Adelaide Statistical Division	1922	320165
Western Australia		
Geraldton (T) LGA	150	5572
Esperance (S) LGA + Ravensthorpe (S) LGA	98	3087
Kalgoorlie (T) LGA	73	2811
Rest of state	321	105467
Perth Statistical Division	2038	298133
Tasmania		
Burnie-Devonport Statistical District	171	22884
Rest of state	171	79961
Hobart Statistical Division	412	55637
Northern Territory		
Alice Springs LGA	62	4413
Rest of state	62	12547
Darwin Statistical Division	402	16502

Sources: Day-to-Day Travel Survey in Australia 1985-86 (FORS)
1981 Population Census (ABS)

4.3.2 Persons

Given the structure of the survey, the household weights derived in Section 4.3.1 would suffice for persons. However, it was decided to post-stratify by age group and sex within each statistical area. It was also assumed that non-response depended primarily on whether the person travelled or not, and that the internal weights derived by Socialdata adequately adjust for this (Section 4.2). The benchmark population was taken as the population in each statistical area at the 1981 Population Census, classified by sex (male, female) and age group (9-14, 15-29, 30-49, 50-64 and 65 or older). Benchmarking to sex and age group would take account of any differential non-response between the age and sex groups. Non-coverage would be analogous to that for households (see Section 4.3.1).

As for households, the fixed weighting factor was the ratio of the number of persons with the same sex and in the same age group as the particular person in the achieved sample to the internal weight for the person, corrected for the limited precision of these weights on the data file.

The age group and sex distribution of persons in each statistical area at the 1981 Population Census are given in Table 4.2. The corresponding weights are called PWEIGHT1 on the persons added weights file. Use of these weights assumes that daily statistics are being analysed. For some purposes, the weights will need to be multiplied by 365 (for annual statistics), or some other factor.

4.3.3 Trips

For trips, the persons weighting factor should be applied. No separate benchmark population exists for trips. To calculate this weight, PWEIGHT1 must be divided by PWEIGHT (on the persons data file) and multiplied by TRPWEIGHT (on the trips data file). Use of these weights assumes that daily statistics are being analysed. For some purposes, the weights will need to be multiplied by 365 (for annual statistics), or some other factor.

Table 4.2 Number of persons in the benchmark population for the Day-to-Day Travel Survey. **The benchmark populations are persons, classified** by sex (male, female) and age group (9-14, 15-29, 30-49, 50-64, 65 or older) at the 1981 Population Census. **For the 'Rest of** state' stratum, the benchmark population relates to the state or territory less the capital city; otherwise the benchmark population relates only to the given statistical area. **See Table 4.1 for** exact definitions of the statistical areaas.

Statistical area		Age group 9-14	15-29	30-49	50-64	65->	Total
Australian Capital Territory							
Canberra Stat Dis	M	13628	28395	32547	11707	3824	90101
	F	12956	29767	32194	11455	5678	92050
ACT	M	13687	28784	32689	11752	3839	90751
	F	13015	29901	32297	11489	5686	92388
New South Wales							
Newcastle Stat Dis	M	20868	50163	47754	30186	17392	166363
	F	20140	47210	46715	30949	23974	168988
Dubbo (C) LGA	M	1815	3623	3604	1692	971	11705
	F	1743	3696	3555	1676	1300	11970
Shoalhaven (C) LGA	M	2327	5575	5365	3891	2925	20083
	F	2285	4910	5303	4236	3012	19746
Brewarrina (S) LGA	M	180	349	286	159	80	1054
	F	152	279	250	140	79	900
Rest of state	M	108522	241846	241636	143153	87329	822486
	F	104222	226687	228919	142691	108788	811307
Sydney Stat Div	M	162695	401135	428771	235075	128418	1356094
	F	155032	401709	417851	242429	195060	1412081

Table 4.2 (continued)

