BUILDING OD MATRICES FROM A WIDE VARIETY OF DATA SOURCES – EXAMPLE OF THE WEST MIDLANDS STRATEGIC TRANSPORT MODEL

Tom van Vuren
Mott MacDonald
Philip Clarke
Peter Davidson
Peter Davidson Consultancy

1. INTRODUCTION

Large scale transport models are making a come-back in England, if only to support the increasing demands for quantitative analysis for the monitoring of Local Transport Plans. Following guidance from the Department for Transport, it is expected that many of these models will be multi-stage demand response models, operating on a pivot-point basis. As with more conventional network models used in England, the base year matrix and its validation using real-life observations will determine the credibility of the overall forecasting system. Techniques are required to estimate OD matrices efficiently using the wide variety of available data sources containing information about origin and destination patterns.

This paper describes the work undertaken to develop a new matrix for the West Midlands metropolitan conurbation, with a population of around 5M, centred on Birmingham, concentrating on the highway matrices for private car[1]. A number of data sources were available: origin – destination survey data from roadside interviews and city centre car park surveys, household interviews, airport passenger surveys and partial trips from number plate recognition cameras, with a large number of supporting data such as traffic counts and journey time surveys. These data sources were successfully merged together using the ERICA4 approach. The methodology is based on variance weighting, which is the first time this method has been tried on such different data sources in combination. This paper describes the methodology used, showing strengths, weaknesses and lessons learned.

2. USING DIFFERENT DATA SOURCES

2.1 Concepts

Origin-destination data are expensive and often disruptive to collect. As a result, many studies have resorted to updating existing, outdated OD matrices on the basis of cheaper traffic counts. Sometimes, part of the matrix is updated using selected OD surveys, such as roadside interviews.

In essence, the existing matrix is used as a prior estimate of the 'best' matrix, and new data are introduced only to improve the model fit to changing circumstances, without proper reflection on the reliability of each of the data sources.

[1] Public transport matrices were readily available from a different source – but the techniques described here are applicable also to PT matrix estimation.

The reliability of travel data is related to the sample rate that an interview has been able to achieve. For example, trip patterns observed in a roadside interview with a 10% average sample rate will be more reliable than those derived from a household survey with a sample rate of, say, 1-2%. This reliability can be reflected through the index of dispersion for each cell in the matrix, and statistically soundly calculated using the Variance Weighting technique described in Appendix 1.

2.2 Data sources

The OD estimation application in the West Midlands had access to a large number of data sources. The key characteristic of these data sources is that they all describe travel patterns at the OD-level, and therefore contain direct information for the estimation of origin-destination matrices:

- household interviews (approx. 1% sample rate)
- roadside interviews from around 200 sites, including all motorway on-slips in the region (approx. 10% sample rate)
- car park surveys (approx. 6-7% sample rate)
- airport surveys (approx. 5% sample rate)

In addition, traffic counts were available for estimation. These contain indirect, aggregate information on OD movements, as they do not identify individual matrix cells, but grouped values. When individual counts are used, the OD flows contributing to each count must be established from model assignments, whereas cordon or screenline flows provide information on trip ends and hence on total in- and outgoing OD movements for groups of cells.

Cordon and screenline count data were available for all the roadside interview sites, with additional information for each of the urban centres in the area.

2.3 Alternative approaches

Alternative methodologies are available for estimating matrices from large and diverse datasets. In Gunn et al (1999) the so-called combined calibration method is described. The method uses similar principles to estimate the OD matrix from a variety of data sources, properly reflecting the statistical quality of each, but with an important difference: the combined calibration method aims to improve a prior matrix (typically generated by a transport model) using direct and indirect information, whereas the ERICA approach estimates a matrix directly from surveys, which may then be improved further by more traditional matrix updating or improvement techniques (as described by eg. Van Zuylen and Willumsen, 1982).

The combined calibration technique, as described by Gunn et al, takes the modelled OD matrix as it starting point and focus and seeks to use real-life observations to improve this. This approach is more appropriate when the observations available are limited in scope, and has added advantages when applying the resulting matrix in pivot point operation (as the structure of the transport model is embedded in the prior matrix). The ERICA4 approach described here focuses on the direct observations to generate as good as possible an observed matrix, which requires substantial and reliable data sources. Hence, the ERICA4 approach is most suited to a data-rich environment, and when the model to which the base matrix will be applied is not (yet) available.

2.4 Advantage of Database Approach

Due to the cost of origin-destination surveys, many OD matrix estimation applications estimate a matrix for a specific application and year, and update this over time using ad-hoc additional surveys and easily obtainable indirect observations (mainly counts) in future years. There are many OD matrices in existence that have their heart in base data collected over a decade ago or more… The ERICA4 approach differs in essence, in that the method is based around a database of OD surveys, which can be expanded with new observations, after which a fully new matrix can be estimated, again acknowledging sample rates of each of the (existing and new) data sources in the database. Hence, the investment in the original data can be preserved.

