ATKINS

Specialist Professional and Technical Services Framework

Task 151:
Data Fusion Reporting

Technical Report

Highways England

30 November 2017

Notice

This document and its contents have been prepared and are intended solely for Highways England's information and use in relation to Task 151 of the Specialist Professional and Technical Services Framework

ATKINS Limited assumes no responsibility to any other party in respect of or arising out of or in connection with this document and/or its contents.

This document has 109 pages including the cover.

Document history

Job number: 5120184			Document ref: Data Fusion Technical Report - v2.0			
Version	Purpose description	Originated	Checked	Reviewed	Authorised	Date
1.0	Draft for comment	ACM	RR/RS	JB	ACM	05/10/17
2.0	Reflecting HE comment	ACM	RR	ACM	ACM	30/11/17

Client signoff

Client	Highways England
Project	Data Fusion Reporting
Document title	Technical Report
Job no.	5120184
Copy no.	
Document reference	Data Fusion Technical Report - v2.0

Table of contents

Tables

Figures

1. Introduction

The work documented in this report was initially undertaken on behalf of the Highways Agency (HA) under Lot 2 of the Framework for Transport Related Technical and Engineering Advice as *Task 145: A1 Data Fusion*. The work was led by Atkins with support from a broad supply chain. It builds upon earlier work on OD data fusion undertaken for the HA and documented in Skrobanski *et al*.[1]. Equations presented in that earlier work are reproduced herein without details of their development/proof.

Further work was undertaken under the subsequent *Task 386* and mostly recently the work has been completed for what is now Highways England (HE) under Lot 1 of the Specialist Professional and Technical Services Framework, as *Task 151 Data Fusion Reporting*.

1.1. Background

1.1.1. Prior Matrices and Data Sources

For a whole host of reasons - planning, option development, scheme justification, operational, economic and environmental appraisal, monitoring and evaluation – it is often necessary to develop traffic models which reflect current observed conditions and can be used to forecast the effects of different infrastructure or operational arrangements in a variety of alternative scenarios. Such models are costly, complex and time consuming to develop and often represent a significant programme risk to HE as a scheme promoter. One of the significant tasks required during model development is the development of prior[2] matrices of travel demand. The most common method of collecting demand data has been through road side interview (RSI) surveys. The major advantage of RSI surveys is that intercept data is collected with sufficient information from which 'partial matrices' can be built. Most other forms of demand data collection depend on information reported by respondents, and are thus subject to misreporting, bias, or under-reporting. However, RSI surveys suffer from the following:

- Some sites may be deemed to be too busy and lack suitable locations for RSI surveys to be undertaken. Public outcry can result from queues due to RSI surveys on heavily trafficked roads;
- The Highway Authority may not be able to authorise some locations because of local issues;
- In relation to trunk roads, HE will not allow surveys on motorways and many high speed dual carriageways; often not even on slip roads;
- To capture a reasonable proportion of demand, numerous RSI survey sites will be required along cordons and screenlines; and
- The police (or HE traffic officers) operate the sites (such officers are legally required to stop traffic) and if they are unwilling to operate a site they can prevent the survey being undertaken.

The result of the above is that often the most important and heavily trafficked routes such as A roads and motorways cannot be surveyed. Additionally, even where successful RSI surveys have been completed, the

[1] Skrobanski, G., Logie, M., Black, I., Fearon, J., Dong, Y. and Gilliam, C, (2012) *Further Developments in OD Data Fusion Methodologies*, Proceedings of European Transport Conference ,which builds upon Skrobanski, G., Logie, M., Black, I., Fearon, J., and Gilliam, C. (2010) *OD Data Fusion*, Proceedings of European Transport Conference, Glasgow

[2] 'prior' in this context meaning before any automated methods of matrix calibration have been applied (i.e. pre-matrix estimation)

resulting 'partial matrices' require infilling for movements not observed by the surveys in order to create a complete 'prior' matrix.

At the same time as RSI surveys are becoming increasingly problematic, new sources of data are becoming available, typically as by-products of the mass use of new technology. Principal among these new 'big data' sources are GPS data and mobile phone data. These opportunities are subject to rapid development at present with other research having been undertaken under the same framework as this study and with ongoing HA plans for more standardised data-sets available more widely.

It should be noted that these new datasets suffer from several shortcomings; principally a lack of segmentation (trip purpose, occupancy, income and in the case of mobile phone data, mode of transport and vehicle type), and in the case of certain types of GPS data, can suffer a very low sample rate and potential bias. Nevertheless the improved sample rate, consistency, coverage and, not least, the ease of collecting GPS and mobile phone data in comparison to the onerous processes involved in RSI surveys makes these extremely valuable data sources. If it can be demonstrated that this data can significantly enhance a model prior matrix, this could be a ground breaking and seminal development in the process adopted for building matrices, potentially saving significant resources and adding to accuracy.

Ultimately it can be envisaged that RSI survey data may be completely replaced with data from 'big data' sources. However, in commissioning this study HE is looking to investigate the impact of *adding* new data sources from GPS and mobile phone to the modelled prior matrix. If the combination of the new data sets with the prior matrix improves the modelled assignment, guidance for practitioners and developers could be written so the approach can be adopted by others.

1.1.2. Fusion

The challenge is therefore to combine new data sources with existing prior matrix data. Often in model development one dataset is used to 'control' another such that, for instance, long-term automatic traffic count data is used to constrain the turning movements observed at an adjacent junction where short-term turning counts have been manually collected – thereby retaining the short-term manually observed pattern of movements but ensuring that the level of demand is consistent with more reliable long-term data. Such techniques are intuitive and hence commonplace. However, they are not necessarily statistically optimal.

Occasionally, demand data for one type of movement might come from one data-source and that for another might be selected from another source, while in reality both sources contain data on both types of movement, only to differing degrees of confidence. The optimal way of combining two or more such data sources is to fuse them statistically, taking due account of their relative reliability with a view to minimising the overall uncertainty in the final fused result. Standard tools already exist to merge two sets of matrix data in this manner – for instance ERICA[3] - which uses a minimum variance approach by which alternative values are combined in inverse proportion to their variance.

A paper was published in the 2010 European Transport Conference by G Skrobanski (ex HA), M Logie, I Black, J Fearon, Y Dong, and C Gilliam entitled "*OD Data Fusion"* and a more recent 2012 paper by the same authors entitled "*Further Developments in OD Data Fusion Methodologies"* which outlined a new approach to data fusion. The initial paper was a largely theoretical treatise on the blending of matrix data using Bayesian methods based on MCMC simulation. The subsequent work developed the theory further using more efficient analytical modelling and applied this in demonstration to a wider area. Both papers gave a sound basis for combining matrix data that can come from a variety of sources in order to improve relatively conventional prior matrices.

[3] Matrix-building software developed on behalf of the Department for Transport. Latest variant is *ERICA 5*.

The mathematics associated with the methodology is introduced in the appropriate 'Theory' sections below. It should be noted that while the paper by Skrobanski *et al* (2012) sets out a generalised data fusion approach, it concentrates on the combination of demand matrices and count data – i.e. an alternative form of matrix estimation (ME). This is certainly one form in which the proposed Data Fusion methodology may be implemented, however it can equally be applied to the fusion of two matrices. Indeed, this '*Matrix Fusion'* is somewhat more straightforward than the fusion of matrices and link counts, which for the purposes of this report we refer to as '*Link Fusion'*.

1.2. Aims of this Study

The objective of this project as set out in the project specification is to use the latest Data Fusion methodology to prove the capability of new data sources, besides RSIs, in improving the prior matrix to produce better assignments that satisfy DMRB validation criteria.

Essentially, this requires investigation into two different facets of the problem:

- First, how the new methodology may be best applied;
- Second, how the different datasets should be cleaned, processed and corrected for bias and which performs best.

The study uses the original Prior matrix from an existing model – the 'test-bed model' – as well as "new data": GPS data sourced from *TrafficMaster* and mobile-phone data from *Telefonica* purchased via *INRIX*.

A key question is how to identify a 'better' assignment. The assumption is that an appropriate use of "big data" should reduce the level of adjustment to the prior travel matrices brought about by ME to counts. In other words, an improved prior matrix would produce closer agreement with the count data before any ME is undertaken. Metrics used to determine the quality of the updated prior matrix are explained in detail subsequently.

As the tests documented in the current study were undertaken to investigate further the Skrobanski et al. Data Fusion approach, the concepts and mathematics introduced in subsequent sections of this report are necessarily generalised to accommodate this. It is important to note that with certain simplifications the approach condenses to one of weighting in inverse proportion to the variances of the alternative sources. This is far more straightforward, is commonly implemented already and will be seen to have significant implications for the conclusions of this study.

The scope of the study was extended to ensure that both Matrix Fusion and Link Fusion methodologies were tested using the same test-bed model. The Link Fusion methodology is essentially an alternative method of ME so that part of the project is concerned with comparing the outputs of Link Fusion with that from conventional ME.

1.3. Scope change

It is noteworthy that the scope of this research project developed considerably over the course of the study. The complexity exceeded the study team's and Highways England's expectations; there were very many issues where the way forward was not clear, and assumptions had to be proposed and agreed in order to make progress. These issues are introduced in turn in the following chapters and were largely unforeseen at the outset. Collectively, they account for the significant delays in bringing the study to a conclusion. Recurring themes were:

- The practical difficulties of using the Cambridge Sub-Regional Model as a testbed, despite being chosen ahead of other candidate models on the basis of its specification against objective selection criteria
- The paucity of guidance and documented best practice in the areas of matrix building and statistical uncertainty.

It is reasonable to assume that elements of the study have now been overtaken by events. Some of the shortcomings in the data provided, particularly the mobile phone data, have now been addressed, at least in part. Consequently, were the study to have been commissioned more recently it is likely that many of the practical difficulties encountered would no longer exist. That said, reasonable assumptions were made at the time and there is no reason to suggest that the primary findings of the study – in terms of the efficacy of Matrix and Link Fusion techniques – would be any different.

1.4. Report Structure

This introductory chapter is followed in Chapter 2 by a commentary on the selection of a test-bed model to use as a practical context for the subsequent tests. Chapter 3, 4 and 5 consider in turn the data and associated unreliability (variance) for the Prior matrix, Trafficmaster GPS data and INRIX mobile-phone data. The methodology for and implementation of the Matrix Fusion methodology is discussed in Chapters 6 and 7 respectively, while the results of Matrix Fusion tests and conclusions are presented in Chapters 8 and 9, respectively.

Similarly, the methodology for the Link Fusion methodology is discussed in Chapter 10 with details of the count data and associated variances explained in Chapter 11. The implementation of the Link Fusion methodology is discussed in Chapter 12, while the results of the Link Fusion test and conclusions are presented in Chapters 13 and 14, respectively.

2. Selection of a Test-Bed Model

2.1. Assessment of the Improvements to the Prior Trip Matrices

The two main tests that may be used to assess the improvements in the prior trip matrices are:

- validation of the prior trip matrices at screenline level, without and with improvement; and
- comparison of the changes brought about by matrix estimation, without and with improvement.

An additional check, but not a primary test, would be the validation of the post-matrix estimation (post-ME) trip matrices at screenline level. Given that the same set of target counts would be used in the ME in the without and with improvement cases, the validation of the post-ME matrices at screenline level should be similar. The changes introduced by ME are liable to 'swamp' the improvements made to the prior matrices.

As this research is focused on the improvement of trip matrices rather than the improvement of the assignment, the focus should be on comparisons of modelled flows with counts at the total screenline level, rather than on comparisons of modelled flows with counts on individual links. (Link flow and journey time assignment calibration usually involves adjustments to networks as well as trip matrices.)

For the validation of the prior (and post-ME) trip matrices at screenline level, a set of screenlines is required which provides good coverage of the main movements in the modelled area. The screenlines should be entirely watertight, with the total count including flows on all roads crossing the screenline, including those not represented in the model (albeit these will usually be low-flow roads). The screenlines should be sufficiently long to avoid leakage around the ends and they should not involve significant multiple crossings. Cordons will not generally be useful for this purpose as they require accurate assignments to be useful in matrix validation. The screenlines used should include those across which roadside interviews were conducted, those on which the counts were used as constraints in the matrix estimation, and those reserved for independent validation.

The changes brought about by matrix estimation may be quantified using the measurements specified in Table 5 in TAG Unit M3-1.