Statistical area		Age group 9-14	15-29	30-49	50-64	65->	Total
Victoria							
Geelong Stat Dis	M	7888	17830	16510	9789	5396	57413
	F	7566	17426	16536	10137	8005	59670
Bendigo Stat Dis	M	3324	7530	6560	3754	2854	24022
	F	3128	7790	6571	4286	4428	26203
Warrnambool (C) LGA	M	1238	2659	2406	1460	974	8737
	F	1198	2937	2424	1524	1439	9522
Rest of non-Melb Vic	M	54661	109977	110422	63717	41754	380531
	F	51540	103321	103473	63024	51531	372889
Rest of state	M	67111	137996	135898	78720	50978	470703
	F	63432	131474	129004	78971	65403	468284
Melbourne Stat Div	M	148146	351596	357693	193683	102751	1153869
	F	143164	351699	354323	195147	155124	1199457
Queensland							
Townsville	M	5456	14872	12906	6205	3214	42653
	F	5169	13612	11682	5973	4213	40649
Charters Towers	M	581	1254	1291	777	670	4573
	F	559	1111	1090	615	657	4032
Gold Coast	M	6460	15751	15075	11279	10346	58911
	F	6000	15212	15320	13044	12421	61997
Rest of state	M	73166	162691	160811	90662	61092	548422
	F	69573	151226	147838	88852	69171	526660
Brisbane Stat Div	M	54136	133118	129845	71515	42623	431237
	F	52109	133198	128062	74821	62241	450431

Table 4.2 (continued)

Statistical area		Age group 9-14	15-29	30-49	50-64	65->	Total
South Australia							
Whyalla (C) LGA	M	2031	4600	3981	1891	651	13154
	F	2013	4264	3839	1687	792	12595
Central Yorke	M	236	502	488	337	244	1807
	F	232	429	437	336	278	1712
Rest of state	M	20995	45400	45637	26252	15323	153607
	F	20006	41951	41663	24611	17611	145842
Adelaide Stat Div	M	48299	119216	115083	70675	40949	394222
	F	45917	121470	118190	73419	61006	420002
Western Australia							
Geraldton (T) LGA	M	1272	2655	2255	1164	707	8053
	F	1267	2525	2124	1131	776	7823
Esperance	M	729	1600	1555	543	306	4733
	F	655	1296	1318	445	295	4009
Kalgoorlie (T) LGA	M	495	1372	1284	584	433	4168
	F	482	1268	998	439	434	3621
Rest of state	M	22314	53925	53588	23742	13622	167191
	F	20828	46014	42591	20389	13641	143463
Perth Stat Div	M	51370	116285	118206	57754	34840	378455
	F	48860	119357	118145	59183	49287	394832

Table 4.2 (continued)

Statistical area		Age group 9-14	15-29	30-49	50-64	65->	Total
Tasmania							
Burnie-Devonport	M	4339	8548	8368	4690	3064	29009
	F	4433	9169	8528	4983	3920	31033
Rest of state	M	15009	31677	31450	17459	10599	106194
	F	14448	31220	29731	16977	13467	105843
Hobart Stat Div	M	9154	22408	20403	11749	6871	70585
	F	8697	22695	20559	12161	9927	74039
Northern Territory							
Alice Springs LGA	M	1184	2653	2763	859	335	7794
	F	1153	2681	2383	749	388	7354
Rest of state	M	3904	9708	9036	3167	1096	26911
	F	3626	8658	6891	2370	1057	22602
Darwin Stat Div	M	3699	8974	10204	2741	791	26409
	F	3588	8943	8291	2152	809	23783

Source: 1981 Population Census (1981)

4.4 Weighting to the target population

4.4.1 Households

The sampled and target populations are the same for the capital cities.