In the case of the West Midlands more than €4M was invested in data collection, and preservation of that investment was important

3. APPLICATION TO ROADSIDE INTERVIEWS

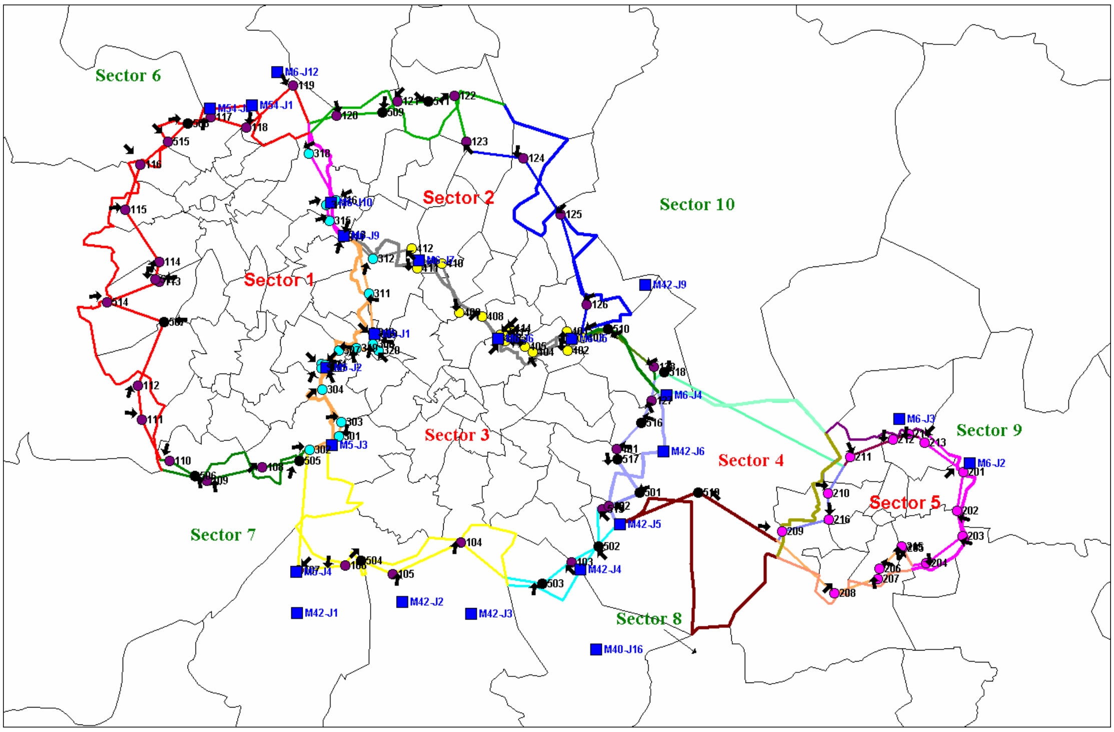
Roadside interview data were the base source for building the new matrices for the West Midlands – as their sample size and geographical spread are greatest. RSIs are used in ERICA4 by combining them into cordons and screenlines. There are two cordons and two screenlines in the RSI data as shown in Figure 1:

- Cordons around the West Midlands Metropolitan County and Coventry;
- Screenlines along the M5 and M6 Motorways.

The cordons and screenlines naturally divide the West Midland County into 5 geographical areas - 5 internal sectors (sector 3 is Birmingham and sector 5 is Coventry).

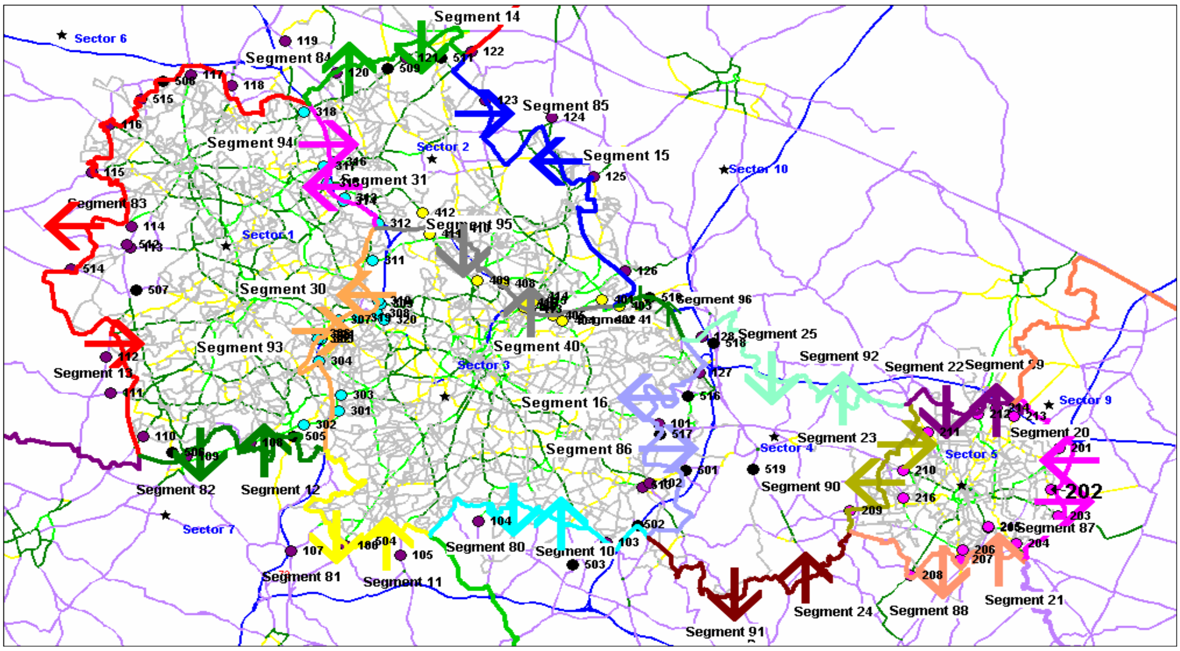
Figure 1 also shows that there are 5 external sectors from sector 6 to sector 10 that cover the remaining areas of the country.