2.2. Required Attributes of the Test Bed Model

The primary requirement is that the prior trip matrices were changed significantly by matrix estimation, as measured using the criteria in Table 5 in TAG Unit M3-1. If the changes brought about by matrix estimation met the TAG criteria, the scope for enhancement of the prior trip matrices would be limited. In our experience, it will be rare for a model to show non-significant changes and therefore this is a weak, but necessary, primary test of model suitability as a test-bed.

Matrix estimation should have been applied at mini-screenline level (as advised in TAG Unit M3-1) rather than on an individual link basis. In our experience, this may rule out many models.

The screenlines used for matrix estimation should include the roadside interview surveys screenlines and cordons (broken down into mini-screenlines) and also a separate set of screenlines away from the survey screenlines and cordons. The mini-screenlines should, ideally, intercept all the main movements in the modelled area.

In summary, the Test Bed model should have been developed using matrix estimation according to good practice and with the result that the changes brought about were significant.

The requirement that the changes brought about by matrix estimation should be significant implies that the prior trip matrices should be deficient. However, models for which the prior trip matrices have been developed in sub-standard ways, where good practice in building the partial trip matrices from the intercept surveys and in creating the synthetic matrices has not been followed, should be avoided. Otherwise the test of whether the data from the GPS and mobile phone systems could make a material difference to the quality of the prior trip matrices would be a false one.

So the Test Bed model should use partial trip matrices developed from the intercept survey data using good practice methods, such as that specified in the ERICA Manual. However, zonal level partial trip matrices will have many cell values with very wide 95% confidence intervals and models which have used the partial matrices at zonal level, this is, without aggregation, should be avoided. Thus a suitable Test Bed model should have been developed using partial trip matrices which have been used at an aggregate, sector level, and preferably with some recognition of the statistical reliability of the sector level movements. In our experience, this rules out many models.

During the development of the Test Bed model trip matrices should have been synthesized by good practice means. Good practice involves:

- understanding the accuracy of trip end estimates available and using the 'least bad' ones;
- using reasonably accurate estimates of inter-zonal and intra-zonal costs; and
- calibrating deterrence functions so that the modelled matrices match the partial matrix trip cost distributions and the reliable movements (at sector level) reasonably well.

A suitable Test Bed model will also have had the partial and synthetic trip matrices combined in a sensible manner. The advice in guidance is incomplete and the commonly used approach[4] of weighting the partial matrices by (say) 90% and the synthetic matrices by 10% is quite inappropriate if applied (as is commonly the case) at the zonal level, given the very low accuracy of the zonal level cell values in the partial matrices. The best practice method, in our opinion, is to carry out a three-dimensional Furness, to origins, destinations and statistically reliable movements at sector level, with the process terminating with a final balance to the movements as these are likely to be much more reliable than the zonal level trip end estimates. In our experience, this requirement rules out many models.

To summarise, the Test Bed model for use in this study should be selected such that:

- matrix estimation has been applied at mini-screenline level and has made significant changes to the prior trip matrices; and
- the partial trip matrices have been developed from the intercept survey data using good practice methods; and
- the synthetic trip matrices have been developed using good practice methods; and
- the partial and synthetic trip matrices have been combined in a way that takes account of their statistical reliability; and
- there is a set of screenlines with counts on every road and which intercept all the main movements in the modelled area and which are suitable for matrix validation.

[4] *Design Manual for Roads and Bridges, Volume 12, Section 1, Part 1*, paragraph 8.7.7 refers to the use of simple weights where the statistical accuracy of the data to be combined is unknown.

These five criteria are all important and, ideally, a suitable Test Bed model should meet them all, at least to some acceptable extent, accepting that there might be some variation between the extents to which the criteria are met.

2.2.1. Additional Criteria

An over-riding factor in making the final choice of Test Bed model is that mobile phone data relating to movements should be available at a sufficient level of spatial detail for the area covered by the chosen model. This judgement requires input from the mobile phone data provider, in this case INRIX, based on the traffic model's zone boundary GIS file. As a benchmark, the mobile phone data is usually readily available for areas the size of post code sectors – i.e. areas identified by any postcode of the form "AB12 3xx". Data for smaller areas is certainly possible but is constrained as the number of records ('probes') in each cell reduces, as small numbers are required to be aggregated for data protection reasons. The consequence is that if more rural areas are selected, the expectation is that the spatial detail will be quite coarse.

A known area of potential difficulty is the removal from mobile phone data of records ('probes') for trips using public transport. For this reason areas with dense railway networks should be avoided. Buses give rise to similar practical difficulties but it is understood that, for future studies, data brokers such as INRIX are developing techniques to systematically identify tight clusters of 'probes' and relabel them as just one 'probe', translating to one vehicle in this context. For this study no such rationalisation will take place; hence a high proportion of bus travel is undesirable.

It is impractical to process data on a national basis in the context of this study and to avoid this it is necessary to be able to deal with external zones in an aggregate way, as zones representing trips to/from each of the crossing-points on the study area boundary; i.e. in the manner of a cordon model drawn from a bigger model. Thus, models requiring an extended external network to accommodate strategic route choice outside of the study area are not appropriate for the purposes of this study.

Regarding survey timing, INRIX had a comprehensive (national) sample for one week during May 2013 and were in a position to collect new data for this study in any specific area for the first neutral period in Autumn 2013, i.e. in September 2013. The selected Test Bed model would therefore ideally be based on May or September RSI and count data, though this was not deemed to be an important requirement.

It is noted that the mobile phone data is not segmented by vehicle type or purpose, so although such disaggregation is conventionally regarded as good practice in model development, it is not a requirement in this context.

The search for suitable Test Bed models which meet the above criteria was focused on those models which were created for the development and appraisal of Highways Agency schemes. The intention was that if there were a choice of suitable models, those based on more recent datasets and those of more modest areas (as opposed to large-scale models covering whole regions or sub-regions) should be preferred, mainly to make the data processing more tractable.

2.2.2. Summary of Criteria

Taking all the above issues together:

- The model should include prior demand matrices, built using traditional roadside interview methods, according to reasonably good practice. Unobserved movements should have been estimated in some way, ensuring that the matrix is complete. The preference is for models that have not used the roadside interview data at a zonal level, but only to calibrate a suitable synthetic model at a sectoral

level; it is considered that sample errors on roadside interview data at a model zonal level are too large to make them reliable;

- Matrix estimation should have been carried out on the prior matrices to incorporate observed count data into the process. This should have been done in a robust way; preferably not starting with a matrix which had been derived using ME. The preference is for models that have carried out matrix estimation at a short screenline level, rather than by individual site; this reduces the risk of incorrectly calibrating away network or assignment errors by adjusting the matrix. Matrix estimation should have had a significant effect on the prior matrices (if it didn't, the existing prior matrices are probably good enough so that significant improvement from mobile data is unlikely);
- A large and robust set of validation data, arranged into screenlines with minimal holes where possible (again, to reduce the effect of network and assignment problems), will be needed to assess the quality of the matrix;
- The model should not cover a very large area if possible to reduce the volume of data required. It should be a highway model; ideally used to assess Highways Agency or at least major trunk-road schemes. A UK model is required;
- If possible, the model should have been built using reasonably new observed data (roadside interviews and counts);
- The area covered by the model must be acceptable to INRIX, the mobile data supplier. From discussion with them it appears that areas with dense rail networks (e.g. London) should be avoided, as should models with very detailed zoning, especially in rural areas, if possible.

2.3. Initial Review and Shortlist

2.3.1. The Initial Long-List

From discussion with the Highways Agency, Atkins and AECOM's transport modelling staff, a list was prepared of 18 models that appeared to have broadly appropriate scope. The list includes both modest models of Highways Agency schemes and larger more strategic models, as summarised in Appendix A.

The appendix contains discussion of some of the pros and cons of the various options

2.3.2. Selection of a Short-List

Of the long-listed models considered, the A14 and A21 models look like the most appropriate small Highways Agency-scheme-specific models. The A21 model has the most recent RSI data of the models considered.

Considering larger models, the Tyne and Wear TPM, SYSTM+ and CSRM looked reasonably promising. The CSRM and Tyne and Wear TPM have the most compliant matrix build methodology of all the models reviewed; however both use relatively old RSI data and the TPM has a dense rail network in its modelled area.

It was considered that none of the other models on the long list were worthy of further more detailed consideration for use in this study, for various reasons including:

- impractically large (e.g. EERM);
- convoluted matrix build history / lack of provenance (all of the smaller models suffered this to some extent);

- unusual matrix build methodology not primarily based on RSIs (A11 local model);
- lack of trunk road focus (e.g. Stevenage-Hitchin model); and
- little validation data organised into screenlines (A30 model).

Candidate models that satisfied the selection criteria less well than one of the five models identified above (A14, A21, Tyne and Wear, SYSTM+ and CSRM) were rejected. For example, NATS was considered reasonably appropriate upon first review, but when it was compared with the CSRM it uses older data, is less strategically focused, and did not use short screenlines in estimation; consequently it was judged less suitable for use as the Test Bed model for this study.

The candidate model zoning and study areas were reviewed by INRIX and their views were captured and are shown below in Table 2-3. INRIX felt that the level of spatial detail available for the CSRM (i.e. in Cambridge) would not be as fine as other areas such as the A21 model, where zones can be potentially about 300m by 300m square. This would result in traffic model zones in Cambridge being aggregated somewhat[5] for the purposes of providing mobile phone data.

2.4. Comparison of Short Listed Models

The five shortlisted models are summarised in terms of the criteria below. Each model was scored from 1 to 5 on each criterion in addition to describing the suitability (5/green being good, 1/red being bad). It is noted that this scoring was somewhat subjective and that the criteria are not necessarily of equal weight, so adding up the scores is unlikely to be very helpful.

Several distinctions can be made in addition to the tabulated criteria. In particular a choice was required between a larger, more complex model with a more robust matrix build methodology (CSRM or Tyne and Wear TPM), and a smaller, more Highways Agency-scheme-specific model with a less robust matrix build methodology (A21 or A14).

The shortlisted models were considered further by the study team and in particular the nominated Technical Reviewer for this work, Dr. Denvil Coombe. The use of RSI data at a zonal level was considered a significant negative point by Dr. Coombe, and thus the A21, A14 and SYSTM+ models were rejected. Although the Tyne and Wear TPM did use RSI data at an aggregate level, the methodology (control of synthetic data to extremes of RSI uncertainty) is an unusual approach, not applied in any other model familiar to the study team, and this, combined with the age of the data, suggested that the Tyne and Wear TPM was unlikely to be the best candidate either.

This left only the CSRM (Cambridge Sub-Regional Model) as a suitable candidate; this was considered the most promising model, subject to the caveat from the mobile phone data supplier, INRIX.

[5] The initial view was that Cambridge city zones would need aggregating approximately three to one, though in the end the ratio transpired to be five to one.

