WORKING PAPER
ITS-WP-02-15

A Spatial and Statistical Approach for Imputing Origin-Destination Matrices from Household Travel Survey Data: A Sydney Case Study

By

Tu T. Ton & David A. Hensher

October, 2002

ISSN 1440-3501

INSTITUTE OF TRANSPORT STUDIES
The Australian Key Centre in Transport Management

The University of Sydney and Monash University
Established under the Australian Research Council's Key Centre Program.

NUMBER: Working Paper ITS-WP-02-15

TITLE: **A Spatial and Statistical Approach for Imputing Origin-Destination Matrices from Household Travel Survey Data: A Sydney Case Study**

ABSTRACT:

A Household Travel Survey (HTS) is a valuable instrument for collecting data suitable for studying the travel behaviour of a sample of households in a specific geographical context. One important output from the trip data after expansion to the population is a set of origin-destination (O-D) trip matrices for combinations of trip purpose, time of day and mode of transport. However, the O-D matrices generally take the form of sparse matrices (ie cell values are mostly zero). The degree of sparseness of these matrices is a function of sample size (a consequence of cost constraints), segmentation requirements and the spatial resolution of a geographical zoning system. Another factor contributing to the sparseness is the non-revelation of information in some cells in order to protect the privacy of households who live in those cells where their total amount of travel in a cell is less than a cut-off criterion (eg < 200 trips).

Establishing an appropriate value to assign to a 'zero value' cell is a non-trivial task. There are two key issues to work through. The first is how to set up a classification rule to determine either if zero value cells have no travel related activity at all (ie genuine zeroes) or the travel values are truly missing. The second issue is the development of a trip allocation rule to assign the number of trips to each missing value cell within the constraint of a given total number of trips to be allocated to each missing value cell (given knowledge of marginals).

This paper shows how spatial and statistical techniques can be implemented to estimate the number of missing value cells and the number of trips associated with each missing value cell. The classification rule is a spatial one in locating missing value cells for any travel activities between each origin and destination. It is driven by the mean trip length distribution of the origin and destination distance among traffic zones. The trip allocation rule is constructed to allocate the number of trips to missing value cells using a distribution assumption (such as the uniform). The two rules are then combined in a process based on the proportion of trip purposes and modes of travel for a whole sample of household travel records. We implement the method for Sydney for the period 1998-2000 to obtain total passenger trip movements for linked trips by five purposes, six modes and six times of day.

KEYWORDS: Household Travel Surveys, OD Matrix, Transport Planning Techniques.

AUTHORS: Tu T. Ton and David A. Hensher

CONTACT: Institute of Transport Studies (Sydney & Monash)
The Australian Key Centre in Transport Management, C37
The University of Sydney NSW 2006, Australia

Telephone: +61 9351 0071
Facsimile: +61 9351 0088
Email: itsinfo@its.usyd.edu.au
Internet: http://www.its.usyd.edu.au

DATE: October 2002

1. Introduction

A Household Travel Survey (HTS) is a valuable instrument for collecting data suitable for studying the travel behaviour of a sample of households in a specific geographical context. One important output from the trip data after expansion to the population is a set of origin-destination (O-D) trip matrices for combinations of trip purpose, time of day and mode of transport. However, the O-D matrices generally take the form of sparse matrices (ie cell values are mostly zero). The degree of sparseness of these matrices is a function of sample size (a consequence of cost constraints), segmentation requirements and the spatial resolution of a geographical zoning system. Another factor contributing to the sparseness is the non-revelation of information in some cells in order to protect the privacy of households who live in those cells where their total amount of travel in a cell is less than a cut-off criterion (eg < 200 trips).

Establishing an appropriate value to assign to a 'zero value' cell is a non-trivial task. There are two key issues to work through. The first is how to set up a classification rule to determine either if zero value cells have no travel related activity at all (ie genuine zeroes) or the travel values are truly missing. The second issue is the development of a trip allocation rule to assign the number of trips to each missing value cell within the constraint of a given total number of trips to be allocated to each missing value cell (given knowledge of marginals).