Figure 1 : Sectors, cordons, and screenlines



A screenline is divided into segments when other screenlines or cordons cross through. The part of the M6 in the West Midlands is divided into two segments by the M5 into segments 94 and 40 (eastbound), and segments 31 and 95 (westbound). Each segment refers to specific RSI sites; for example, segment 94 represents RSI sites 313^{T}, 315, 316^{T}, and 318.

Figure 2: Screenline Segments



© Association for European Transport 2004

3.1 Positive and negative trips

ERICA4 utilises the concept of positive trips and negative trips to tackle the problem of double counting. For origin sectors any inter-sector trips have their outbound cordons in the positive direction and inbound cordons in the negative direction. For example, for the movements from sector 3 to other sectors, any trips picked up by RSI sites on segments 30 and 40 are positive as they are on the outbound cordon of sector 3, but any trips picked up by RSI sites on segments 93 and 95 are negative as they are on the inbound cordon of sector 3.

Each ERICA4 run generates a positive matrix and a negative matrix containing trips in positive directions and negative directions respectively. A net matrix is obtained by taking away a negative matrix from a positive matrix. Theoretically all cell values should be positive or 0, but in practice negativity always occurs due to the nature of random sampling. The negatives in the net matrix are called residual negatives.

Checking the negative and residual negative trips is at the heart of the ERICA4 process and it is directly related to the quality of the generated matrices.

3.2 Building the RSI Matrices

Three different types of roadside interview were carried out (for a number of reasons, not all connected with this project):

- RSIs on an external cordon around the conurbation (JDT RSI non-motorway)
- RSIs on roads approaching Motorway junctions, creating Motorway screenlines (JDT RSI motorway)
- RSIs on Motorway on- or off-slips, aimed at capturing a full Motorway matrix (HA RSI)

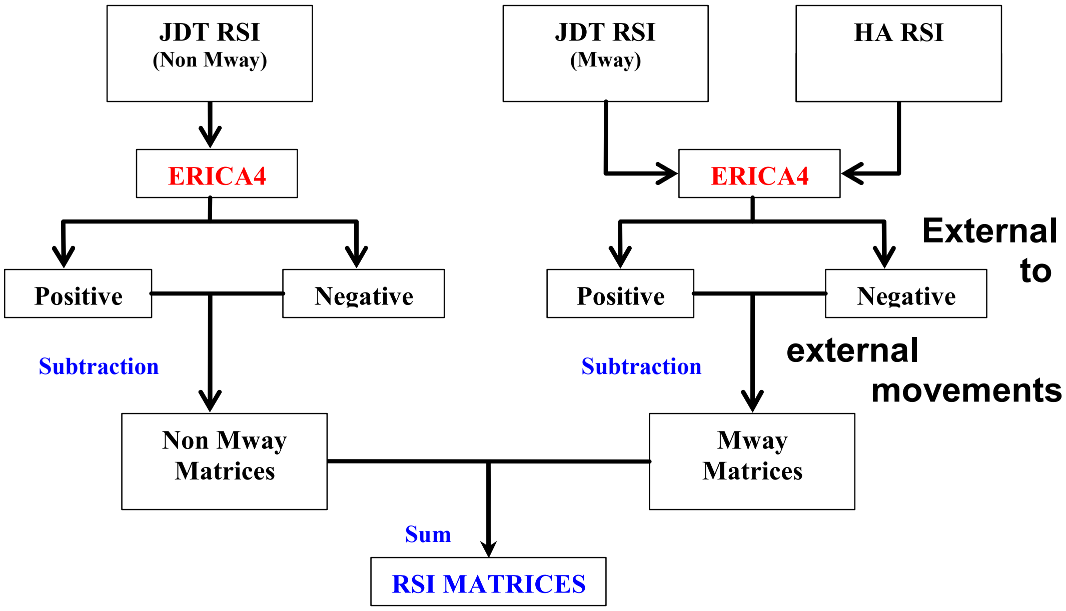
The latter 2 data-sets substantially overlap, and double-counting was a serious issue for the methodology to contend with. The RSI matrices were built in two parts, one with just the JDT RSI (non motorway) and one combining both JDT RSI (motorway) and HA RSI matrices. Figure 3 shows the schematic diagram of the matrix building process of the RSI matrices. The steps of the process are as follows:

- The ERICA4 matrix building option was run to read the trip records of the datasets and accumulate the trips and variances for each cell in the matrix. It would output two trench files, one with positive (trip matrix) and one with negative (‘wiggly trips’) matrix cells.
- The negative matrix cells were subtracted from the trip matrix to eliminate the possible ‘wiggly trips’ that cross the boundary more than once.

- This procedure was carried for both the non-motorway and motorway matrices

- The two matrices were combined together to produce the final RSI matrices.

Figure 3: RSI Matrix Building Schematic Diagram



3.3 External Movements

External-to-external movements are only partially observed. A special approach was developed to deal with these trips. The methodology is illustrated in Figure 4, with the aim of trying to pick up as many external-to-external trips as possible crossing through the West Midlands.