Table 2-1 Shortlist of Promising Models

	Cambridge Sub-Regional Model	Tyne and Wear Model	A21 Tonbridge to Pembury traffic model	A14 Local Model	SYSTM+
Prior Matrix Build	RSI cordons around Cambridge and Huntingdon, with an RSI screenline through Cambridge. Observed matrices were combined using a barrier method, but with considerably less weight given to transposed trips. RSI data were used only at a (42) sector level, with synthetic and JTW data used to split it to zonal level. External demand was taken from EERM, which did use RSI data at a zonal level, but given that external CSRM zoning is highly aggregate, this is not a significant concern. 4/5.	Considerable RSI coverage, not well grouped into cordons or screenlines; about 50% of movements were thought to be observed. RSI data were used only at a (around 40) sectoral level, and was merged using variance method. A complex synthetic gravity model was used as the starting point for the matrix, calibrated to NTS and other household survey data. This was adjusted where sector-sector movements lay outside the confidence interval of relevant RSI data. 4/5.	11 RSI sites were used to build an observed matrix. The barrier method was used to eliminate double counting, and the matrices were used at zonal level. An ANPR survey was used to estimate through trips on the A21. Unobserved movements were filled in using data from the 2009 Tunbridge Wells Traffic model. 2/5.	Built using two cordons of RSIs, with gaps on A14. The data were used at the zonal level, and the barrier method was used to combine the matrices. External demand was taken from NNTM and EERM. No synthetic demand was used. 2/5.	Around 300 RSI sites. Merged using barrier method, but with transposed trips used only where no forward-direction data available. Cordons/ screenlines exist, but not watertight. Data used at zonal level. Assignment used to identify observed movements. Synthetic gravity model calibrated to observed trip distance profile for observed movements. Merged with observed data using a variance method; however assumptions mean greater weight is given to the observed estimate. 2/5
Size and Scope	Covers Cambridge, Huntingdon and area between. Moderate size, therefore. Quite focussed on the A14, but not explicitly an HA-specific model. 3/5.	Covers Newcastle and environs. Not specifically trunk-road focussed, but does include several trunk roads. 3/5.	Tunbridge Wells and Tonbridge. Small size, and focus is on trunk roads in the area. 5/5.	Kettering, Corby, and Wellingborough. Fairly small, and focus is on the A14. 4/5.	South and West Yorkshire, as well as parts of Derbyshire and Nottinghamshire. Fairly large. 2/5
Age of Data	RSI Data is from 2005-2006. Model base year is 2006. Count data is 2006-2008. 2/5.	RSI Data is from 2000-5. Base year 2005. Count data mainly 2005, though getting hold of updated counts at the same sites should be fairly easy. 1/5	Base year is 2012. RSI data is from 2009-2012. Many counts from 2012; some older. 4/5.	RSI data from 2005. Count data from 2004-2008. Base year is 2005. 2/5.	RSI 2001-2008 (most around 2005). Base year is 2007. 2/5
INRIX Perspective	Relatively rural area with fairly detailed zoning. Poor granularity matching to the cellular network. Little rail network in area. 2/5.	Highly detailed zoning in urban area, with fairly dense rail network. Probably not very suitable. 1/5.	Rail network is fairly sparse. Zoning moderately detailed but area is mostly urban. Good coverage within the cellular network 5/5.	Reasonable granularity within towns, but worsening quickly into more 'rural' areas. Rail network fairly sparse. 2/5.	Moderately dense rail network, with fairly detailed zoning in urban areas. Large population leading to significant processing requirements 3/5

	Cambridge Sub-Regional Model	Tyne and Wear Model	A21 Tonbridge to Pembury traffic model	A14 Local Model	SYSTM+
Matrix Estimation	Significant number of counts, including both MCC and ATC data. Some data (MCCs) were grouped into short screenlines. ATCs were used at link level. Changes were significant; there were 5-10% changes in sector-sector movements. Some intra-sector movements (where the input demand was wholly synthetic) changed by more than 30%. 4/5.	Some sites were grouped into short screenlines, where possible. Some were used as individual sites. Strangely given the detail of the rest of the documentation and general focus of the approach on minimising the intervention of ME, changes brought about by ME do not seem to be reported. Text suggests changes are likely to be quite small, but they're probably still not negligible. 3/5.	Five screenlines and two cordons were used; coverage looks fairly comprehensive, but they appear to have been used at a site level (it isn't entirely clear; it's possible some short screenlines were used). Some RSI data was "frozen" in matrix estimation process. Significant changes were observed; r-squared values for pre against post cell-level demand were mostly around 0.75-0.80. 3/5.	One calibration screenline was used in combination with the RSI cordons. There were also a number of sites not combined into screenlines. Estimation was done by site. Some significant changes (>5%) were observed in matrix totals following estimation. 2/5.	80 screenlines and cordons and 600 sites, offering wide coverage of the model area. Some sites were grouped into short screenlines (all sites were so grouped for HGV and LGV), but some car counts were used at a site level. No data were saved for validation. Some changes were observed; average trip lengths changed by more than 5% for some periods and some sector-sector changes were quite large; generally matrix does not seem to have been greatly distorted, however. 4/5
Validation Data	There are relatively few independent validation (not used in calibration) counts, and these are not well grouped into screenlines. However, the calibration data is quite extensive and the scope of estimation could be reduced for the research if felt useful. The calibration data is mostly cordons, however; there are relatively few real screenlines. 2/5.	Very large number of (mostly fairly short) validation screenlines. Generally free of significant holes. 5/5.	Reasonable amount of independent validation data; combined with the calibration screenlines the coverage is considerable. However, the validation data is not organised into screenlines. 4/5.	Three screenlines; two RSI cordons, A14 slip-roads, and a set of stand-alone validation sites. Fairly comprehensive; standalone sites are difficult to combine into screenlines. All screenlines contain some holes. 3/5.	No independent validation data, but calibration screenline coverage is extensive. Not possible to assess screenlines for holes based on documentation. 4/5
Other Issues	May be political issues reporting results as it is being used on a live scheme. 3/5.	None known. N/A	None known. N/A	None known. N/A	None known.

3. Prior Matrix Data

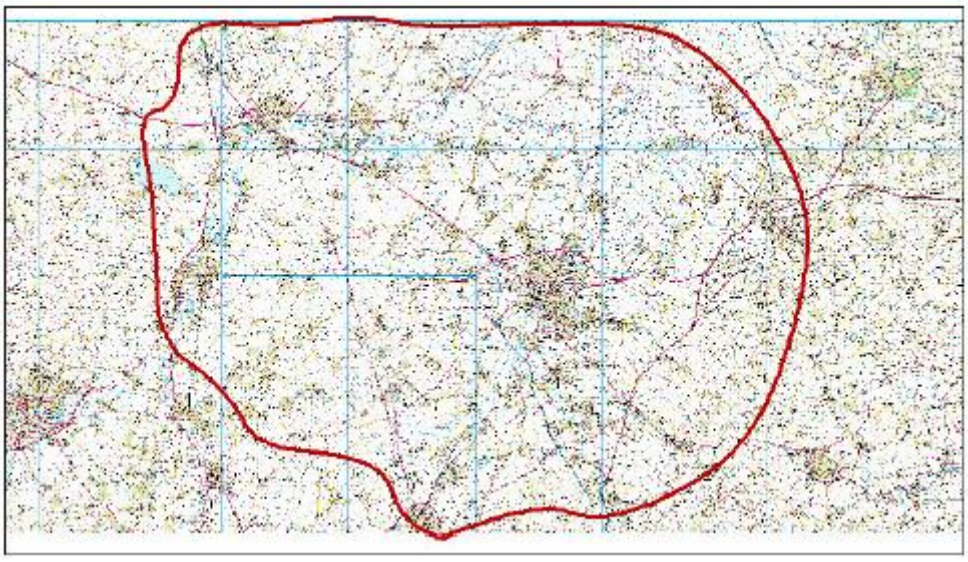
This chapter considers the data sources and structure of the original Prior matrix developed for the CSRM. It is this Prior against which any improvements due to the fusion of new data is to be compared. The chapter introduces the CSRM and goes into more detail about the development of the original Prior matrix, before discussing the processing of this matrix for use in Data Fusion (i.e. both Matrix Fusion and Link Fusion) and finally discussing the calculation of variances representing the uncertainty associated with the Prior matrix in the fusion processes.

At an early stage in the study the decision was made to undertake Matrix Fusion using matrices of proportions rather than matrices of trips. The “new data” being fused does not contribute in any useful way to an estimate of total traffic, since there is no clear basis on which either the TM or the INRIX data has been sampled, meaning that the effective sampling fractions cannot be determined. Hence the fusion was undertaken at the sample level (of proportions), rather than the expanded level (of trips), recognising that after fusing, the output proportions would need to be expanded. The variances considered later in this chapter were calculated accordingly.

3.1. The Cambridge Sub-Regional Model

The CSRM traffic model was constructed using SATURN software to represent traffic movements in the area of Cambridge, Huntingdon, and the region between, during a typical weekday in October 2006. It forms part of a variable demand model and was developed from earlier models of the area going back to the Cambridge to Huntingdon Multi-Modal Study (CHUMMS) model. The purpose of the 2006 CSRM as used in this study was to support Cambridgeshire County Council’s Transport Innovation Fund bid. It includes AM peak hour, average inter-peak hour and PM peak hour models. The study area is shown below, with Cambridge as the main urban area and Huntingdon included towards the north-west corner of the area.

Figure 3-1 Extent of CSRM Modelled Area



The model development and validation is documented in the Local Model Validation Report[6]. Greater detail on the Prior matrix development is provided in a supplementary Technical Note 5[7].

The model area comprises 325 zones. More is provided on the development of the Prior matrix below, with this subsequently being calibrated using counts as illustrated in Figure 3-2.

There are ten user classes, representing different vehicle types, journey purposes and (for consumer light vehicles) different degrees of willingness to pay. There is no distinction between cars and LGVs – both vehicle types are included in all eight light vehicle user classes. This is not ideal given the segmentation available from the National Travel Survey (NTS) and Trafficmaster GPS data and is an acknowledged shortcoming of the model. The user class definitions are reproduced below.

Table 3-1 CSRM User Class Definitions

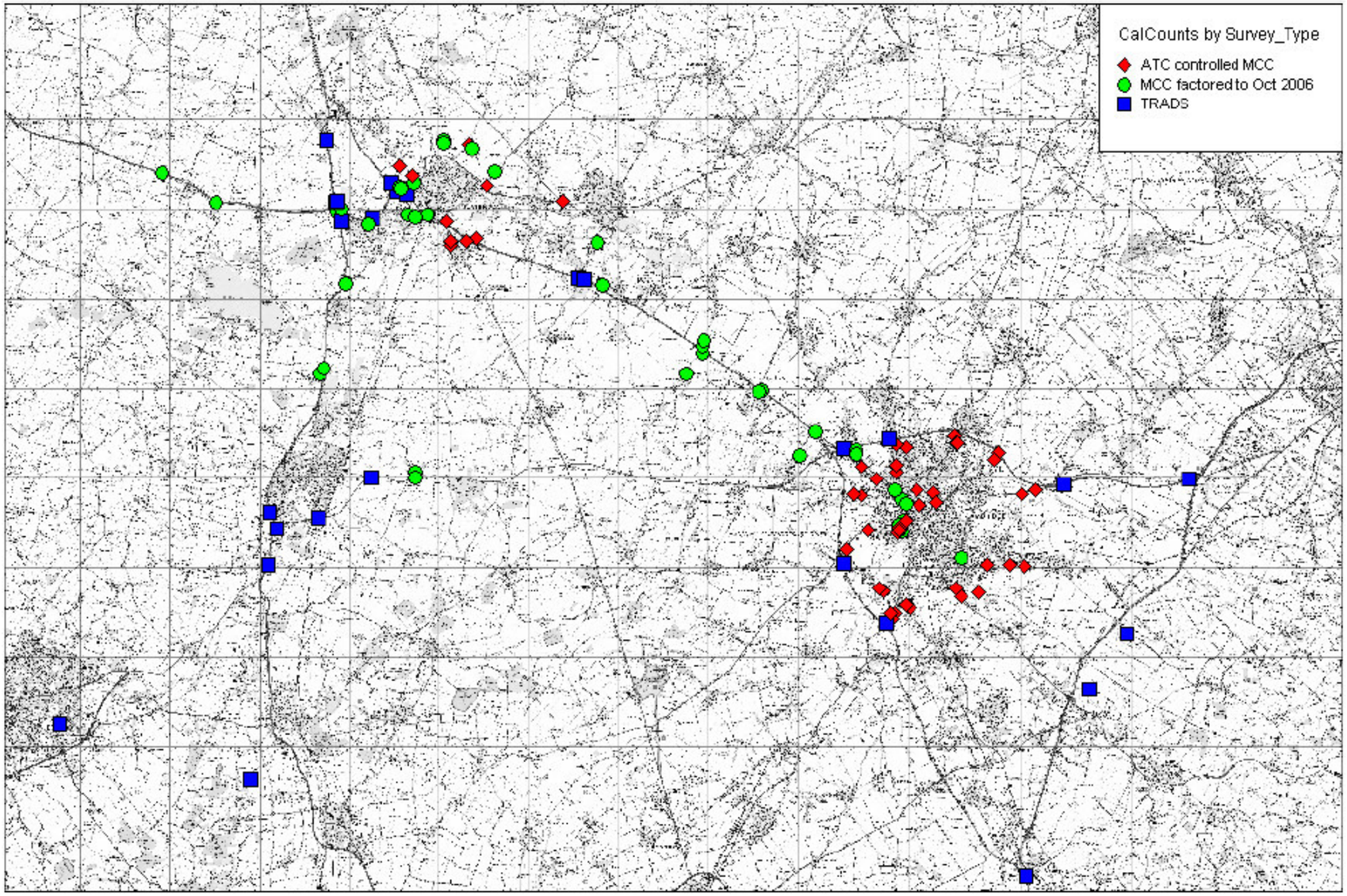
User Class	Vehicle Type	Purpose	Income	Identifier
1	Light	Home Based Work (commute)	Low	HBW Low
2	Light	Home Based Work (commute)	Medium	HBW Medium
3	Light	Home Based Work (commute)	High	HBW High
4	Light	Home Based Education	Not applicable	HBEd
5	Light	Employers Business	Not applicable	EB
6	Light	Other (Discretionary)	Low	HBO Low
7	Light	Other (Discretionary)	Medium	HBO Medium
8	Light	Other (Discretionary)	High	HBO High
9	HGV	HGV>7.5t Without a Huntingdon trip end		HGV1
10	HGV	All other HGVs		HGV2

The final base year traffic model was developed after the application of matrix estimation. Somewhat unconventionally, this was undertaken in two stages. Before the final ME stage, the original Prior matrix for each time period was further modified on the basis of a very simple initial ME relating (in the AM peak) to counts at the A1/B1043 at Sawtry. This "was required to address the discrepancy that would otherwise arise between the largely EERM-sourced flows predicted by the model and the observed October 2006 flows". The initial ME stage can therefore be regarded as a 'correction' (and is referred to as the 'A1 correction' hereafter), with the final ME, using all calibration counts, considered a conventional use of ME as a calibration tool.

