

Task 1.7 Combined Data Processing

The general objective is to make the best use of all data sources in model development and application. We need to combine that data sources for a number of purposes:

- providing a base road matrix for assigning to obtain costs for model calibration, and in the process checking for bias in the observed trip data;
- providing trip data for model estimation¹, which requires a decision on what data to use for this purpose.

As illustrated in Table 1 there is duplication between the data sources, which needs to be overcome when the data are combined. The processes are described below.

■ Table 1 Duplication between Data Sources

Purpose/Mode	Car	Public Transport
All	Double-counting between roadside interview locations	Double-counting between train, bus and ferry surveys
HBE _d	Household Roadside	Household School Bus Public Transport
Other purposes	Household Roadside	Household Public Transport
Resident external	Household External roadside	Household Public Transport
Non-Residents	External roadside	Public Transport

Note: cars used to access public transport (park-&-ride) in the main survey should normally be coded as main mode public transport though the trip linking procedure.

¹ We anticipate using expanded trip data for the attraction and distribution mode choice calibration, (adjusting the statistical tests to allow for the interview samples used).

Task 1.7.1 Double-Counting between Roadside Survey Sites

Inputs

Roadside survey data

Processing

- 1) The potential double-counting should be identified using the locations of the interview sites and the geometry of the road system. Examples are (if Internal OD Survey undertaken):
 - external-external trips are observed at two study area entry points;
 - through-isthmus trips are observed at two isthmus entry points;
 - external-isthmus trips are observed on both external and isthmus sites;
 - there will be other double-counting related to local network geometry some of which may be identified by counter-intuitive movements (eg outbound trips at an inbound site, indicating a multiple crossing).
- 2) Using each data set, the potentially multiple counted movements should be tabulated and compared between data sets. In principle the volumes should be similar.
- 3) The multiple counting should be removed by appropriate factoring of the movements for each survey (by 0.5, 0.33 etc). These additional factors should be appended to each survey trip record.

Outputs

Roadside survey data with multiple counting factors appended

Task 1.7.2 Multiple-Counting between Public Transport Surveys

Inputs

Public transport survey data

Processing

- 1) The access/egress mode data provides evidence of where more than 1 public transport mode is used on a single trip. Using each data set, the potentially multiple counted movements should be tabulated and compared between data sets. In principle the volumes should be similar.
- 2) The multiple counting should be removed by appropriate factoring of the movements for each survey (by 0.5, 0.33 etc). These additional factors should be appended to each survey trip record.

Outputs

Public transport survey data with multiple counting factors appended

Task 1.7.3 Combining Estimates from Different Surveys

Inputs

Survey data files

Processing

Establish General Data Consistency

- 3) All of the comparisons of data sources which follow are envisaged to be based on TA to TA expanded matrices². Comparisons of overlapping data from different sources would allow for the sampling errors:
 - assume variance of household survey is 50% larger than equivalent Poisson variance; a binomial sampling distribution applies to the intercept surveys;
 - the variance of the difference between two samples equals the sum of the variances of the individual samples (and use a simple t-test of 95% significance);
 - a frequency distribution of matrix cells by number of sampled (unexpanded) trips would also be useful to compare survey coverage.

- 4) The comparisons (specified in Table 2) would be for those parts of the trip matrices which overlap:
 - this is for residents' trips only and in principle should be done for each modelled trip; purpose separately but some merging of trip purposes is inevitable where data is sparse;
 - for car, this is for car driver trips only; for HBEd, as these will mostly be escort trips, we should recognise that direct comparisons may be subject to uncertainties relating to the treatment of escorts in different surveys;
 - for car (from the HTS) and the Internal OD survey (if undertaken), this is trips crossing the Isthmus cordon; caution should be exercised for “through” trips between the West and the North which can occur via the Isthmus or via the Upper Harbour corridor;
 - for car (from the HTS) and the external cordon, this is for external trips;
 - for bus and rail, the full residents' matrices from the HTS and the PTIS can be compared; because the main mode coding effectively suppresses some bus access trips to rail, these comparisons by individual PT mode should be made excluding the public transport multiple counting factors;
 - for school bus, the volumes should only be compared in locations where school bus intercept data has been obtained and only apply to secondary school (as distinct from primary and tertiary education trips).

² Small TAs could be merged with others.

- 5) This analysis will establish/confirm:
- that the intercept data provide much better matrix coverage, and will quantify this;
 - whether the surveys are consistent, at the TLA level, ie the numbers of trips are not statistically different.

■ **Table 2 Duplication between Data Sources**

Purpose/Mode	Car Driver	Bus	Rail
HBEEd (residents)	Household Isthmus cordon	Household School Bus Public Bus	Household Rail
Each other purpose (residents)	Household Isthmus cordon	Household Public Bus	Household Rail
External trips by purpose (residents)	Household External roadside	Household Public Bus (external)	Household Rail (external)

Determine Strategy and Implement

- 6) The following presumes that the outcomes of the consistency checks are as we would wish – the intercept surveys provide better coverage and are consistent with the household data. If either of these proves to be untrue, in particular, if the trip numbers are significantly different, then we shall need to re-think our approach.
- 7) The simplest method of combining the best matrices from these surveys would be to:
- use only the non-residents’ trip data from the external roadside cordon;
 - for cells for which intercept data is available (from the isthmus cordon or public transport and school bus surveys), use the intercept data;
 - for all other matrix cells use household survey data.
- 8) This is a good method where sample sizes for the intercept surveys far exceed those from the household surveys, but, where the samples are similar, the variance of the observed matrix cells can be reduced by a weighted combination of the alternative sources. For sparse matrices (as these will be), this is not particularly constructive and we will aggregate cells to ensure sufficient observations to determine an appropriate average weight for the group of cells.

Outputs

Best estimate trip matrices

Note