However, extrapolation from the sampled population to the target population for most other areas is difficult. The areas sampled were chosen purposively, and not at random. Furthermore, the basis on which they were chosen seems to be that the largest possible town was selected. Even for the 'sparse' areas, Kalgoorlie and Brewarrina, the largest town in the general area rather than the whole area was sampled. Although it is true that most people do live in towns, it is likely that these people will have different travel patterns to those who live away from towns. This problem is exacerbated by the selection of provincial centres, which service the other towns and rural areas nearby.

In these circumstances, many analysts of this data set will be reluctant to extrapolate from the sampled population of households to the target population. However, it is important to define weights that may give reasonable estimates for the target population, as this is the base population for most accident data.

For households, the target population is all households in Australia, and a desirable benchmark population is all occupied private dwellings at the 1981 Population Census. For the capital cities, the sampled and target populations are the same, and so the same weights can be used. For the rest of each State and of the Northern Territory, the proposed benchmark population is the number of occupied private dwellings at the 1981 Population Census in the given area. For the Australian Capital Territory, the few households not in the Canberra Statistical District (ACT Part) have been included in the benchmark population.

One consequence is that, before adjustment by the internal weights (see Section 4.2), all sampled non-capital city households in a State are given equal weight. Since the sampling fraction was higher for statistical areas with fewer households, this weighting scheme results in a bias to smaller statistical areas. The disparity in sampling fractions is large, for example, the ratio of benchmark households to achieved households was 446 in the Newcastle Statistical District, but only 32 in the Dubbo LGA. Some analysts will find this choice of benchmark unacceptable. However, given the inherent bias in the selection of the sampled population, it is unlikely that a universally acceptable weighting scheme could be devised.

These weights are called HWEIGHT2 on the household added weights file.

4.4.2 Persons

For persons, the target population is all persons in Australia aged 9 years or older. Since the survey sample was drawn from a census of ordinary households, it misses some groups of persons. Persons in institutions, such as prisons, nursing homes, boarding houses, colleges, hotels, motels or defence establishments are not covered by the survey. Some of these institutions contain persons who are unlikely to have any trips (for example, prisons and nursing homes). However, students, tourists and defence personnel are examples of persons who may be in these other institutions and whose travel patterns may differ from the rest of the population. Only 4.5% of the population of Australia at the 1981 Population Census did not live in a private household, and no adjustment has been made for this source of bias.

The benchmark population for persons was defined similarly to that for households, with weights calculated separately for the capital cities and for the rest of each State and the Northern Territory. The benchmark population was the number of persons at the 1981 Population Census, classified by sex (male, female), age group (age 9 to 14, 15 to 29, 30 to 49, 50 to 64 and 65 or older), State/Territory and capital city/rest of State.

These weights are called PWEIGHT2 on the persons added weights file.

4.4.3 Trips

The weighting factor for trips was taken to be the same as that for persons, before adjusting for the internal weight (see Section 4.2). These weights must be derived by the user from PWEIGHT2 on the persons added weights file, and the internal weights (PWEIGHT and TRPWEIGHT) on the existing data base. So TWEIGHT2 is given by PWEIGHT2 muliplied by TRPWEIGHT divided by PWEIGHT.

4.5 Summary

There are three sets of weights. The first adjusts for non-response within the sample, the second extrapolates to the sampled statistical areas and the third extrapolates to the target population of all Australia. The sampled statistical areas includes all capital cities, but only selected Statistical Districts and LGAs from elsewhere.

For households, the first set of weights adjusts for known non-response with season of the year, and the weighted number of households in each season in each sampled area is exactly one quarter of the achieved sample in that area. The second set of weights ensures that the weighted number of households in each sampled area agrees with the number recorded at the 1981 Population Census, while the third set benchmarks to the number of households in Australia at the 1981 Population Census, classified by State or Territory and capital city/rest of State.

For persons, the first set of weights adjusts for the lower response rate of non-travellers compared with travellers. The second set of weights benchmarks to the 1981 population, classified by sex and age group (9-14, 15-29, 30-49, 50-64 and 65+), in each of the sampled areas, while the third set of weights benchmarks to the 1981 population of Australia, classified by sex, age group, State or Territory and capital city/rest of State.