The ME processes were undertaken sequentially for heavy vehicles and then for light vehicles (across all user classes simultaneously). As the estimated matrix leads to alternative routeings, ME is an iterative process and was repeated up to six times. In the case of the morning peak hour model, the fifth iteration was found to yield the best calibration results, and for consistency all matrix estimation comparisons and further work undertaken with the CSRM for the purposes of this study also employed five iterations.

[6] "Cambridge Sub-Regional Model, Highway Model Component: Model Development and Validation Report", July 2009. Document *5075394-LMVR-v7-draft.doc*.
[7] "RSI Data in Integrated Model Highway Matrices", October 2008. Document *5075394 TN5 Matrix Building v2.doc*.

Figure 3-2 Location of Counts used in Calibration



In line with best practice, counts used in the main ME were in some cases 'grouped' to form mini-screenlines. The counts that were grouped were all manual classified counts which have a reduced reliability compared to the automatic traffic counts comprising the remainder of the count-set. The grouping was undertaken in such a way as to increase the reliability of the grouped counts until the standard error (relative to the mean) was roughly equal across all counts used.

An integral part of the main ME process was the constraint to trips ends as well as link counts. The trip end constraints relate to 88 zones comprising the entirety of Cambridge. In addition, several zones were completely 'frozen' (i.e the relevant matrix row and column were fixed)[8]. All these constraints were taken from the CSRM demand model for consistency with other elements of the CSRM model system.

A large number of calibration and validation measures were derived and reported. These have been reviewed and form the basis of the metrics for comparing test results documented later in this report.

3.1.1. Scope of Use in the Data Fusion Study

The study in hand concerns the identification and comparison of improvements in an example prior matrix. Only one such matrix is required for this purpose. Given the lack of segmentation available in mobile phone and GPS data this study is focused on the development of the light vehicle (all purposes) CSRM Prior matrix for the 2006 AM peak hour model. For the purposes of the tests, all other inputs to the CSRM model system are assumed to be unchanged. This includes both the pre-ME and post-ME heavy vehicle matrices, and this has required edits to the original ME process in order to effectively 'freeze' the HGV component in all subsequent tests.

The application of ME and the basis for comparison between the original Prior and those matrices developed in subsequent tests is significantly complicated by the aforementioned 'A1 correction'. In all cases it would be simpler if this could regarded as part of the development of the original priors, but as the basic properties of the matrices are changed by this adjustment, it would be difficult to assess the impact on the variances. Instead, at least in the case of the Matrix Fusion tests, it is preferable to regard the ME process as occurring in two stages – the initial adjustment for the 'A1 correction' and then the full ME. That means omitting the 'A1 correction' from the definition of the priors. The new post-Matrix Fusion matrices will then go through both ME processes.

In the case of the Link Fusion tests, the study is attempting to establish whether the fusion of a prior matrix with a set of link counts is any better than a conventional ME to link counts. The 'A1 correction' in this instance can therefore be regarded as part of the Prior matrix development such that the comparison is between (solely) the main ME process and the Link Fusion test[9]. The existence of trip end constraints applied in the original CSRM main ME process is a further complication, but these have been retained to allow direct comparison with the original CSRM modelling.

Thus the study used two versions of the CSRM prior matrices (and their associated variances) – the **Matrix Fusion Prior** excludes the 'A1 correction', but to facilitate a like-for-like comparison with ME, the **Link Fusion Prior** is the 'post A1 correction' matrix from the CSRM model development. That said, for comparison purposes some of the Link Fusion tests were undertaken using the Matrix Fusion Prior as well as the Link Fusion Prior, as described in section 13.

[8] these relate to 14 special zones, comprising park & ride sites, motorway service areas and certain business parks.
[9] This has the added complication of needing to adjust the input prior matrix variances to reflect the bias introduced by the 'A1 correction', as described in the relevant Link Fusion section.

3.2. Data Sources

3.2.1. Development of the Original CSRM Prior Matrix

The original Prior matrix included the use of data from a variety of sources:

- RSI data for 10 cordon sites around Huntingdon collected in 2005;
- RSI data for 15 cordon sites around Cambridge and 6 sites along the River Cam within Cambridge, collected in 2006;
- 'synthetic' data from the demand model component of the CSRM model system - a land-use / transport interaction model developed by WSP;
- externally modelled data for long-distance trips taken from the highway component of the East of England Regional Model (EERM).

The RSI data was processed and expanded to associated count data in a conventional manner. It forms the most reliable origin-destination data in the Prior matrix but only relates to those trips crossing an RSI cordon.

The Synthetic data was used in the prior matrix development to:

- seed movements that otherwise would not cross an observed Roadside Interview cordon;
- smooth the distribution of movements that do cross an RSI cordon.

The source of the synthetic data used varies by journey purpose:

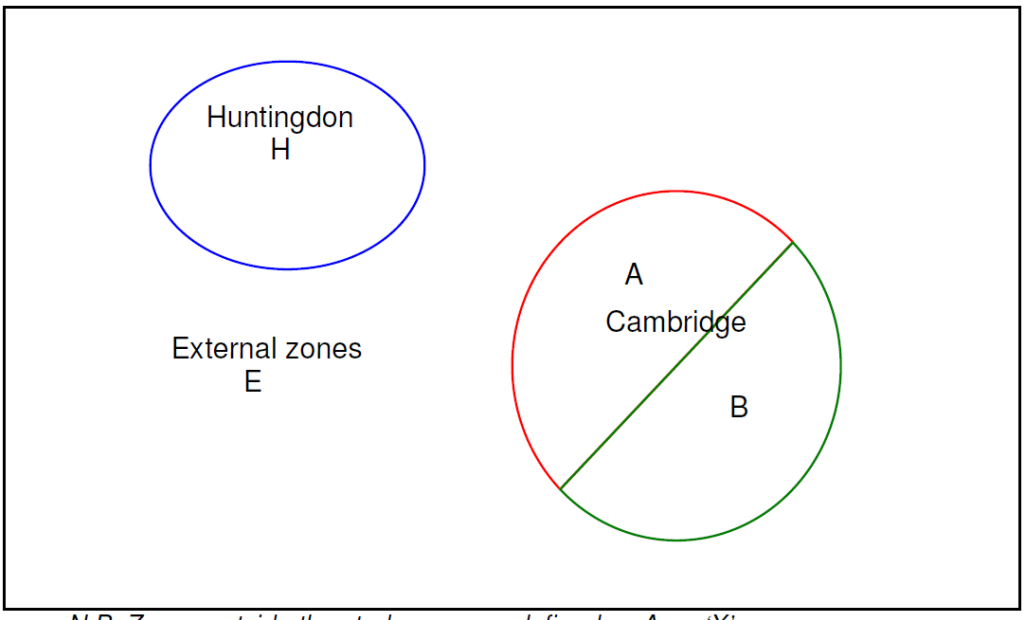
- Home Based Work - derived from census data supplied by ONS;
- Home Based Education - taken from the CSRM demand model;
- Employers' Business – taken from the CSRM demand model;
- Other – taken from the CSRM demand model.

The EERM is a strategic transport model that covers the eastern region. The CSRM highway model has the advantage of being able to draw upon this larger area model to provide a representation of strategic flows though the study area. Data from the EERM was supplied in the form of cordoned trip matrices to provide a representation of the following movements for both light and heavy vehicles:

- movements that have both an origin and destination outside the study area but which travel through the study area. An example of such a trip would be a movement from Felixstowe to Birmingham – reflecting the number of trips through the study area - excluding those which choose to use the M25/M40 route;
- movements that have an origin or destination outside the study area (external to internal or internal to external). There is an exception for those trips crossing the Cambridge or Huntingdon cordons as matrix building for these trips is based on observed RSI data rather than EERM model data.

The Prior matrix is built from these data sources, with different sources used for movements between five different geographic areas (A, B, H, E, X) as defined below. These areas are referred to as 'super-sectors' hereafter.

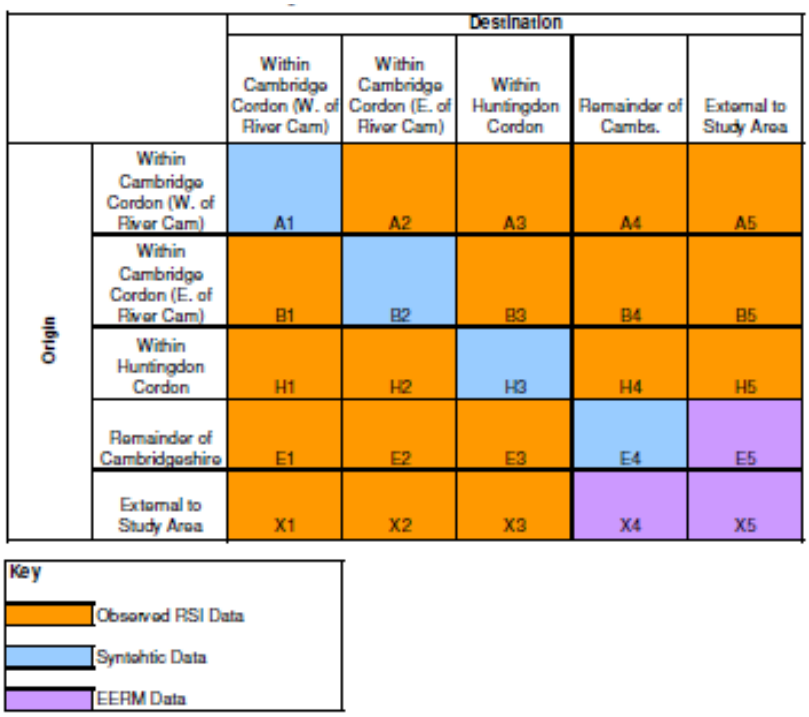
Figure 3-3 Definition of Matrix Building Areas



N.B. Zones outside the study area are defined as Area 'X'

The sources used for each of these areas of the Prior matrix are illustrated in the diagram below:

Figure 3-4 Sources of Prior Matrix Data



3.2.2. Processing for use in Data Fusion

Light vehicle OD matrices at the sector level are required: no purpose or other breakdown is necessary for the data fusion, so data can be combined over the first 8 user classes in Table 3-1.

The data comprising the Prior is most reliable for the RSI-observed portion and, as the RSI data is too "lumpy" to use at the zone level, the data fusion should ideally be undertaken at a sector level. It is possible that the new data sources will allow a greater level of reliable spatial detail, and this can be used subsequently to disaggregate the sector matrices; this is addressed by the Matrix Fusion test programme, documented later.

For the construction of the original CSRM RSI matrix, a 38 sector system was used. This was designed such that sector boundaries were constrained by the screenlines, to preserve totals of trips between areas, ensuring that the unobserved trips remain well-defined. However, there was a concern that this might give rise to many cells in the Prior trip matrix with large variances. Accordingly, a new sector system was designed to spread the records more evenly such that, for each sector-level cell, there is a sufficient number of records but not too many for a certain level of reliability.

Note that the level at which fusion is applied need not be individual cells in a particular sector system. In other words, if there are N sectors, it is not necessary to apply the fusion separately to the N^2 cells of the sector matrix: a further amalgamation of cells can be applied to obtain reliable results. (Note that the cells referred to here and below are cells of the sector matrix not of the underlying matrix of trips between zones). The sole caveats are that the (amalgamated) cells at which fusion is carried out should respect the A/B/H/E/X boundaries and need to be geographically contiguous.