This paper shows how spatial and statistical techniques can be implemented to estimate the number of missing value cells and the number of trips associated with each missing value cell.

2. An overview of the methodology

The method proposed for imputing OD matrices involves the use of four key steps. Table 1 provides a description of these four steps and associated aims in the imputation process.

Table 1: Key Steps Used in Imputing O-D Trip Matrices

Key Steps	Aims
Step 1 - Process OD-trip matrices	to identify the degree of sparseness of OD matrices
Step 2 - Construct Missing Cell Classification Rule	to identify all missing value cells
Step 3 - Construct Trip Generation Rate Assignment Rule	to assign different trip generation rates to each missing value cell
Step 4 – Construct Missing Cell and Trips Allocation Rule	to allocate missing value cells and associated trip generation rates to different segmentations (eg by trip purposes, transport modes and time of days)

Step 1 involves processing the database of OD trip matrices. The aim in implementing this step is to identify the degree of sparseness of OD matrices. In Step 2, a classification rule is developed to locate missing value cells for travel activities between each origin and destination. The classification rule is driven by the mean trip length (MTL) distribution of the origin and destination distance among traffic zones. For every missing value cell, we need to determine the number of trips associated with it. The trip generation rate assignment rule in Step 3 is developed to serve this aim. The trip allocation rule is constructed as a final step (Step 4) to allocate missing value cells and associated trip generation rates to different segments of OD matrices (eg by trip purposes, transport modes and time of days).

With this overview we now present in more detail the underlying implementation procedure.

3. Process OD-trip matrices

The aim of this step is to identify the degree of sparseness of OD matrices. This step involves two key tasks. The first task was to represent and structure of the imputing problem as a multi-dimensional matrix. The second task is to identify the degree of sparseness of OD matrices.

Representing the imputing problem as a multi-dimensional matrix based environment

A three dimensional (3D) array of matrix objects was used to represent the complete set of OD matrices segmented by trip purpose, time of day and transport mode. Each cell of this array is a two dimensional OD matrix. In the Sydney case study, a 3D array represents 180 OD matrices (5 trip purposes x 6 times of day x 6 transport modes). The dimension of every OD matrix is 904 origins (rows) and 904 destinations (columns) (see Figure 1).