For trips, the first set of weights adjusts for the lower response rate of non-travellers compared with travellers. The second and thrid sets of weights follow those for persons.

For most purposes, the third set of weights would be used to estimate quantities from the survey.

5. System documentation

5.1 Location of the data

The exposure data is stored as four survey files on the IP Sharp system. The four files are as follows :

System number	Description
40	Household file
41	Persons file
42	Trip file
43	Trip mode file

The files are stored in user number 2635020 and can be accessed (but not changed or added to) from user number 3547260. There are other MABRA systems available, but these are beyond the scope of this documentation.

On the IP Sharp system, each file is accessed by its survey file number. Information from one file can be linked to information in another file by using the Link Systems command and specifying the appropriate key fields (this is described in detail below).

MABRA is used to manipulate the files, list records and produce preliminary frequency tables and crosstabulations. XTABS is used to define variables based on the MABRA fields and produce more sophisticated analyses with appropriate labelling on tables.

Relevant documentation for the IP Sharp system is found in the following manuals:

MABRA manual (1980)
Changes to MABRA (1981)
MABRAXT (1987)
XTABS manual (1980)

5.2 MABRA field descriptions

The MABRA fields for each file are described below. Except for the key fields and an occupation field for part of the Victorian sample (see below), they correspond exactly with the data items set provided by Socialdata. The key fields (used for linking two systems) were created by the IP Sharp programmers.

The following list gives the name and description of all fields in each of the four files. The column numbers and lengths are provided for reference only (from documentation provided by Socialdata) as they have no meaning within the four survey files on the IP Sharp system.

Frequency tables for each field can be found in the documentation listings along with descriptions of the MABRA fields and their format on the IP Sharp system.

HOUSEHOLD FILE

Columns	**MABRA Fieldname**	Length	Contents
1-7	**HKEY**	7	Household key (1 State, 2-4 Region, 5-7 Dwelling)
1	STATE	1	State
2-4	REG	3	Region within state
8-9	NOP	2	Number of people
10-11	NOL5	2	Number in household younger than 5.
12-13	NO5TO8	2	Number in household aged 5 to 8
14-15	NO9O	2	Number in household aged 9 or over
16	TEL	1	Telephone (1=yes, 0=no)
17-18	NOBY	2	Number of bicycles
19-20	NOMBY	2	Number of motorbikes
21-22	NOCARS	2	Number of cars
23-24	NOOTHER	2	Number of other vehicles
25-27	TRDAYO	3	Travel day originally assigned
28-30	TRDAYA	3	Travel day actually reported
31-33	QDAY	3	Questionnaire returned day number
34-36	COMM	3	Comment code
37-38	RESP	2	Response status
39-42	POST	4	Postcode
43-46	HWEIGHT	4	Household weighting factor (3 dec places)
	HWEIGHT1		
	HWEIGHT2		

PERSONS FILE

Columns	MABRA Fieldname	Length	Contents
1-7	HKEY	7	Household key (as in Household file)
8-9	PNO	2	Person number (within household)
10-11	YOB	2	Year of Birth (00 if pre 1900)
12-13	HOWOLD	2	Age at survey date (99 if 100+)
14	SEX	1	Sex (1=male, 2=female)
15	COB	1	Country of birth (1-8)
16	EDUC	1	Highest level of education attended (0-4)
17	EMP1	1	Employment - Home duties (1=yes)
18	EMP2	1	Employment - Looking for work "
19	EMP3	1	Employment - Retired,age pensioner"
20	EMP4	1	Employment - Other not employed "
21	EMP5	1	Employment - Full time student "
22	EMP6	1	Employment - Part time student "
23	EMP7	1	Employment - Part time,casual work
24	EMP81	1	Employment - Full time work "
25	CARLIC	1	Car licence (1=yes,0=no)
26	TRUCKLIC	1	Truck licence (1=yes,0=no)
27	MBYLIC	1	Motorcycle licence (1=yes,0=no)
28-29	NOTRIPS	2	Number of trips
30	TRIPO	1	Trip origin (1=home,2=elsewhere)
31	REASON	1	Reason for no travel (0-9)
32-34	TRDAY	3	Travel day for this person
35-38	PWEIGHT	4	Person weighting factor (3 dec pl)
	PKEY		Unique person key field (HKEY*100+PNO)
	OCC		Occupation field (added as a separate field for Victorian sample - no coding information available)
	PWEIGHT1		
	PWEIGHT2		