Ideally, the basis for doing this should be to choose the sectors so that the variances of the (amalgamated) cells are reasonably consistent (this was necessarily a trial and error process, as the study team members were not aware of any principles by which this could be optimised). On practical grounds, however, it is much more straightforward to work with the RSI sample sizes and ensure that these are distributed in a consistent way, and this was the approach adopted.

As explained in the following section, if the aim is to have a 95% confidence interval on the expanded estimate of ±20%, then this implies a sample size of just under 100, and if it is acceptable to increase the confidence interval to ±30%, this falls to 43. On this basis, the sectors and the amalgamations have been selected so that the (amalgamated) cells contain RSI samples in the range 50-100. In all places amalgamated movements have been kept within the same A/B/H/E/X movement with the exception of trips between Cambridge (A or B) and Huntingdon (H), where A and B have been grouped together since observations are quite sparse in this case.

There are 29 identified sectors (shown in Figure 3-5 compared to the underlying zoning system, and in the wider geographical context) and many more amalgamated 'blocks' of RSI observations[10]; the resulting RSI sample sizes, are shown in Figure 3-6 below.

For the unobserved areas of the matrix, blocks have also been defined in order to facilitate data fusion. In these areas the blocks have been chosen to yield Prior matrix values of approximately 300-400 (N.B. the RSI expansion factors for the observed movements was typically about 4). Figure 3-7 shows the definition of these amalgamated blocks and the AM peak Prior matrix movement totals in each. The total number of observed and unobserved amalgamated 'blocks' is 184.

[10] Note that for practical coding purposes the RSI observed 'blocks' have been defined as rectangular areas.

Figure 3-5 Location of 29 Sectors Used for Defining Amalgamated Fusion Areas

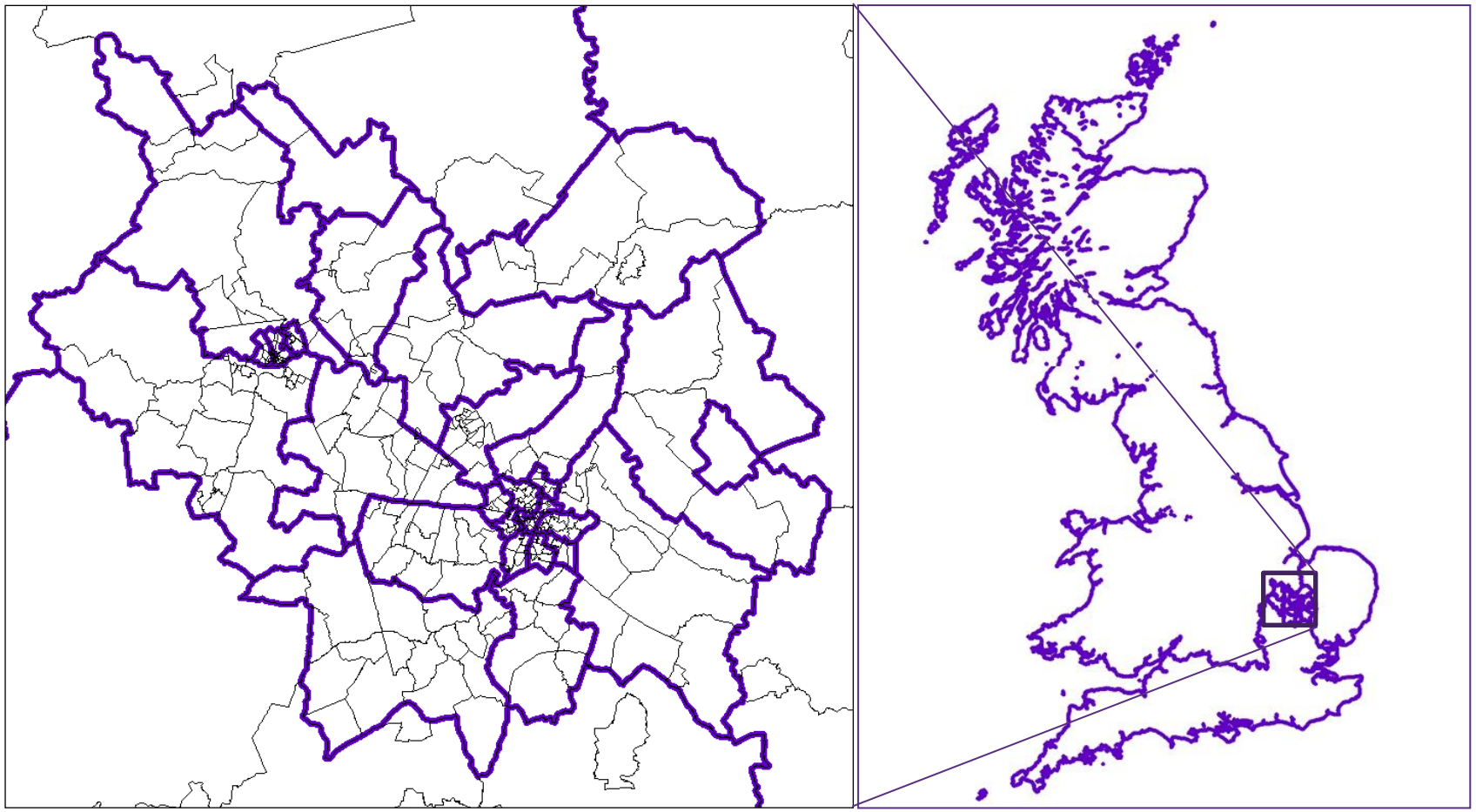
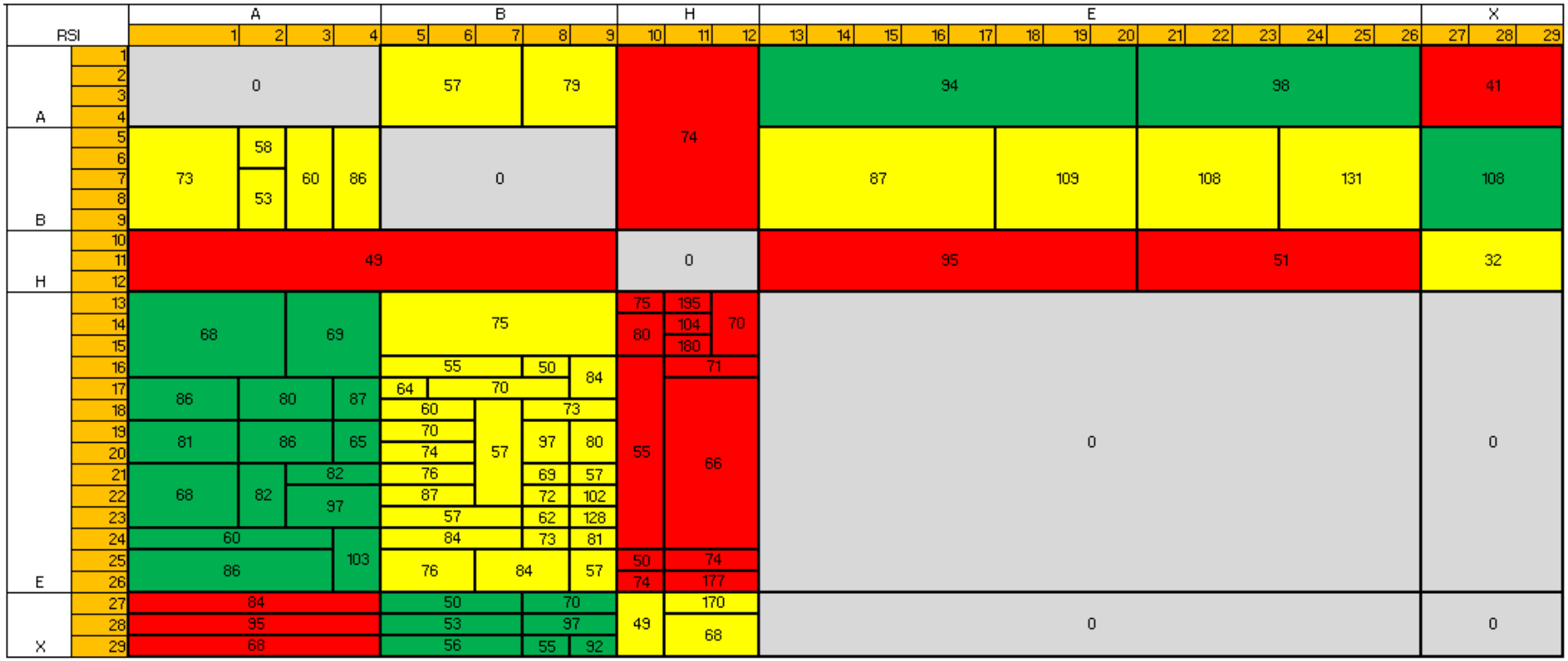
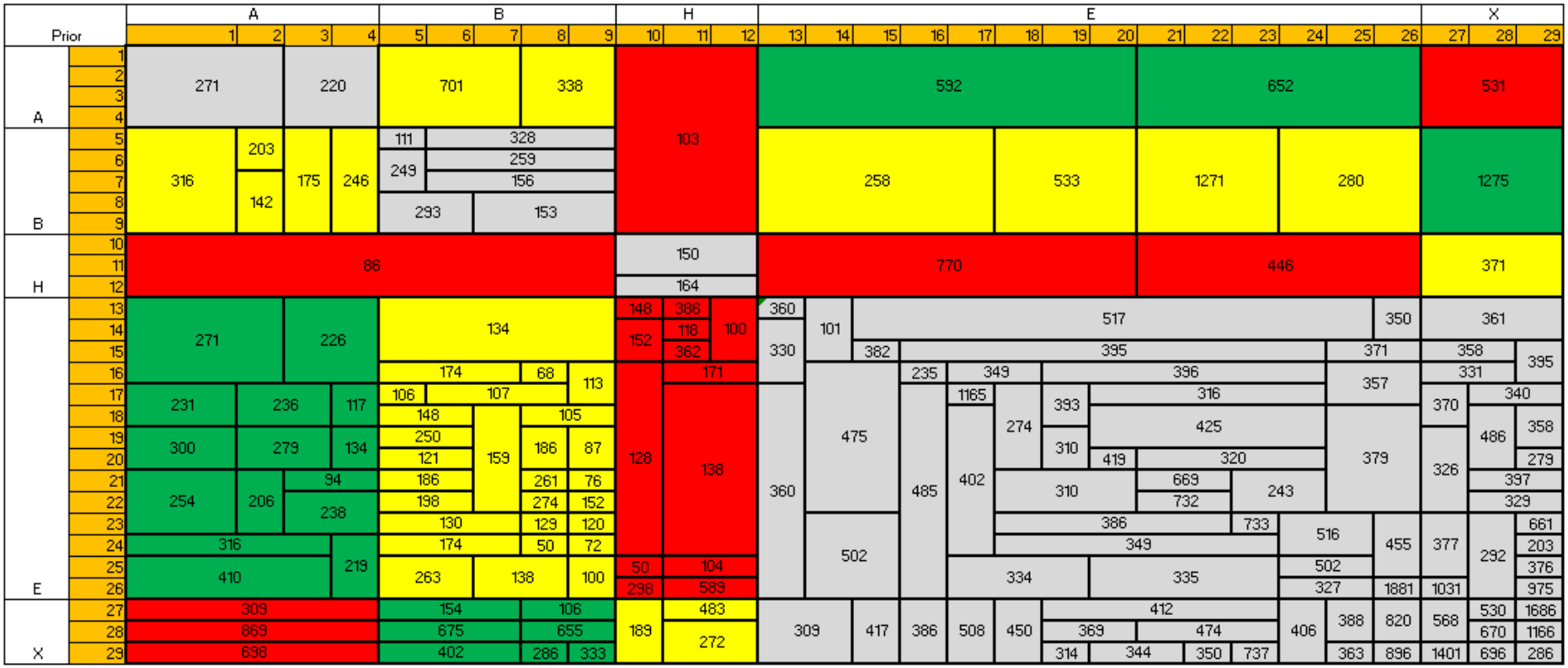


Figure 3-6 AM Peak RSI Sample Sizes for Amalgamated Fusion Areas



Note that there is no meaning to the red/yellow/green other than distinguishing movements between different super-sectors.