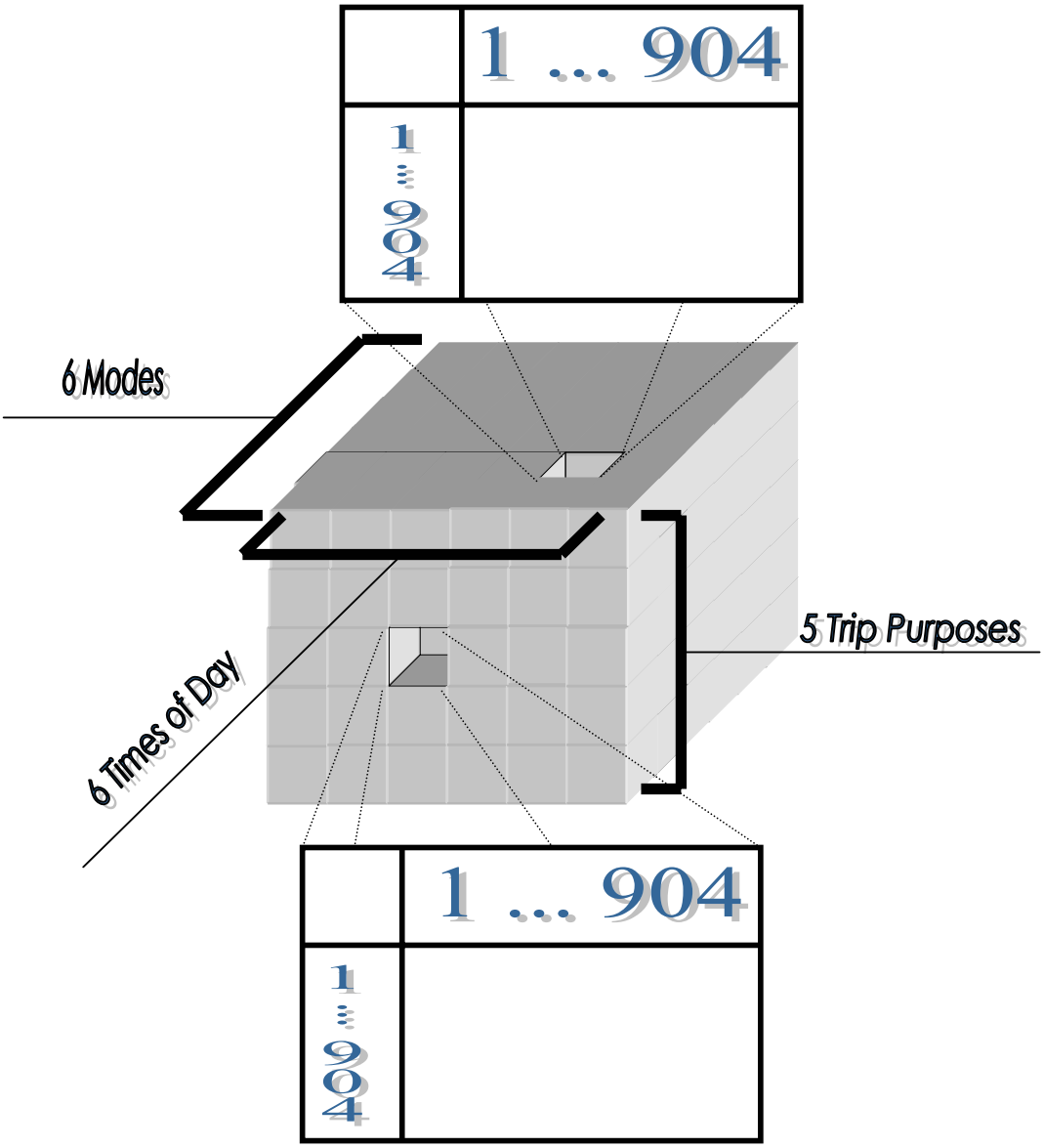
3.1 Identifying the degree of sparseness of OD matrices

In general, the degree of sparseness of a matrix can be measured by a number of zero values cells of the matrix. In the context of a household travel survey, the degree of sparseness derived from household travel records is a function of sample size (a consequence of cost constraints), segmentation requirements and the spatial resolution of a geographical zoning system. Another factor contributing to the sparseness is the non-revelation of information (from agencies who distribute the data) in some cells in order to protect the privacy of households who live in those cells where the total amount of travel in a cell is less than a cut-off criterion (eg < 200 trips).

Given the size and complexity of the OD matrices, we initially ignore the segmentation constraint. Specifically for the Sydney case study set out below, instead of scanning all 180 OD matrices, we only need to locate the zero value cells in a single total trip matrix for all trip purposes, all times of day and all transport modes. In order to do that, we need to calculate the OD matrix from the 180 OD matrices. The calculation involves aggregating the number of trips with the corresponding cell across all trip purposes, all times of day and all modes of transport. The advantage of working with the total trip matrix is that all zero value cells can be quickly located.

Having found zero value cells, the challenge is to determine if these zero value cells have either no travel related activity at all (ie genuine zeroes) or the travel values are truly missing. The next section describes how classification rules are developed to locate truly missing value cells for any travel activities between each origin and destination.

Figure 1: Representation of OD matrices by trip purposes, times of day and transport modes



3.2 Construct Missing Cells Classification Rule

As mentioned previously, one factor contributing to the sparseness of OD matrices is the non-revelation of information in some cells in order to protect the privacy of households who live in those cells where their total amount of travel in a cell is less than a cut-off criterion (eg < 200 trips).