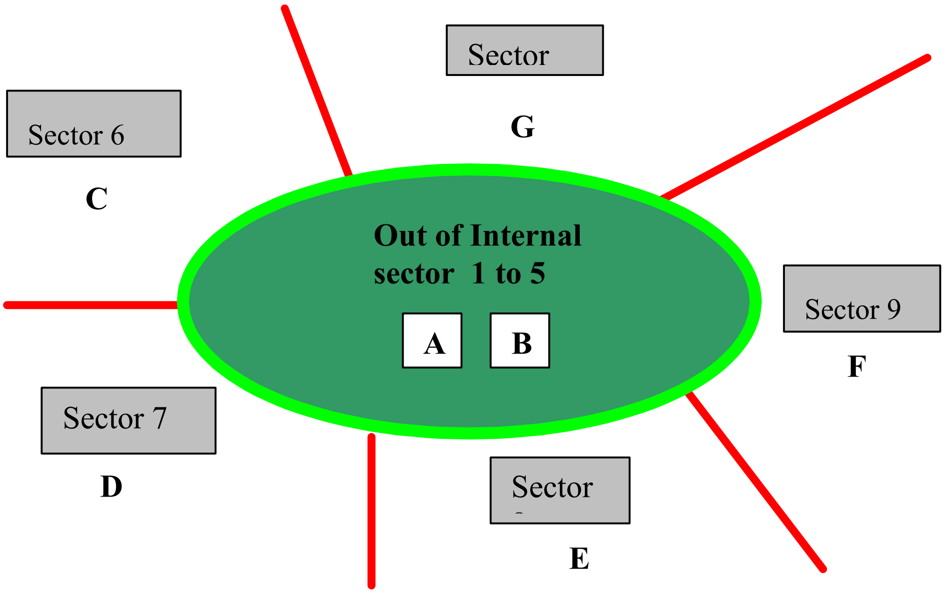
The motorway and non-motorway survey sites were considered together when dealing with external-to-external trips. The 5 internal sectors were grouped together into one internal sector “1 to 5” labelled A and B and the remaining 7 outbound cordons as C, D, E, F, and G shown in the figure below.

- Outbound cordon A: the set of all screenline segments at the boundary of the internal sector “1to5” with the 5 external sectors;

- Outbound cordon B: the set of motorway on-slip sites in the internal sector “1to5”.

- Outbound cordon C, D, E, F, and G: each represents the set of motorway on-slip sites in the corresponding external sector from 6 to sector 10, respectively.

Figure 4: External to External Movements



Sector 6 to all other external sectors:	outbound cordon=A+B+C
Sector 7 to all other external sectors:	outbound cordon=A+B+D
Sector 8 to all other external sectors:	outbound cordon=A+B+E
Sector 9 to all other external sectors:	outbound cordon=A+B+F
Sector 10 to all other external sectors:	outbound cordon=A+B+G

4. MERGING OTHER COMPONENT MATRICES

The other components to the overall highway matrix are the car park matrix, the airport matrix and the household interview (HHI) matrix. The car park and airport matrices were estimated using very similar procedures as described for the roadside interviews. The calculation of expansion factors for the car park matrix is complex, as three different types of parking must be accounted for, each of which had been surveyed differently:

- Off-street parking
- On-street parking
- Private non-residential parking (PNR)

Only off-street and on-street car park surveys had been carried out, in the main urban centres in the conurbation. PNR trips were derived indirectly from these sources.

4.1 Calculation of Expansion Factors for Car Parking Surveys

The calculations of expansion factors for observed and unobserved car park survey data was done as follows[2]:

Off-Street Data

- The observed data, i.e records, counts and capacities of the sites of a particular model zone were grouped together.
- The observed percentage arrival rates by time period for each zone were calculated.
- We identified the unobserved capacity for each zone and factored the calculated observed percentage arrival rates to the unobserved capacity to derive the predicted unobserved arrival flow.
- We combined the observed and unobserved arrival flow.
- Finally, we calculated the expansion factor for each time period and applied it to the records.

PNR (Private Non Residential)

No specific interviews had been undertaken within PNR car parks, due to costs and privacy issues. The methodology adopted to estimate PNR trips is as outlined below.

1. All O-D surveys by urban centre were grouped together and trips with destination purpose work and employer's business were selected.
2. Factors were calculated for each urban centre. These factors were based on selected long stay car parks and were based on the proportion of trips with these 2 purposes exiting and entering each zone for each hour of the day between 0700 and 1900 hours.
3. PNR capacities for each zone were available from a separate desktop study using proxies such as floor-space or numbers of workers, with the exception of Birmingham where a dedicated field-study had been undertaken.
4. The commute and business factors calculated in 2. were applied to the estimated zonal capacities to create PNR trips.
5. The OD pattern was assumed identical; expansion factors were calculated for each hour separately.

[2] First, five additional parking sectors were created, one in each of the internal sectors, making the total 15.

On-Street Data

1. Surveys were carried out in parts of the centres only
2. In addition, we identified the non observed on-street capacities by area. These areas were chosen so as to represent consistent parking areas.
3. The observed data were grouped to zonal level.
4. The observed percentage arrival rates by time period were calculated and used to derive the predicted unobserved arrival flow.
5. Expansion factors were calculated for each of the interview records based on the total arriving flows, by time period, as calculated in 4.

4.2 Household Interview Data

The value of household interviews is that in principle all trips, however short, should be captured, unlike roadside interviews which have a bias towards longer trips. Hence, HHI data are particularly relevant for intra-sector movements, which were not observed by the RSIs, and only partially by the car park surveys.

Unfortunately, because of the costs, sample rates for household surveys tend to be low (around 1%), whilst also the sampling approach tends to lead to clustered spatial patterns.

In such situations, the modeller has 2 options:

- Try to estimate observed matrices directly from the data, accepting that a lumpy resulting pattern is likely – and dealing with the unobserved parts of the matrix through partial matrix techniques (as described by e.g. Kirby (1979));
- Estimate a model from the household data and use the model to create synthetic matrices that represent the information in the household data, but applied across the whole of the study area.