Figure 3-7 AM Peak Prior Matrix Movements for Amalgamated Fusion Areas



As above but with movements rather than samples and figures included for all the ‘blocks’

3.3. Variance Assumptions

While the Data Fusion formulae are not presented until section 6, it is important to note here that an estimate is required of the variance of the value provided for each Prior matrix element (and indeed the equivalent elements from the new data sources). This variance is a measure of the statistical uncertainty associated with the value concerned and can be extremely challenging to calculate. The means of deriving appropriate variance estimates varies according to the source of data and, consequently, in the case of the Prior matrix the variance calculations vary by *area* of the matrix.

3.3.1. Observed Movements

To determine the variance of the RSI cells (shaded orange in Figure 3-4), the unweighted sample size N_{ij} as well as the implied expansion factor $e_{ij} = T_{ij} / N_{ij}$ are known. For some movements, these will derive from more than one RSI (including both the Huntingdon and Cambridge cordons): and it is assumed the combination has been done correctly. This allows us to calculate the variance for each sector-sector movement. If N is the total RSI (AM peak) sample[11], and we assume a multinomial distribution for each ij cell, with $p_{ij} = N_{ij}/N$ as the probability of observing cell ij on each event, then the variance of N_{ij} is given as:

$$var(N_{ij}) = N * p_{ij} * (1 – p_{ij}) = N_{ij}.(1–N_{ij}/N)$$

Correspondingly, since the expansion factor e_{ij} is a constant, by the properties of variance that:

$$var (a * X) = a^2 * var(X)$$

for any constant a, then the variance of the expanded quantity $T_{ij} = e_{ij}.N_{ij}$ is given as:

$$var(T_{ij}) = e_{ij}^2 * var(N_{ij}) = e^2_{ij}.N_{ij}.(1–N_{ij}/N)$$

and the variance of the proportions pij = Nij/N is given as:

$$var(p_{ij}) = 1/N^2 * var(N_{ij}) = p_{ij}.(1 – p_{ij})/N$$

To help define the ‘fusion blocks’, the following approach was used. The estimate of a sector-sector movement is $T_{ij} = e_{ij}.N_{ij}$. If we want to have a 95% confidence interval on this of (say) ±20%, then this implies approximately:

$$1.96\ \sigma = 0.2\ e_{ij}.N_{ij}$$

where σ is the estimated standard deviation of T_{ij}, and we have $\sigma^2 = V_{ij}$.

Squaring the equation and substituting for σ^2, we get:

$$1.96^2.\ \sigma^2 = 1.96^2.\ e^2_{ij}.N_{ij}.(1– N_{ij}/N) = 0.04\ .\ e^2_{ij}.N_{ij}^2$$

Cancelling terms gives approximately: $(1– N_{ij}/N) = 0.0104\ .\ N_{ij}$

Hence $N_{ij}.(0.0104 + 1/N) = 1$ and N_{ij} is about 96. If we were willing to increase the confidence interval to ±30%, then N_{ij} would fall to about 43. These figures guided the amalgamation of the sector-cells to yield the ‘fusion blocks’ introduced above.

Although further “smoothing” of the CSRM Prior took place, this was only done at a lower spatial level: hence, at the fusion block level, the variances do not need any further adjustment.

[11] with reference to Figure 3-6 there are 5534 from the Cambridge RSIs, and 1859 from the Huntingdon RSIs.

3.3.2. Un-observed Movements

For the non-RSI movements, it is much less clear how to proceed in terms of the variance calculations. In the EERDM MDR[12], a method is proposed based on very early work by Haskey (1973) but this seems to relate to the additional variance which might be introduced by fitting a gravity model to RSI data, and is not obviously relevant to an entirely (independent) synthetic source, as is the case with CSRM.

Based on analysis of the CSRM documentation[13], there are 24,022 light vehicle trips in the RSI-based movements, 29,165 in the synthetic (WSP) movements and 23,068 in the EERM-based movements. There is no further information on the reliability of the latter two sources which might be of help, so reasonable assumptions are required.

In principle, it might seem that – given a reasonable RSI coverage, as here – the remaining cells should be less reliable. Nonetheless, given that the majority of synthetic cells are short-distance, and this is where the majority of the trips occur[14], it is possible that the synthetic movements are relatively accurately estimated, particularly for the morning peak with commuting trips dominating. The quality of such movement data is far from clear and will vary with the quality of the underlying trip end data and the subsequent distribution modelling.

It was proposed that a range of alternative assumptions be made. First a general relationship was estimated between the variances, V_{ij} and the (expanded) trips, T_{ij} for the RSI cells, plotting to see the shape of the curve, and fitting the following function:

$$\ln(V_{ij}) = f(\ln(T_{ij}))$$

where f is fitted by polynomial regression (i.e. f is the n^{th} degree polynomial that maximises the R^2 goodness of fit statistic, which is the coefficient of determination in the regression analysis).

This relationship was used to give an "RSI-equivalent" variance calculation for all the other cells (i.e. an estimate of the variance which might apply if the cell had been derived from appropriate RSI data). For both the synthetic movements (blue in Figure 3-4) and the EERM-based movements (mauve), as a central assumption these estimates were multiplied by 2 to indicate a greater level of unreliability. However, to investigate the impact of this assumption, further versions were tested using:

- a factor of 4 for both synthetic and EERM-based movements;
- a factor of 2 for EERM-based movements and of ½ for synthetic movements, reflecting the above possibility that the synthetic movements might actually be reasonably accurately estimated.

Three further alternative proposals were tested. One is to confine the fusion to areas of the matrix where the values come from RSI data: this effectively implies a zero variance for the non-RSI Prior matrix movements, so that they will simply be copied into the fused matrix. Another alternative is again to confine the fusion to the "RSI-observed" areas and then to replace the other areas of the matrix, the details for which are lacking, with values estimated directly from the new data. This effectively implies an infinite variance for the non-RSI Prior matrix movements, so that they will play no part in the fused output. The final option considered here (again to address only the unobserved parts of the matrix), is to use the mean of the prior and new data to constrain the fusion (which would be consistent with the variances of the two sources being equal, if they were known).

Whichever approach is adopted, a number of trips, T_{IJ}, and a variance, V_{IJ}, can be calculated in an array for each IJ 'block' (each comprising a number of ij cells) over which fusion is to be undertaken. Thus T_{IJ} and V_{IJ} are of the same dimension, 1 x n, where n is the number of IJ 'blocks', in this case 184. Dividing each value of T_{IJ} by the total number of trips T in the fusion array **T**, we transform from trips T_{IJ} to **proportions** p_{IJ}. This

[12] East of England Regional Demand Model, Model Development Report.
[13] See Table 6.2 in the CSRM Technical Note 5, relating to the AM peak.
[14] according to National Travel Survey statistics, only 23% of all car trips are greater than 10 miles.

division by T divides all the variance terms by T^2. This provides the diagonal (variance) terms for the covariance matrix $\Omega_{IJ,IJ}$, which is much larger.

To clarify, the CSRM trip matrix comprises 325x325 ij cells. These have been aggregated into 29x29 sectors and further amalgamated into 184 IJ fusion blocks of amalgamated ij cells. Each of the 184 elements in **T** generates a row or column in **Ω**, so that **Ω** is a 184x184 matrix, with typical term $\Omega_{IJ,RS}$ say, where "IJ" and "RS" are two different IJ blocks of amalgamated cells of the prior trip matrix. We can view this as if we "string out" the IJ blocks in **T**/T into a column vector form with 184 elements. These are the 184 proportions to be fused with equivalent data from the new data sources. The diagonal elements in **Ω** represent the variances of the proportions p_{IJ}, while the off-diagonal terms are calculated as the co-variances between different IJ blocks (e.g. p_{IJ},p_{RS}).

For a pure multinomial distribution, the correlation[15] between the terms p_{IJ}, p_{RS} is given as:

$$\rho_{IJ,RS} = \frac{-p_{IJ}.p_{RS}}{\sqrt{p_{IJ}.p_{RS}(1-p_{IJ}).(1-p_{RS})}}$$

This can be calculated on the basis of the actual proportions, and then re-applied to the calculated variances to give the off-diagonal terms for the covariance matrix **Ω**, using the formula:

$$\Omega_{IJ,RS} = \rho_{IJ,RS}.\sqrt{\Omega_{IJ,IJ}.\Omega_{RS,RS}}\ .$$

3.4. Summary of Prior Matrix Processing

In summary, the prior matrix from the test-bed model was processed in the following manner:

1. sum original CSRM prior matrices across user classes 1-8, and sum further from zones to sectors, respecting the five CSRM matrix development areas A B H E X to give the light vehicle prior matrix **T**
2. allocate and amalgamate sectors to ensure statically significant and broadly consistent RSI sample sizes, making appropriate assumptions for non-observed areas, to yield the 184 "fusion blocks" shown in Figure 3-7
3. For the fusion blocks containing RSI data, obtain the sector-to-sector samples N_{IJ} for each fusion block and calculate the variances V_{IJ} using the formula given
4. Plot and estimate the relationship between V_{IJ} and T_{IJ} for the RSI-fusion blocks
5. Apply this relationship to the remaining (un-observed) fusion blocks (synthetic and EERM-based), further multiplying the resulting variances by two as an initial estimate of the reduced reliability, creating variants for sensitivity testing as introduced above and reported subsequently
6. Divide all 184 fusion block elements by T ($\sum_{IJ} T_{IJ}$) to give the proportions matrix p, and divide all corresponding variance elements of the fusion blocks (V_{IJ}) by T^2 to give the diagonal (variance) elements of the covariance matrix Ω.
7. Calculate the off-diagonal elements of the covariance matrix Ω, using the formula given.

The resulting array of 184 prior matrix proportions **p** and the associated 184x184 covariance matrix **Ω** were then used in the fusion process, with the corresponding elements from the new data sources, as described below.

[15] correlation(X_i,X_j) = covariance(X_i,X_j) / (variance(X_i) * variance(X_j))$^{1/2}$

4. TrafficMaster GPS Data

4.1. Data Sources

The available TrafficMaster (TM) matrix data consists of origins and destinations for the following vehicle types:

- Car
- LGV
- HGV

The TM datasets are distinguished by their excellent spatial accuracy but compromised by biases in trip purpose, as described below. A trip in the TM data is deemed to start and end according to the vehicle's ignition with mixed impacts on data quality – there is welcome certainty in knowing when the driver is 'in transit', but while the destination of a short drop-off trip may not be registered, some longer trips may be prematurely terminated at stop-off points such as service areas.

In each case the dataset comprises uncorrected sample data, which was obtained for the whole day but subsequently processed for the AM peak hour only. This relates to a year (2011-2) of weekday observations[16]. The dataset was obtained from the DfT coded to the National Transport Model PASS3 zone system[17], retaining only those X-X movements which have been identified as passing through the study area. However, to avoid identification, the first and last 500m of every trip had been removed at source such that any trips under 1 Km in length were missing from the data prior to receiving it. For all other trips, the distance (after adding back the missing 1 Km) reflects the average actual distance travelled between each O-D pair by the vehicle type concerned. The PASS3 zones were then re-coded to the CSRM system.

Given the focus of the study on improving the CSRM light vehicle matrix, HGV data was not processed further.

For cars and LGVs there are biases inherent in the data due to the way it is collected, ie largely from vehicle fleets, which undertake a relatively high proportion of business-related trips that are typically longer than average, as can be determined from comparisons with NTS data. The car dataset was analysed and corrected for trip length bias as described further below. However, unlike cars there is no firm basis for correcting the LGV matrices for trip length bias. The principal options comprise:

a. ignore LGVs, and treat cars as representing "all light vehicles", or

b. assume any bias corrections for car representativity apply to LGVs as well, and combine.

The problem with option b) is that to combine the LGV and Car samples one would need to weight for differential representativity. Although the total number of vehicles in each vehicle class in the TM sample is known and can be related to the total stock, we have no way of knowing whether a correction based on this information is appropriate for the Cambridge data. There is no guarantee that the Cambridge fleet composition (or more precisely, the composition of the fleet travelling through the study area) is consistent with the national data. Note that the downstream assignment and ME tests in this study require the use a single class of "light vehicles", so that no distinction is required for the matrix fusion: the aim is merely to get the best estimate of the zone-to-zone pattern. In this context option a) was chosen for this study[18].

[16] September 2011 to August 2012 school days only, comprising approximately 680,000 all-day records. Of these approximately 10% were shown as starting in the AM peak hour, 0800 to 0900.
[17] approximately 11,000 zones in Great Britain.
[18] Given that WebTAG unit M3.1 recommends LGVs to be modelled separately from cars, it is likely that in other practical studies there would be a need to check the trip length distribution of TM LGV data and adjust

A number of checks on the car matrix were carried out at the all day level, comprising an analysis of the approximate symmetry of the all day matrices (O-D) and a limited ratio comparison of the (uncorrected) trip ends with TEMPRO, with other checks and corrections undertaken specifically for the AM peak period. These comparisons are documented below.