To overcome this problem of suppressed data, we need to obtain trip length distributions by transport modes and trip purposes for those missing value cells. For Sydney, there are 33,678 missing values cells (with less than 200 trips). Table 2 provides trip length distributions by 5 trip purposes and all transport modes for all missing value cells. For example, for all modes associated with the journey to and from work, there were 4,602 cells out of 33,687 total missing value cells with a mean trip length of 14.25 Km and a standard deviation of 12.78 Km.

Table 2 Trip Length Distributions by Five Trip Purposes and All Transport Modes for Cells with less than 200 trips (Transport Data Centre 2001)

Trip Purposes	Mean Trip Length (Kms)	# of Missing Value Cells	Std. Deviation of Mean Trip Length (Kms)
To and from work	14.25	4602	12.78
To and from education	8.58	1738	10.28
To and from personal business	8.62	2200	11.12
To and from shops	7.23	4693	10.54
Other	11.58	20454	15.87
Total	10.99	33687	14.44

Table 3 provides a more detailed breakdown of trip length distribution by 5 trip purposes and 6 transport modes. By relating Tables 2 and 3 we can explore in more detail the trip length distribution by different trip purposes and transport modes. For example, the 4,602 cells (in Table 2) associated with all modes for the journey to and from work consist of 2646, 456, 750, 360, 270 and 120 cells respectively for car as driver, car as passenger, train, bus, walk and other transport modes.

These trip length distributions by transport modes and trips purposes form the key input data in establishing a classification rule to identify missing value cells out of the total number of zero value cells found in Step 1 above.

Table 3 Trip Length Distributions by 6 Transport Modes and 5 Trip Purposes for Cells with less than 200 trips (Transport Data Centre 2001)

Transport Modes	Trip Purposes	Mean Trip Length (Kms)	# of Missing Value Cells	Std. Deviation of Mean Trip Length (Kms)
Car as driver	To and from work	15.68	2646	13.01
	To and from education	17.59	150	19.23
	To and from personal business	9.34	1226	11.78
	To and from shops	7.78	2589	10.59
	Other	12.65	11633	15.86
	Total	12.22	18244	14.79
Car as passenger	To and from work	11.19	456	11.26
	To and from education	6.42	572	6.46
	To and from personal business	9.39	471	11.07
	To and from shops	8.95	995	13.13
	Other	12.22	5169	17.23
	Total	11.13	7663	15.62
Train	To and from work	18.84	750	14.17
	To and from education	16.17	250	12.47
	To and from personal business	13.86	126	14.44
	To and from shops	14.16	123	13.51
	Other	18.03	720	20.28
	Total	17.60	1969	16.54
Bus	To and from work	8.83	360	5.60
	To and from education	8.24	436	6.24
	To and from personal business	5.87	118	3.78
	To and from shops	5.49	247	4.30
	Other	8.79	614	11.43
	Total	8.01	1775	8.13
Walk	To and from work	2.36	270	2.10
	To and from education	2.23	303	1.98
	To and from personal business	2.07	223	1.54
	To and from shops	2.08	682	1.71
	Other	2.48	1877	4.72
	Total	2.34	3355	3.74
Other	To and from work	8.66	120	6.41
	To and from education	10.38	27	15.36
	To and from personal business	5.30	36	5.04
	To and from shops	6.11	57	6.47
	Other	7.81	441	12.10
	Total	7.79	681	10.80

There are five specific tasks in establishing the missing cells classification rule. The first task is to measure distance between every pair of origin and destination zones. Secondly, a classification scheme for trip lengths by distance values is required. Thirdly, every cell in the total OD trip matrix needs to be classified by their distance between the corresponding origin and destination. Fourth, a zero-value cells database is developed to provide specific detail of every cell that has a zero value in terms of the

total number of trips. Having identified the zero value cells, the challenge is to determine whether these zero value cells have no travel related activity at all (ie genuine zeroes) or the travel values are truly missing. The fifth task involves the use of the trip length distribution of missing value cells as a basis for locating truly missing value cells for any travel activities between OD pairs among a whole set of zero value cells. The following sections provide more detail on each step.