TRIP FILE

Columns	MABRA Fieldname	Length	Contents
1-7	HKEY	7	Household key (as in Household file)
8-9	PNO	2	Person number
10-11	TRPNO	2	Trip number
12-15	TRPST	4	Trip start time (0000 to 2400+)
16-19	TRPDUR	4	Trip duration (in minutes)
20-21	TRPPUR	2	Trip purpose (1-14)
22	TRPM1	1	Trip mode - Walk (1=yes)
23	TRPM2	1	Trip mode - Bicycle (1=yes)
24	TRPM3	1	Trip mode - Bus (1=yes)
25	TRPM4	1	Trip mode - Train (1=yes)
26	TRPM5	1	Trip mode - Tram (1=yes)
27	TRPM6	1	Trip mode - Taxi (1=yes)
28	TRPM7	1	Trip mode - Ferry (1=yes)
29	TRPM8	1	Trip mode - Motor bike (1=yes)
30	TRPM9	1	Trip mode - Car driver (1=yes)
31	TRPM10	1	Trip mode - Car passenger (1=yes)
32	TRPM11	1	Trip mode - Truck (1=yes)
33	TRPM12	1	Trip mode - Semi-trailer (1=yes)
34	TRPM13	1	Trip mode - Other including plane (1=yes)
35	TRPM14	1	NB NOT a trip mode - indicates travel on SA road (refer Socialdata) (1=yes)
36-39	TRPDIST	4	Trip distance (in 100 metre units)
40-43	TRPWEIGHT	4	Trip weighting factor (3 dec places)
	PKEY		Person key (as in Persons file)
	TKEY		Unique Trip key

TRIP MODE FILE

Columns	MABRA Fieldname	Length	Contents
1-7	HKEY	7	Household key (as in household file)
8-9	PNO	2	Person number (as in persons file)
10-11	TRPNO	2	Trip number (as in trip file)
12	MODNO	1	Mode number (1-5)
13-14	TRPMODE	2	Trip mode (01-14)
15-18	DURATION	4	Duration in minutes
19-22	DIST	4	Distance (in 100 metre units)
	TKEY		Unique trip key (as in trip file)

5.3 Code frame changes

The codes for each field are listed in Appendix H of the report "A Data Base for the Evaluation of Road User Risk in Australia" prepared by Socialdata. The following codes were omitted from the original report or documented incorrectly.

Household file

COMM (Comment code)

- 007 Two different days recorded in the household, coders allocated the same day to each person
- 096)
- 097)Households in NSW whose questionnaires were returned
- 098)late because of mail strike (dropped for weighting)

RESP (Response status)

- 09 Non-responding households - no information (recorded as 02 in original documentation)
- 02 This is an invalid code (should have been 09 in original documentation)

Trip file

TRPM14 (Trip mode 14)

- 1 Travel on South Australian road (not a trip mode)

5.4 Linking files

The data is stored as four hierarchical files. Any two files can be linked by using the Link Systems command (LS) in MABRA and specifying the appropriate key field. One file is the host file (the file to be analysed) and the other is the linked file (the file from which fields are to be copied).

The key fields were created by the programmers at IP Sharp and consist of a combination of fields which uniquely identify a record within each particular file. Each file contains its own key and the keys of the files above it in the hierarchy.