4.1.1. All Day Trip End Symmetry

A comparison of Os and Ds over the whole day was carried out for the 325 CSRM zones. The correlation was 99.8%, and the mean absolute percentage difference was 6.3%. Most of the larger errors appeared to be related to adjacent zones and – in particular – service areas and places with overnight accommodation. Overall, the issue is not considered significant (especially given the intention of carrying out the data fusion at a more aggregate spatial level).

4.1.2. All Day Trip End Comparison with TEMPRO

The primary aim here was to "verify" that the "new" data did indeed bear some overall relation to travel patterns, Although TEMPRO Trip Ends cannot be treated as reliable in absolute terms, TEMPRO does include a) the latest data on population, b) a reasonable account of car ownership, c) a generally reliable account of trip rates by all modes, and d) some account of modal propensity. Hence a comparison between zones is useful even if considerable "noise" is expected. It is noted that TEMPRO does include some LGV trips.

All-Day TM Trip Ends were compared with three TEMPRO variables: Trip Ends, Population and Car Ownership. Generally, the closest correlation was found with car ownership, as shown below.

Table 4-1 Regression of All-Day TM Trip Ends against TEMPRO Car Ownership (2012)

y=mx	**TEMPRO Areas**	
	TEMPRO Zones	TEMPRO Districts
Gradient (*m)*	1.169	1.019
Correlation (R^2)	0.935	0.996

This provided some general re-assurance about the zonal variation in the TM data.

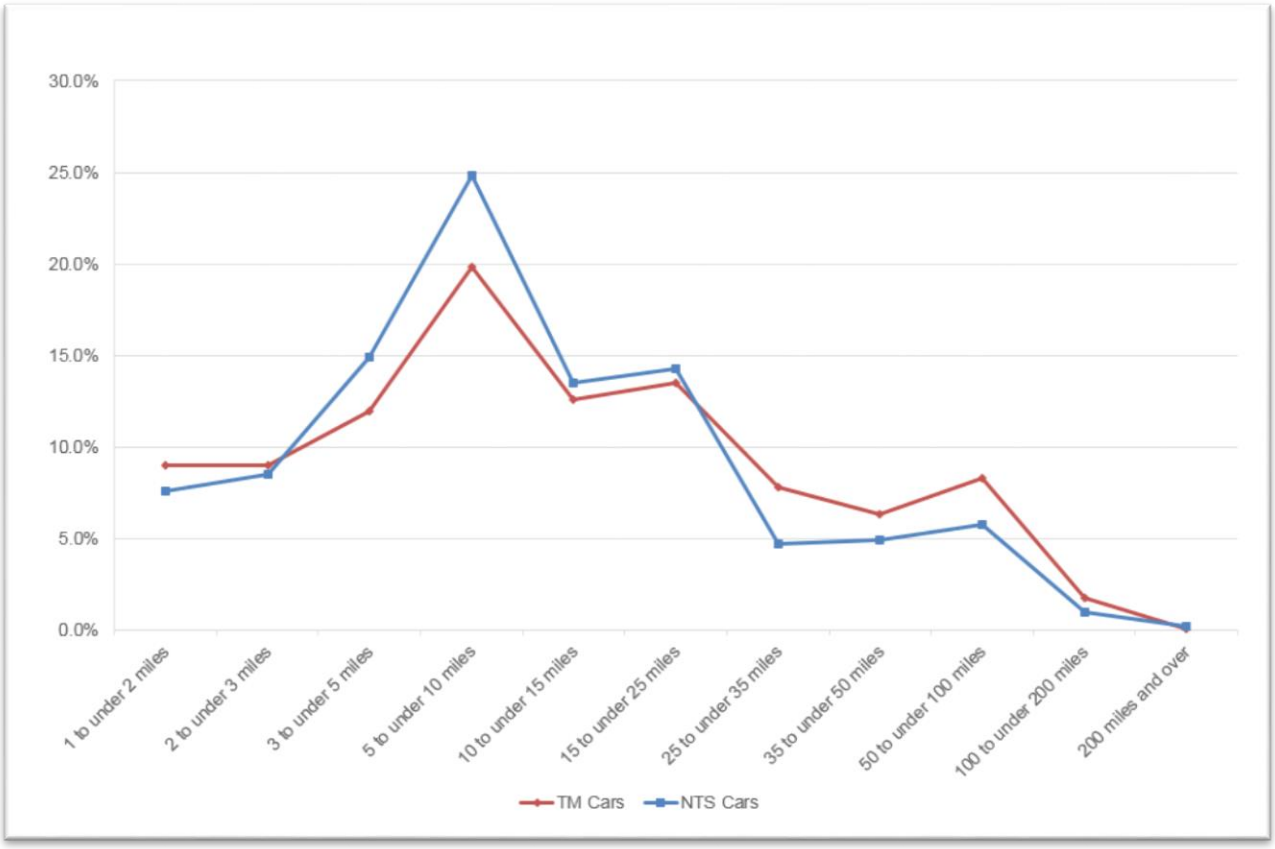
4.2. Correction of AM Peak Car Trip Length Distribution

2002-11 NTS data for AM peak car driver trips with at least one trip end in Cambridgeshire[19] has been compiled into the standard NTS distance bands. While it has been claimed that nationally the TM TLD is very similar to that from NTS, the analysis suggests that the TM data for Cambridgeshire (NB excluding the X-X movements in the TM car matrix for compatibility) tends to under-represent shorter distance trips under 25 miles (particularly so for trips under 1 mile) and over-represents trips over 25 miles, as shown in the figure below, reflecting the fact that TM is biased towards fleet vehicles typically travelling further than average:

for bias as necessary. It remains unclear what could be used as a reliable data source for this, but it is noted that the work-related biases observed for TM car data would be much reduced when considering LGV trips, as a far higher proportion of LGV trips are for work purposes anyway.

[19] This is the closest approximation to the study area available using NTS data

Figure 4-1 Car Trip Length Distributions (AM Peak) for TM (2011-2, Study Area) and NTS (2002-11, Cambridgeshire)[20]



On this basis, a correction was applied for all distance bands except those less than one mile – the lowest standard distance category available in NTS (because of the anonymisation referred to above whereby trips less than 1 Km are removed). This correction has been applied to the X-X trips as well. The total of the affected trips has been rescaled to equal the original sample. This constitutes the corrected output.

Correction factors for trips below 100 miles varied between 0.60 and 1.25. At the longest distances the factors could be expected to be greater due to the fact that the TM trip "ends" when the ignition is turned off (because of stops). This can be seen for trips >200miles (though the samples are very small), but for trips 100-200 miles the correction factor is 0.54, implying that at this distance trips are still being over-sampled.

Note that no corrections have been made in relation to movements less than 1 mile; the uncorrected and corrected matrices are identical for these movements–. Most of these will be intra-zonal, and they are likely to be substantially underestimated[21]. Given the fact that the intrazonals do not impact on the network, and that even for interzonals, trips less than 1 mile are unlikely to have any influence on the matrix estimation, it was proposed to ignore them in the fusion. Thus, the proportions of TM trips within each 'fusion block' were calculated excluding trips less than 1 mile[22] in length.

As a result of this, when fusing TM data with the Prior matrix, trips less than 1 mile were excluded from both data-sets. Therefore any matrix cells representing trips under 1 mile within the network model were removed from the fusion process – i.e. their cell values were omitted from the appropriate fusion-block figures, with

[20] Excludes external trips (i.e. between areas X and X) even if they pass through the study area.
[21] an analysis of weekday car driver trips in the 2010 NTS data set, where distance is recorded to the nearest 1/10th of a mile showed that 79% of trips under 1 mile are less than 0.6 mile (approx. 1 Km)
[22] The Trafficmaster AM peak hour sample excluding trips under 1 mile and before correction of the trip length distribution was 68,141

sample sizes, proportions and resulting variances recalculated for both the Prior matrix and the TM matrix[23]. After fusing the Prior and TM fusion-block proportions, the resulting outputs were distributed within fusion blocks, using whichever distribution had been defined for that test, but excluding the short distance cells. Then the short distance trips from the original Prior were added back in. In practice, this is a minor variation: excluding intrazonals (which are of course not assigned), about 4% of trips in the AM peak prior matrix are less than one mile.

For the purpose of determining which trips fall below the 1 mile threshold, distances can be calculated either from the traffic model network or using crowfly distances. Both methods were tested, leading to only a 0.2% variation in the number of trips lost. Ultimately, the crowfly distance was used for the data processing, on the grounds of convenience.

4.2.1. AM Peak Trip End Comparison with TEMPRO

After applying the TLD correction the AM Peak TM Trip Ends were again compared with the three TEMPRO variables, and again the highest correlation was found with car ownership, as shown below: The figures are very similar to those presented above for the regression of unprocessed TM data with TEMPRO car ownership data.

Table 4-2 Regression of AM Peak TM Trip Ends against TEMPRO Car Ownership (2012)

y=mx	TEMPRO Areas	
	TEMPRO Zones	TEMPRO Districts
Gradient (*m*)	1.168	1.030
Correlation (R^2)	0.902	0.991

4.3. Variance Assumptions

Because of the correction to the sample proportions, the standard calculations for sampling variance are not valid. Instinctively, it would seem that the bias correction should produce more accurate central estimates, while at the same time increasing the variance. A full treatment would require the variance of the bias correction factors, and it is not clear how this could be derived. Hence, on practical grounds, the following approach was followed:

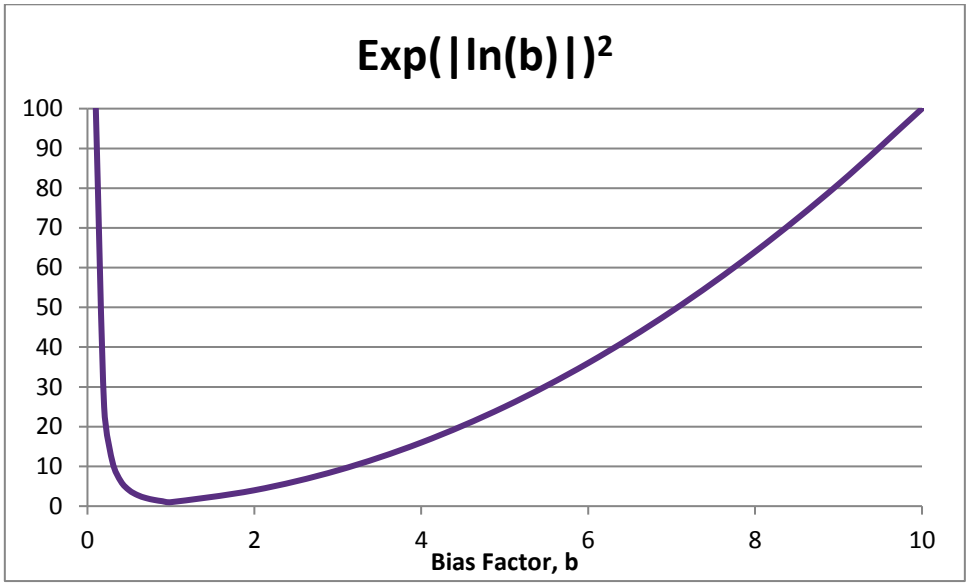
Assume that the corrected proportions matrix has the form $b_{IJ}.p_{IJ}$ (ensuring that it adds to 1), with the unprocessed proportions being p_{IJ}, so that b_{IJ} can be derived as the quotient of the processed and unprocessed proportions matrices. Then it is reasonable to suggest that b_{IJ} will affect the variance. Indeed, if b_{IJ} were a constant then the variance of $b_{IJ}.p_{IJ}$ would be calculated as $b_{IJ}^2.Var(p_{IJ})$. Instinctively factors of, say, 2 and ½ should be treated in the same way, suggesting that the adjustment to the variance should be based on $|\ln(b_{IJ})|$. Consequently, the following relationship has been assumed:

$$Var(b_{IJ}.p_{IJ}) = \exp(|\ln(b_{IJ})|)^2.Var(p_{IJ})$$

The function $\exp(|\ln(b_{IJ})|)^2$ has reasonable properties, with a minimum value of one for b = 1, but then increasing as the square of b (or 1/b for values of b<1), so that for b = 2 (or ½) the value is 4, as shown in the figure below.