The distance between every pair of ODs is measured as a road network distance in kilometres between the centroids of the origin and destination zones.

By taking the maximum OD distance value in the Sydney study area as a threshold, a classification scheme to classify trip length by distance can then be specified. It includes 31 classes ranging from class 1 within less than 5 kilometres, increasing by an increment of 5 kilometres to class 31 with a distance value equal to or greater than 150 kilometres.

3.3 Classification of every cell of the total OD trip matrix by their trip length distance

This step involves reading the distance between every OD pair and looking up the classification table for a trip length by distance class. A database is constructed to provide the OD distance classification information for every cell in the OD matrix. For the Sydney case study, this database has 817,216 records (904 x 904) and four fields:

- Cell ID. This field is automatically generated by the database as a unique reference. This reference will be used later in the imputation process.
- Origin zone.
- Destination zone.
- Trip Length Distance Class ID.
- Trip Length Distance Range Value in Kilometres.

3.4 Create the Zero Value Cells Database

Given the total number of zero value cells found from Step 1 above, a database is automatically generated in the imputation process to provide a complete enumeration of every identified zero-value cell. The building of zero-value cells database requires two data inputs. The first is the total OD trip matrix generated from the previous step; the second is the OD distance classification of every cell of the total OD trip matrix. The zero-value cell data base contains the following updated data:

- Cell ID. This field will be automatically generated by the database as a unique reference. This reference will be used later in the imputation process.
- Origin zone. This field stores the ID of the origin zone of each zero value cell found.
- Destination zone. This field stores the ID of destination zone of zero value cell found.

- OD Distance Class ID.
- OD Distance value.
- Selected (Y=1/N=0). This field is used to notify whether the current zero value cell has been chosen as a missing value cell.

3.5 Using Trip Length Distributions to Locate Missing Value Cells

Given the total number of missing value cells and the associated trip length distribution of these missing value cells, we can use them to locate missing value cells based on the specific trip length distance between each origin and destination. For Sydney, the trip length distribution for all trips and modes (mean trip length = 10.99 Km and standard deviation = 14.44 Km) for 33,686 missing value cells are generated as shown in Table 4.

Table 4: Trip Length Distribution for Total Trips of Missing Value Cells

Class ID	Trip Length by distance (Kms)	Fx	xFx	Px	# of Missing Value Cells
(1)	(2)	(3)	(4)	(5)	(6)
1	0	0.020681	0.223305	0.223305	7522
2	5	0.025350	0.339137	0.115832	3902
3	10	0.027563	0.472670	0.133533	4498
4	15	0.026583	0.609379	0.136709	4605
5	20	0.022741	0.733673	0.124294	4187
6	25	0.017256	0.834032	0.100359	3381
7	30	0.011615	0.905994	0.071962	2424
8	35	0.006934	0.951818	0.045824	1544
9	40	0.003672	0.977732	0.025914	873
10	45	0.001725	0.990745	0.013014	438
11	50	0.000719	0.996549	0.005804	196
12	55	0.000266	0.998847	0.002298	77
13	60	8.71E-05	0.999656	0.000808	27
14	65	2.53E-05	0.999908	0.000252	9
15	70	6.53E-06	0.999978	7.00E-05	2
16	75	1.49E-06	0.999995	1.72E-05	1
17	80	3.03E-07	0.999999	3.77E-06	0
18	85	5.46E-08	1	7.32E-07	0
.	.	.	.	.	0
.	.	.	.	.	0
31	150	2.08E-22	1	0	0
				SUM=	33686

*Note*s:
(1) ClassID
(2) Trip Length by distance in Kilometres
(3) Fx - Given trip length by distance value, the normal distribution specified by the mean and variance of trip length will evaluate the density curve.
(4) xFx - Cumulative probability (calculated cumulative area under the density curve)
(5) Px - Probability (or proportion) for each class. This value is calculated by subtracting the two successive values in column (4)
(6) Number of Missing Value Cells in each Class – The absolute value for each class is calculated by multiplying column (5) by total number of missing value cells.