Hierarchy of files	Unique key field
Household (40)	HKEY
Persons (41)	PKEY
Trips (42)	TKEY
Tripmode (43)	

Two examples of linking systems can be found in the documentation listings. One example shows linking to a file lower in the hierarchy (Person-Trip) and the other to a file higher in the hierarchy (Person-Household).

Key fields. When linking two files, the appropriate key to use is the one belonging to the file which is highest in the hierarchy.

Totalling. Usually, when linking from the host file to a file higher in the hierarchy, totalling across records is not requested (for example, person to household). Usually, when linking from the host file to a file lower in the hierarchy, totalling across records is requested (for example, person to trip). If totalling is not requested in this case, only fields from the first matching record of the linked file will be copied across.

Fields to be copied. Any field from the linked system can be copied across to the host system and given the same or a new name. The field specification DELTA COUNT (from the linked system) will give the total number of records in the linked system for each of the records in the host system (for example, number of trip records per person). An example is shown in the documentation listings.

5.5 Identifying the Victorian sample selected from SEC records

The records of households in Victoria selected from Electricity Authority records, rather than ABS Census Collector's districts, can be identified within the data base. These records have a state code of 3 (STATE EQ 3) and region codes 400-499 (REGION GE 400 AND REGION LE 499).

5.6 Changes made to the data base

Conversion of binary fields to integer. The fields for Employment (EMPL, Persons file) and Tripmode (TRPMODE, Trip file) were originally stored as multi-valued binary fields. These were converted to a series of single valued integer fields for two reasons:

(1) the difficulty of accessing each of the values separately in multi-value binary format, and

(2) problems in the IP Sharp MABRA code relating to binary fields in very large data bases. This led to a series of 'Workspace Full' messages when attempting produce frequency tables.

Recreation of the data base to distinguish blank from zero. Because the original version of the data base failed to distinguish between blanks (indicating missing data) and zeros, the data base was recreated on 12th May to make this distinction. Blanks were converted to a value of '-1' in the data base.

The addition of fields for state and region. The information for state and region were embedded within the Household key. Rino Ciaccia from IP Sharp recommended these fields be created as pseudo fields by performing a mathematical calculation on the household key. This proved impossible using a NOAPLCHARS terminal (see below), so the fields were created as separate fields by the IP Sharp programmers.

5.7 Accessing the IP Sharp system

The IP Sharp system can be accessed with either a directly connected terminal or a microcomputer and a modem. Terminals connected directly to the system may use the APLCHARS character set or the NOAPLCHARS character set. Accessing the system via a microcomputer restricts the user to the NOAPLCHARS set, unless a special communications package and processor board are used. The NOPAPLCHARS set is a subset of the APLCHARS set, so the user is unable to perform some commands using the NOAPLCHARS set (see p29 of the "Changes to MABRA" documentation for a description of the NOAPLCHARS set).

Modem settings. Accessing the IP Sharp system using a modem requires the following settings :

CAPS Lock on
1200 baud
Full Duplex
7 data bits
1 stop bit
Even parity
Local echo on

The Sydney dial-up number is 221 5099.

For information on accessing the system using a modem, phone:

Canberra (062) 73 3700 (David)
Sydney (02) 232 6366

5.8 Logging onto the IP Sharp system and accessing the data base

The following commands are necessary to log onto the system and access the data. **Each command should be followed by a carriage return.**

```
O)                    [Letter O, not zero]
[A crosshatched line appears to obscure the user number and password]
)2635020:pw           [where 2635020 is the user number
                       and pw is replaced by the current password]
[A message indicating you are through to the IP Sharp system appears]
) LOAD 2297815
MABRAXT
NOAPLCHARS            [if using a non-APL terminal]
START [A prompt indicating you are in MABRA system mode appears]
[You are now ready to enter one of the files (for example, household)]
ENTER;40
```

If user number 3547260 is used, the survey file number must be preceded by the user number of the owner of the file, that is, 2635020.