The second option was used – as a tour-based model was developed in parallel as the main demand response component of the PRISM strategic transport model[3]. The 2 questions to answer were how to convert the tour matrices to OD matrices and how to calculate appropriate variances for the synthetic cells for use in the merging process.

Conversion to O-D Matrices

[3] The reader may argue that, if a relevant model was available, the combined calibration method could have been employed. We believe that this would not have done justice to the amount of direct OD-data that had been collected, but a comparison in future would be very interesting.

Each of the production-attraction all day tour matrices was converted into individual O-D peak period (AM-0700-0930; interpeak-0930-1530; PM-1530-1900 & off peak-1900-0700) matrices. 'Time of day' factors calculated from the household surveys (Table 1) were applied to the tour matrix and its transpose. These matrices were combined to produce O-D matrices for the respective time periods. Below is an example of the conversion of an all-day tour matrix to an O-D AM period matrix.

Business_AM = [a*(HB-EmpB) + b*(HB-Empb_T)] + [c*(NHB-EmpB)]

Where

Business_AM - Business matrix for the AM period

HB-EmpB - Home-based Employer's Business all day tour matrix

HB-EmpB_T - Transpose of the HB-EmpB tour matrix

NHB-EmpB - Non-home based Employer Business all day tour matrix

a, b and c - Time of day factors for the tour matrices

Table 1: Time of day factors for tours by purpose

Tour Purpose		**AM**	**IP**	**PM**	**OP**
HBWork	**TP out**	0.704	0.120	0.023	0.153
	TP return	0.005	0.189	0.703	0.103
HBEmpBus	**TP out**	0.540	0.331	0.059	0.071
	TP return	0.016	0.339	0.575	0.070
NHBEmpBus	**TP**	0.150	0.666	0.167	0.017
HBEduc-Secondary	**TP out**	0.956	0.040	0.003	0.001
	TP return	0.000	0.292	0.703	0.005
HBEduc-Tertiary	**TP out**	0.532	0.429	0.035	0.005
	TP return	0.261	0.338	0.359	0.042
HBShop	**TP out**	0.104	0.823	0.059	0.014
	TP return	0.007	0.768	0.190	0.035
HBOther	**TP out**	0.170	0.543	0.159	0.128
	TP return	0.060	0.416	0.280	0.244
NHBOther	**TP**	0.102	0.542	0.281	0.075

Calculation of Variances

The synthetic matrices produced by the PRISM model do not have variances attached. As the model was based on the information obtained from the household interviews, the assumption was made that the representative sample rate for the synthetic cells values was the same as achieved in practice in the HHI survey: 1%. After the transfer process from tours to origin-destination trips was completed, these variances were attached to the trip matrices.

4.3 Final merging process

Merging of the different matrices was undertaken at sector-level, to enable a distinction between contributions of each of the component matrices in different parts of the study area, as follows:

Intra-Sector Movements

General intra-sector movements:

(1) HHI component matrix

Intra-parking-sector movements:

(1) HHI component matrix

(2) Parking component matrix

Intra-airport-sector movement:

(1) HHI component matrix

(2) Airport component matrix

Inter-Sector Movements

General sector to other sector movements:

(1) RSI component matrix

(2) HHI component matrix

Movements from parking sectors to other sectors:

(1) RSI component matrix

(2) HHI component matrix

(3) Parking component matrix

Movements from Airport sector to other sectors:

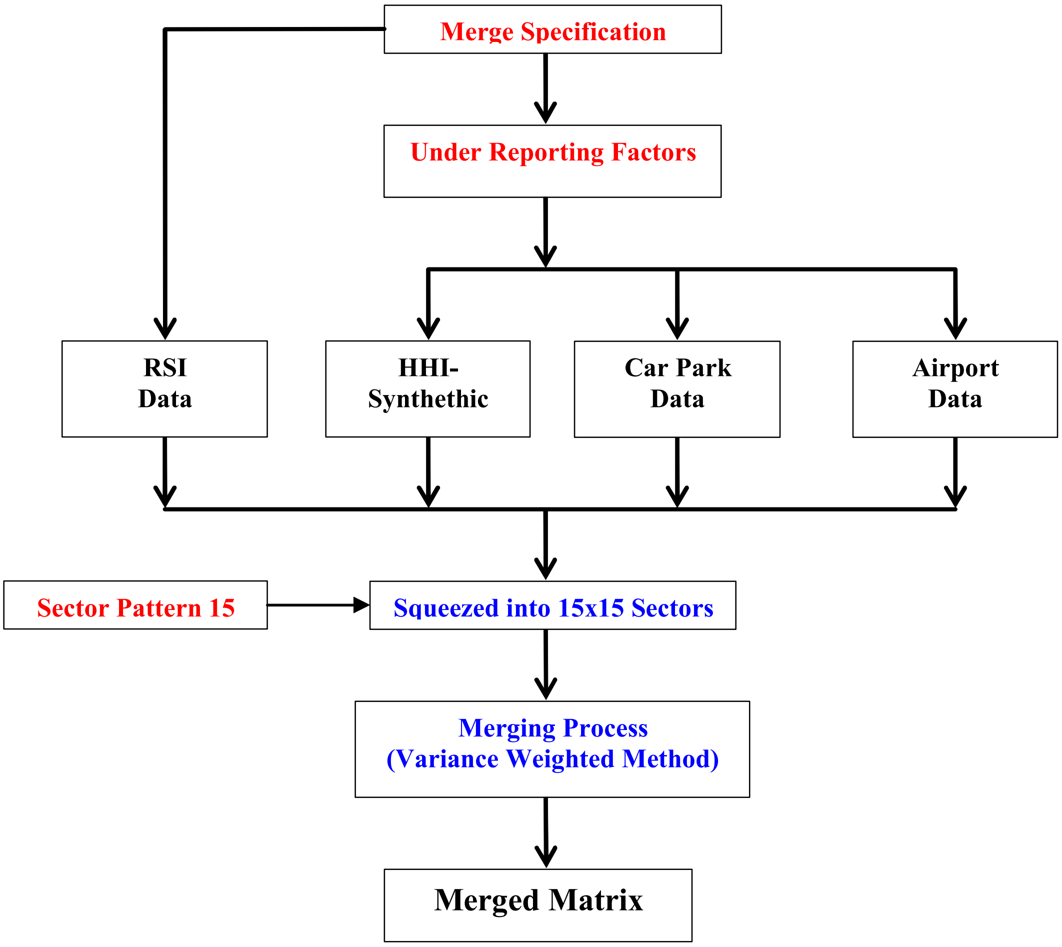
(1) RSI component matrix

(2) HHI component matrix

(3) Airport component matrix

Figure 5 shows the final merging process of all component matrices, reflecting the above secor-based approach.

Figure 5: Final merging process



5. FINAL MATRIX IMPROVEMENTS

The merged matrix was compared with traffic counts on the cordons around the 5 sectors, and on the cordons around the 9 urban centres, shown in Figure 6. It is important to remember that the sector cordons are based on roadside interview sites, whereas the urban cordons had not played a role in the merged matrix estimation at all.

Using acceptance criteria of observed flow +/- 10%, we can see in Table 2 (before TFLOWFUZZY) that the fit to the sector cordons is acceptable, but that there is a structural shortfall at the urban centre cordons. The good fit at the sector cordons is not surprising: the RSIs are the major source for the matrices, with the largest sample sizes. The shortfall at the urban centres indicates a lack in short intra-sector trips, which were estimated from the alternative data sources, with considerably less reliability.

Table 2: Effect of TFLOWFUZZY on matrix fit to observations

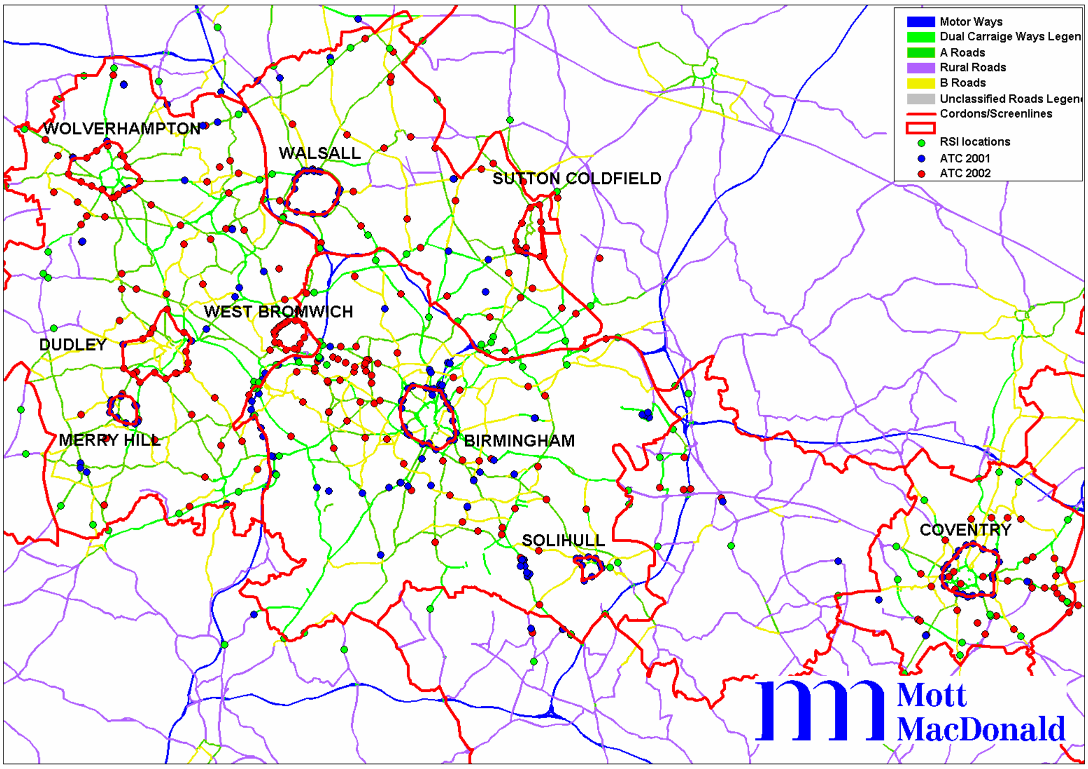
		before TFLOWFUZZY	after TFLOWFUZZY
Sector 1	in	100%	101%
	out	98%	101%
Sector 2	in	97%	104%
	out	105%	104%
Sector 3	in	99%	103%
	out	86%	102%
Sector 4	in	82%	102%
	out	72%	98%
Sector 5	in	95%	101%
	out	90%	101%
Birmingham	in	86%	98%
	out	48%	92%
Coventry	in	64%	95%
	out	68%	96%
Dudley	in	92%	102%
	out	78%	100%
Merry Hill	in	72%	92%
	out	61%	93%
Solihull	in	54%	96%
	out	77%	98%
Sutton Coldfield	in	68%	96%
	out	74%	90%
Walsall	in	76%	97%
	out	78%	90%
West Bromwich	in	84%	92%
	out	62%	87%
Wolverhampton	in	93%	104%
	out	68%	101%

The matrices were compared in terms of purpose split against the Department for Transport's TEMPRO database, and also against observed splits at the more urban roadside interviews; no bias in any of the purposes could be detected, so that it was decided to apply a matrix improvements from counts procedure with the following objectives:

- Improve the fit to observations at the urban centre cordons
- Increase the number of short distance trips in the matrix
- Maintain the fit at the sector cordons

ptv's TFLOWFUZZY procedure was applied. The procedure differs from more common entropy maximisation techniques in that it treats the counts not as constraints but as fuzzy sets with varying bandwidths. The bandwidths reflect differing levels of confidence in the count data, but based on subjective valuation rather than statistical calculations as used in the Variance Weighting procedure.