Table 4 provides basic information to guide the imputation process in locating all true missing value cells by using trip length distance between each origin and destination. As shown in the last column, we found that given a total number of missing value cells of 33,686, there are 7,522 cells classified as having a trip length distance of less than 5 Km, 3,902 cells with trip length distance between 5 and 10 Km, and so on.

Having identified the missing value cells, the next challenge is to determine the trip generation from these cells given the set of constraints. For the Sydney case study, the two constraints are (i) the number of trips generated by each missing value lies between 1 and 200. The threshold of 200 trips per missing value cell is chosen to represent the cut-off point to maintain privacy for local residences living at any cell with less than 200 trips per day, and (ii) the total number of trips generated by all 33,686 missing cells was 4,664,879 trips per day. In the next section we describe how the trip generation rate assignment is developed to determine the amount of trips generated for every missing value cell.

3.6 Construct Trip Generation Rate Assignment Rule

The objective of this step is to establish a rule for associating the trip generation rate to every missing value cell. This step involves six specific tasks.

The first task is to specify the trip generation rate classification scheme for missing value cells. Second, given a known total number of missing value cells n, we use this value to randomly generate n trip generation rate values for n missing value cells, to construct the empirical distribution of trip generation rates for all missing value cells. We use a uniform distribution in this random process. Thirdly, given the constructed empirical distribution of trip generation rates for all missing value cells, we normalise this distribution to estimate the number of missing value cells associated with every trip generation rate class. Fourth, given the estimate of every missing value cell and associated trip generation rates, we can calculate the number of trips generated by missing value cells in each trip generation class. Fifth, given the constraint of the total number of trips generated by all missing value cells, we can then use it to adjust the number of trips generated by missing value cells in each trip generation class (in step four). Sixth, given the known missing value cells in each trip generation class, we can adjust the number of trips associated with each trip generation class. Table 5 can be used to illustrate these steps using the Sydney data.

3.7 Specification of the trip generation rates classification scheme

Taking into account the threshold of 200 trips per missing value cell, we can specify trip generation rates for Sydney by using 41 classes ranging from class 1 with slightly above 0 trips per cell, increased by 5 trips per class up to class 41 with a maximum threshold of 200 trips per cell (see columns 1 and 2, Table 5).

3.8 Randomly generate trip generation rates associating with missing value cells

Given a known total of 33,686 missing value cells, we randomly generate 33,686 trip generation rates whose value ranges from 1 to 199 trips. These trip generation rate values form an empirical distribution of trip generation rate for each missing value cell.

Table 5 Construct Trip Generation Rate Assignment Rule for Sydney Data

Trip Generation Rate Class ID (1)	Trip Gen Rate Value, x (No.of trips per cell) (2)	Fx (3)	xFx (4)	Px (5)	Est. of No.of MCs (6)	Estimate of No.of Trips associated with MCs (7)	Adjusted estimate of No.of trips associated with MCs (8)	Adjusted trip gen rate value (No.of trips per cell) (9)
1	0	0.001510	0.040149	0.040149	1352	3380	4468	3
2	5	0.001753	0.048297	0.008147	274	2055	2716	10
3	10	0.002019	0.057716	0.009419	317	3963	5238	17
4	15	0.002308	0.068522	0.010807	364	6370	8420	23
5	20	0.002618	0.080827	0.012305	414	9315	12312	30
6	25	0.002947	0.094731	0.013904	468	12870	17011	36
7	30	0.003292	0.110322	0.015591	525	17063	22553	43
8	35	0.003650	0.127673	0.017351	584	21900	28946	50
9	40	0.004016	0.146835	0.019162	645	27413	36233	56
10	45	0.004385	0.167837	0.021002	707	33583	44388	63
11	50	0.004752	0.190680	0.022843	770	40425	53432	69
12	55	0.005110	0.215338	0.024658	831	47783	63157	76
13	60	0.005453	0.241752	0.026414	890	55625	73522	83
14	65	0.005775	0.269832	0.028080	946	63855	84400	89
15	70	0.006070	0.299456	0.029624	998	72355	95635	96
16	75	0.006331	0.330472	0.031016	1045	80988	107046	102
17	80	0.006553	0.362699	0.032227	1086	89595	118422	109
18	85	0.006732	0.395930	0.033231	1119	97913	129417	116
.	.	.	.	.	.	.	.	.
.	.		.	.	.	.	.	.
41	200	0.001520	1	0.048664	1639	327800	433270	264
SUM					33683	3529314	4664876	5551

Note:

(1) ClassID: Trip generation class ID
(2) x – Trip generation rate value generated by missing value cells of the current class. (Number of passenger trips per cell)
(3) Fx – Given x (trip generation rate), a random sampling process is used to draw the outcome leading to the form of the density curve. The random sampling process is based on a uniform distribution.
(4) xFx - Cumulative probability (calculated cumulative area under the density curve)
(5) Px - Probability (or proportion) for each class. This value is calculated by subtracting the 2 successive values in column (4)
(6) Estimate number of missing value cells in each class – The absolute value for each class is calculated by multiplying column (5) by the total number of missing value cells (33,686 cells). The sum of this is equal to 33683 (in comparing to 33686 cells) due to rounding error.
(7) Estimate of number of trips associated with estimate of the number of missing value cells.
(8) Adjusted number of trips associated with number of missing value cells in a particular trip generation rate class

(9) Adjusted trip generation value in each trip generation class.

3.9 Normalise the empirical distribution of trip generation associating with missing value cells

The empirical distribution of the trip generation rates for missing value cells is normalised. The result is shown in columns 3, 4 and 5 in Table 5. Consequently, we can estimate the number of missing value cells associated with every trip generation rate class (see column 6, Table 5) by multiplying the corresponding value in column 5 with total number of missing value cells (33,686 cells).

3.10 Estimate the amount of trips generated by missing value cells in each trip generation class

The expected number of trips generated by the associated missing value cells in each trip class is calculated by multiplying the expected number of missing value cells in column (6) with the mean number of trips associated with the missing value cells. For example, as shown on Table 5, for class ID = 1, the number of missing value cells are 1,352 (in column 6), the mean number of associated trips is (0 trip+5trips)/2 = 2.5 trips (calculated from column 2) and the expected number of trips is 3,380 (=1353 cells @2.5trips per cell). In other words, there are 1,353 missing value cells and 3,380 passenger trips associated with these cells at the lowest end of trip generation range (ie @2.5trips per cell). The calculation continues to reach the highest rate of trip generation (ie @200trips per cell). The value of 200 trips per cell represents the threshold (cut-off point) for non-revelation of household data.

3.11 Adjust the amount of trips generated by missing value cells in each trip generation class

The sum of column 7 in Table 5 represents the estimate of the total number of trips generated by all missing value cells. This value is constrained by the observed total number of trips generated by all missing value cells. The observed value was provided by the Transport Data Centre of Transport NSW. The difference between the estimated and observed values is used to adjust the estimated number of trips generated by every missing value cell. The adjustment is based on the following formula.

$$Adjusted\_ENumofTrips(x) = ENumTrips(x)+(ENumTrips(x)/SumofENumTrips)* DeltaNumTrips$$

where:

EnumTrips(x): estimate of number of trips generated by missing value cells of class x. These values are calculated and shown in column (7) of Table 5 for Sydney.

SumofENumTrips: total number of trips generated by all missing value cells. This value is the sum of column (7) of Table 5.

DeltaNumTrips: the difference between the estimate (SumofENumTrips) and observed total number of trips generated by all missing value cells.

Adjusted_EnumTrips(x): adjusted estimate of number of trips generated by missing value cells of class x. These values are calculated and shown in column (7) of Table 5 for Sydney data.

The result of the adjustment process (Adjusted_ENumofTrips(x)) is shown in column 8 of Table 5.