Figure 6: Urban Centre Cordons



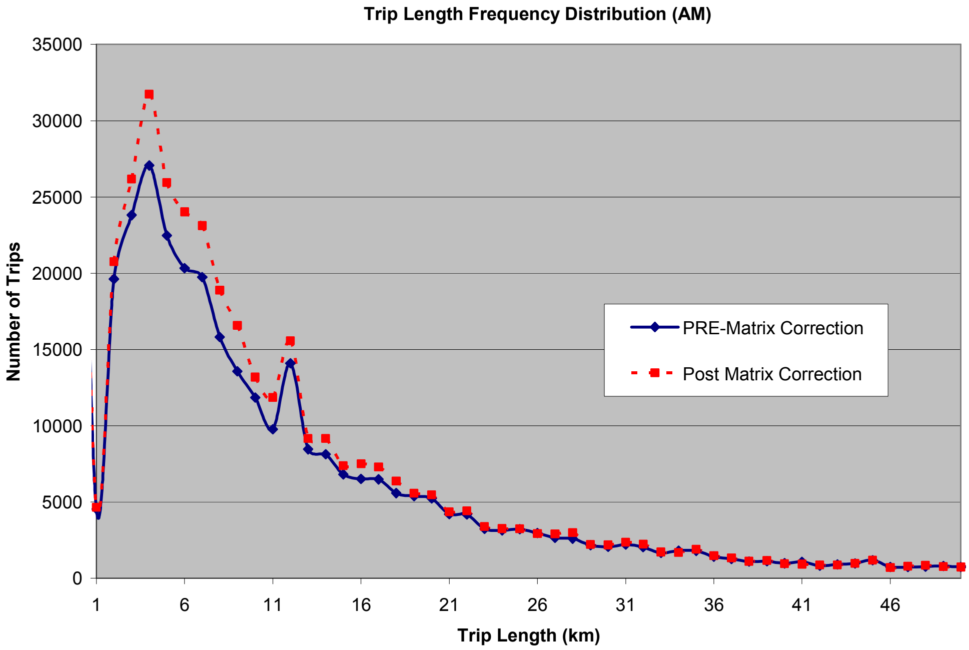
We can observe in table 2 the following:

- Flows on nearly all the sector cordons have increased, so that at all cordons except one the modelled flows exceed the observed values
- Flows on the urban centre cordons have increased to the extent that all cordons except one meet the acceptance criteria

Although we may consider there to be some structural bias on the sector cordons, we accepted this as an inevitable by-product of the significant increases in trips on the urban centre cordons. To support this, an analysis was carried out of the changes in the trip length distribution before and after the TFLOWFUZZY application.

Figure 7 shows that the procedure indeed increased the short-distance trips, with no significant changes in the trip numbers longer than about 20 km. Note that the 'blip' in the 12 km band is a peculiarity in the model area, not a matrix error.

Figure 7: Trip length distribution before/after TFLOWFUZZY



The procedure was considered to be successful, and the final matrix improvements delivered base OD matrices (4 time periods, 4 purposes) that:

- Were firmly based on origin-destination observations;
- Accounted for the statistical reliability of the data sources;
- Contained some of the structure of the transport model to which they were to be applied (although the weight of this part of the matrix was not great due to the small sample size of it household interview source data);
- Fitted observations well.

6. CONSIDERATIONS IN DETAILED APPLICATIONS

6.1 Modification of expansion factors

Even though the average sample rate for roadside interview is 10%, at individual sites or during parts of the day this rate drops, particularly for less common vehicle types, and hence the expansion factors increase. This problem is exacerbated if expansion factors are calculated for short time slices (say, 30 minutes). A suitable correction procedure in those situations is as follows:

- Select records of the same vehicle type from adjacent time slices;
- Copy and paste back to the time period in question (as a rule of thumb copied records should always be taken within the same time period

(AM peak, off-peak etc.) and from a period of expansion factor less than 10);

- Continue until expansion factors are within acceptable bounds.

The assumption made is that the travel patterns in adjacent time slices are a good approximation for the non-interviewed population. The effect of reduced expansion factors is a less lumpy resulting OD matrix.

However, care must be exercised when applying this method, artificially boosting sample rates in individual time periods: no extra information is added, but the variance, as calculated in Appendix 1, is artificially reduced, and hence also the index of dispersion. As a result, too much weight is give to these data in the merging process, unless the variances are corrected back to their original value for the boosted time periods.

6.2 Underreporting Factors

Underreporting factors are factors applied to relevant component matrices prior to merging, so as to bring all the matrices to a consistent level and to build a statistically coherent final matrix. These factors were also applied to variances of each component matrix to create new sets of variances.

The RSI data was chosen as the base data as it is the most fully observed data source. The under-reporting factors for the household interview, car parking and airport matrices were calculated against it. Underreporting factors were calculated for each time period, vehicle type and purpose.