3.12 Adjust the number of trips associated with each trip generation class

The initial trip generation rate values specified in column 2 of Table 5 need to be adjusted to comply with the total trip generation constraint in each trip generation class. The adjusted trip range for every trip generation class is calculated as follows:

AdjustedTripRange(x) = Adjusted_EnumTrips(x) / ENumofCells(x)

where:

Adjusted_EnumTrips(x): adjusted estimate of number of trips generated by missing value cells of class x. These values are calculated and shown in column (7) of Table 5 for Sydney data.

ENumofCells(x): estimate of number of missing value cells in trip generation class x. These values are calculated in previous task and shown in column (7) of Table 5.

AdjustedTripRange(x) : adjusted trip range for trip generation class x.

Having located missing value cells and associated trip generation rates with each cell, the next step is allocate them to different segments of the OD matrices. For Sydney, the segments include trip purposes, transport modes and time of days. The next section describes how the allocation rule is constructed to serve this aim.

3.13 Construct the Missing Value Cells and Trip Allocation Rule

With the complete information on all missing value cells and associated trip generation rates, we are able to address the segmentation requirements of the household travel data. Given the availability of distributions of all missing value cells by transport modes and trip purposes for Sydney (see Table 6), the trip allocation rule uses this knowledge in allocating missing value cells and associated trip generation rates to these segments. As an illustration of the operation (see Table 6), the allocation rule allocates 2646, 150, 1226, 2589 and 11633 missing value cells and associated trip generation rates to OD matrices of car as driver and five different trip purposes in the order indicated in Table 6. The process continues until all corresponding missing value cells in the OD matrices for all transport modes and trip purposes are allocated.

Finally, the allocation of missing value cells to the selected times of day dimension of all OD trip matrices is completed by using proportional fitting to the existing OD trip matrices.

Table 6: Distribution of 33686 Missing Value Cells by Transport Modes and Trip Purposes for Sydney data (source: New South Wales Travel Data Centre, 2001)

	Trip Purposes				
Transport Modes	To and from work	To and from education	To and from personal business	To and from shops	Other
Car as driver	2646	150	1226	2589	11633
Car as passenger	456	572	471	995	5169
Train	360	436	118	247	614
Bus	750	250	126	123	720
Walk	270	303	223	682	1877
Other	120	27	36	57	441

Conclusions

This paper has presented a method for imputing missing trip data using expanded sample data from a household travel survey. It is commonly observed that such data expanded up to the population is deficient as levels of spatial details required for area-wide transport planning applications. Given our interest in developing trip matrices for a number of modes, trip purposes and times of day at a high level of spatial representation, the initial set of trip tables are extremely sparse. In the Sydney context there were over 67% of cells with missing data (either being true zeroes or genuine non-zero trip activity). The paper presents a practical method for imputing OD matrices using Sydney data as a case study. A spatial and statistical approach was used to answer following questions:

- How to identify the degree of sparseness of OD matrices?
- How to identify all missing value cells from zero travel activity cells?
- How to assign different trip generation rates to different missing value cells?
- How to allocate missing value cells and associated trip generation rates to different travel segments (eg by trip purposes, transport modes and time of days)?

An object-oriented software system written in C++ was developed by the authors to automate the imputation process. As an initial effort in the imputation of OD matrices derived from the Sydney household travel survey, the research has revealed ongoing issues for investigation. One particular research area is the implications of this work on influences contributing to the degree of sparseness of OD matrices such as sample size (a consequence of cost constraints), segment requirements, the spatial resolution of a geographical zoning system and the threshold value (cut-off criterion) used in the non-revelation of information in order to protect the privacy of households.

Acknowledgments

We thank Mr Tim Raymond of the Transport Data Centre of Transport NSW for providing constructive discussions on the data used in Tables 2, 3 and 6 of this paper.

References

Transport Data Centre (2001) Mean Trip Length Distribution Data of Missing Value Cells segmented by transport modes and trip purposes in Sydney Household Travel Survey, Transport NSW, Sydney.