The RSI and other component matrices were 'squeezed' to a sector-level matrix by ERICA4.

1. Relevant sector-to-sector movements of the individual matrices were selected. For example, the sector-to-sector movements to and from the car park sectors were selected in both the RSI and car park matrices to calculate the the car park underreporting factor.
2. The formula used in the calculation of the factors for car park, HHI and airport datasets are:

   - <u>Car Park</u>

   The car parking underreporting factor is calculated based on all trips from and to the 4 parking zones. The formula is:

$$\mu_{CP2RSI} = \frac{\text{sum of the cell values from - and - to the 4 parking zones in RSI matrix}}{\text{sum of the cell values from and - to the 4 parking zones in parking matrix}}$$

   - <u>HHI</u>

   The HHI underreporting factor is the ratio between the total trip ends in the RSI matrix and the total trip ends in the HHI matrix (excluding intra-sector trips because the RSI matrix does not contain any intra-sector movements):

$$\mu_{HHI2RSI} = \frac{\text{total trip ends in the RSI matrix}}{\text{total trip ends in HHI matrix excluding intra - sector trips}}$$

- Airport

The airport under-reporting factor is the ratio between the two totals in the two datasets for both to-and-from airport movements:

$$\mu_{BAX2RSI} = \frac{\text{Sum of from and to Airport trips in the RSI matrix}}{\text{Sum of from and to Airport trips in the Airport matrix}}$$

7. CONCLUSIONS

In this paper we have described some of the most interesting elements of the procedures developed for estimating base matrices for the new West Midlands Strategic Transport Model PRISM (see Van Vuren et al, 2004). The ERICA4 technique applied is founded in statistical theory and reflects properly the reliability of each of the data-sources through variance weighting.

In the application lessons have been learned for future applications; many of these are on the operational end of the spectrum, concerned with data handling when matrix cells become sparse (fine zoning system, different time periods and vehicle types).

A major advantage of the procedure is that a database of sources, properly expanded and with variances attached, has been developed and is now available for future use, when new data sources such as new roadside interviews, or Census Journey To Work data become available. This protects the investment made, in data and processes.

The method, though using a substantial amount of direct observed data, was not able to estimate a trip matrix that fitted the observed data (counts) to a satisfactory level. A TFLOWFUZZY matrix enhancement step using counts was required to deal with the final refinement of the base matrix to acceptable standards. However, comparisons of the matrix before and after the TFLOWFUZZY step show that the structure of the observed matrix remains intact.

8. ACKNOWLEDGEMENTS

The matrix development procedure described in this paper was in support of the PRISM West Midlands strategic transport model, and carried out between 2002 and 2004 by Mott MacDonald and Peter Davidson Consultancy on behalf of the 7 West Midlands Metropolitan Authorities, the Highways Agency and CENTRO (the West Midlands Public Transport Executive). The process was overseen by a client-side Project Management Group involving Colin Eastman, Rebecca Ellison, John Hawkins, Chris Haynes, George Parsons and Bryan Whittaker, and we gratefully acknowledge their support. Much of the detailed work was carried out by Yanling Xiang, Firuz Sulaimi, Serbjeet

Kohli, Andri Heriawan and Paul Hoad of Mott MacDonald. The views expressed in this paper are those of the authors alone, and cannot be ascribed to any of the sponsoring organisations or their officers.

9. REFERENCES

Gunn, H, Mijjer, P, Lindveld, K and Hofman, F (1999) Estimating Base Matrices: The Combined Calibration Method, European Transport Conference, Cambridge,

Kirby (1979) Partial Matrix Techniques, Traffic Engineering and Control, Vol 20, No 8/9, pp 422-428

Van Vuren, T, Gordon, A, Daly A, Fox, J and Rohr, C (2004) PRISM: Modelling 21st Century Transport Policies In The West Midlands Region, European Transport Conference, Strasbourg

Van Zuylen, HJ and Willumsen LG (1984) The Most Likely Trip Matrix Estimated From Traffic Counts, Transportation Research Vol 14B, No 3, pp 281-293

APPENDIX: VARIANCE WEIGHTING

The variance weighting technique takes origin-destination (O-D) trip records and accumulates for each zone-to-zone movement the number of trips, the number of surveyed records upon which the trips were based and the variance of the trip estimate.

The number of trips is the sum of the expansion factor (*e*) over the whole dataset. The number of records is the number (*n*) of trip records and the variance is a function of the expansion factor. This is consistent with the statistics developed by the UK DfT funded research underpinning the ERICA4 software, used to gauge the relative precision of each estimate when merging data from different data sources.

The mathematical formula used in the software is given as following:

Total number of trips for cell ij (T_{ij})

$$T_{ij} = \sum_n e_{ij}$$

Variance Var(T_{ij}) associated with T_{ij}

$$Var(T_{ij}) = \sum_n e_{ij}(e_{ij} - 1)$$

Index of dispersion (I_{ij})

$$I_{ij} = Var(T_{ij}) / T_{ij}$$

where

e = Expansion factor for each recorded journey

n = Count of the number of recorded journeys from origin i to destination j

The merging of matrix cells from different data sources is applied in a pair-wise and consecutive fashion. For 2 sources the merged cell trip value is:

$$T_m = \frac{T_1 I_2 + T_2 I_1}{I_1 + I_2}$$

and the merged cell index of dispersion (i.e. the variance over mean)

$$I_m = \frac{I_1 I_2}{I_1 + I_2}$$

where T_1 and T_2 = the cell trip value from matrix 1 or 2, and

I_1 and I_2 = the index of dispersion for these cells