[23] As this correction was not needed when fusing the Prior with INRIX data, this means that two variants of the Prior matrix data – with and without short trips included – were required

Figure 4-2 Hypothesised relationship between the correction factor (b_{IJ}) and the impact on the variance var(p_{IJ})



Hence it was proposed that the variance is multiplied by a factor of $\theta.(\exp(|\ln(b_{IJ})|))^2$ where θ is a further scalar to allow flexibility but provisionally set to 1. In this case, the variance remains as for the multinomial assumption when $b_{IJ} = 1$, but as b_{IJ} moves away from 1 it would be progressively increased. For b = 2 (or ½) the variance would be multiplied by a factor of 4 for b = 4 (or ¼) by 16.

Hence, starting with an initial (sample-based) estimate of p_{IJ} and correcting this for bias by a further b_{IJ}, the variance for the processed proportions matrices $b_{IJ}.p_{IJ}$ would be calculated as:

$$\theta.\left[\exp\left(\left|\ln(b_{IJ})\right|\right)\right]^2.p_{IJ}.(1-p_{IJ})/N$$

It is noted that given the TLD bias correction factors derived for the Cambridgeshire TM data, this correction to the variance would not be expected to have a large impact. The TLD adjustments which were applied vary between 0.54 and 1.25[24]. At sector level these extremes will be very diluted so b will typically not be far from one.

The off-diagonal terms in Ω were calculated along the lines of the prior matrix, as described in section 3.3 of this report: derive the correlation terms (NB in terms of the uncorrected proportions p_{IJ}), and then apply these to the calculated variances.

Hence, the processing of the TM dataset can be summarised as follows:

1. Remove trips < 1 mile
2. Assume car trips alone to be representative of Light Vehicles
3. Sum unprocessed matrices from zones to sectors to fusion-blocks and in the light of Step 1 calculate the proportions, p_{IJ}
4. Sum processed matrices from zones to sectors to fusion-blocks and in the light of Step 1 calculate proportions and hence correction factors b_{IJ}
5. Apply the recommended relationship for the diagonal elements of the covariance matrix **Ω**.

[24] And up to 4.81 for the few trips >200 miles in length

6. Calculate the off-diagonal elements of the covariance matrix **Ω**, using the formula given in Section 3.3.

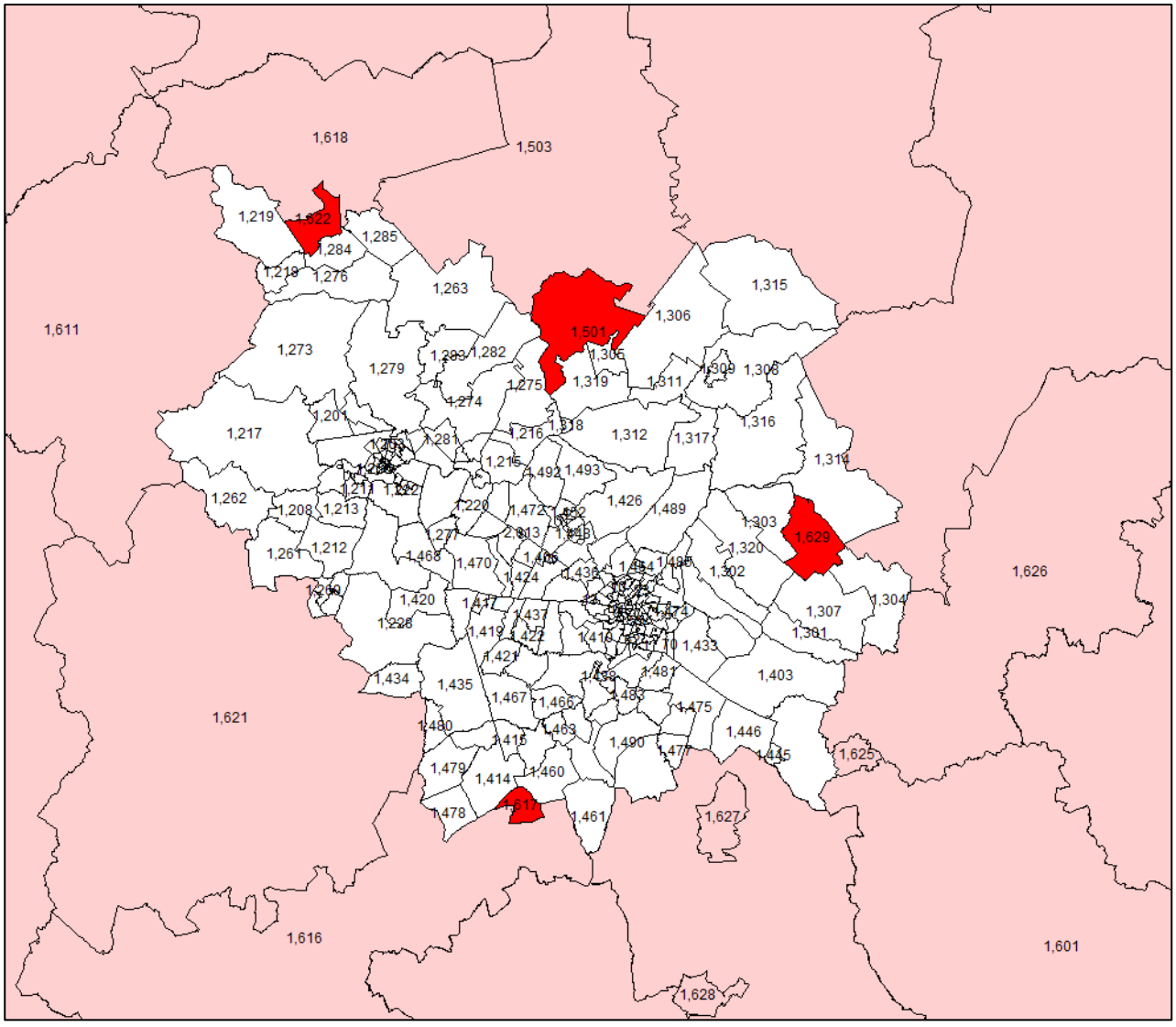
As discussed subsequently, once the TM proportions had been fused with the Prior proportions (also omitting trips <1 km), the output proportions were then expanded and the Prior matrix trips under 1 km were added back in, to yield the newly updated Prior matrix. Thus trips <1 km were ignored in the fusion but not omitted from the wider process.

5. INRIX Mobile-phone Data

5.1. Data Sources

Mobile phone data from Telefonica, sourced via the data brokers INRIX, was received for all movements entering, leaving or within an area similar to the “extent of modelled area” as shown in Figure 5-1 (itself taken from the CSRM LMVR). The precise boundary was defined as that between the pink coloured zones and the other zones shown in the figure below. External trip ends are not identified, but instead the zone at which the trip enters or leaves the study area is identified. This made the data broker’s analysis of the raw mobile phone data more manageable and the data provision more affordable.

Figure 5-1 Geographic Coverage of INRIX Mobile Phone Data[25]

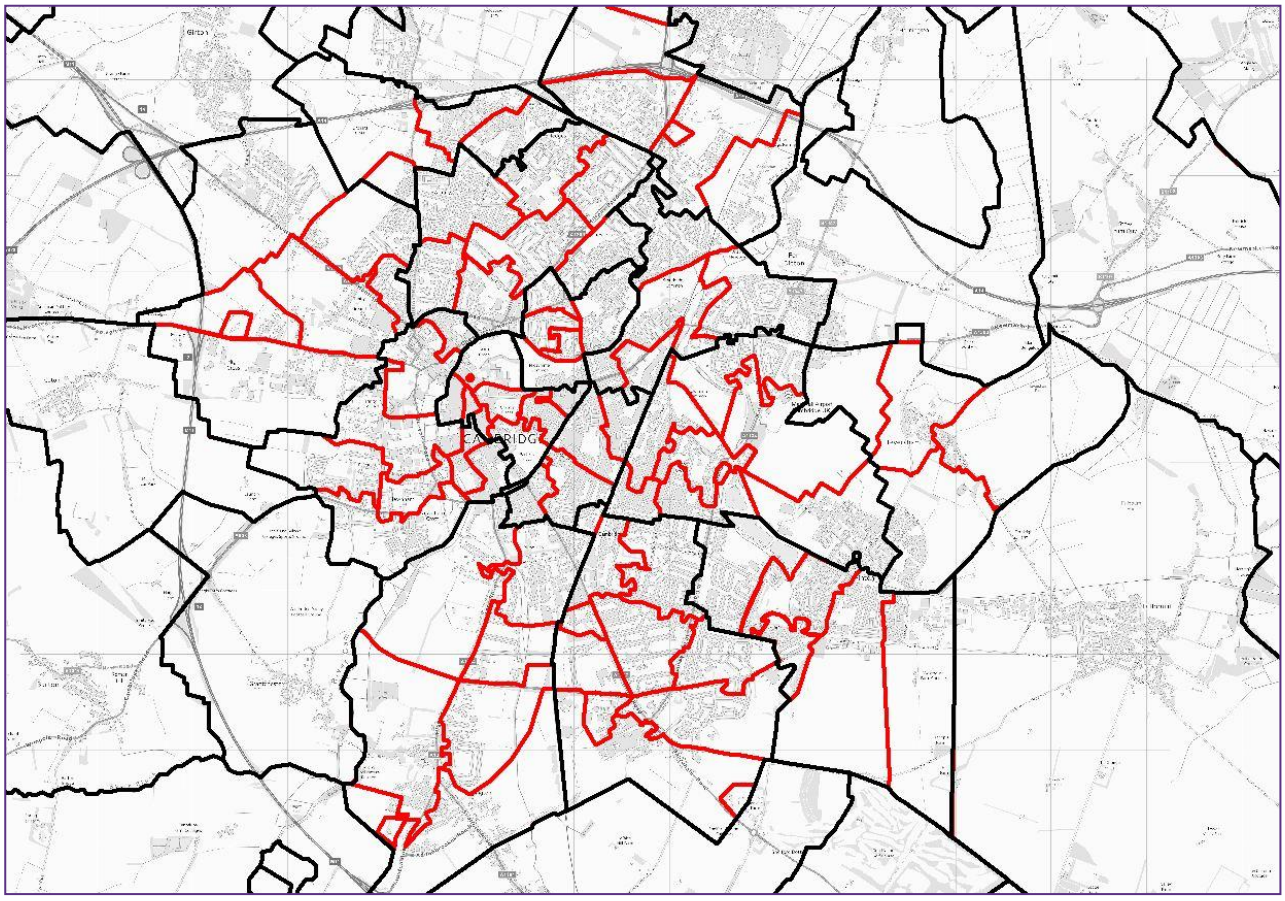


[25] The white zones are those defined as internal (i.e. covering super-sectors A, B, H and E) in the CSRM matrix development process. The red zones are additional (originally external) zones included in this study in order to smooth the boundary and to capture additional observed trip ends in conurbations on the very edge of the internal area.

The data is based on the traced movement of phone handsets, the key benefit of which is a very large sample indeed. However, there are numerous practical problems of interpretation, some similar to those for GPS-based data but with additional problems owing to a lower level of spatial accuracy. By purchasing more recent data and analysis it may now be possible to receive mobile phone data that much more closely approximates highway vehicle movements[26].

Due to the location of mobile phone masts and the consequent level of spatial detail at which the mobile phone data could be provided, there were several locations where two or more CSRM zones were represented by a single mobile phone zone. By far the most significant of these was in Cambridge itself, where 70 CSRM zones were aggregated into 14 mobile phone zones. The figure below shows how the (red) CSRM zones map to the (black) mobile phone zones provided by INRIX.

Figure 5-2 Zone Aggregation for Mobile Phone Data in Cambridge



Initially the data contained all detected movements of phone handset signals (known as 'probes') within the study area, that were *between* the aggregated INRIX zones; intra-zonals were ignored. A trip end was identified after the signal from the handset had failed to move for an hour.

[26] Note that the INRIX data was acquired Q2 of 2013, and that this sphere of use of mobile phone data has moved on substantially since that time. Mobile phone data procured now may have greater spatial detail and the data providers are developing more accurate algorithms to isolate active modes, HGV users and PT passengers. Moreover, the home and work end of the trip can be readily 'inferred'. Examples of the improvements being made are the tracing of multiple 'probes' (corresponding to travellers' handsets) on a single trajectory; the records can be compressed to a single movement, dealing with the cases of multiple occupant cars and – more importantly – public transport. However, passengers making different OD movements in the course of a single bus (or, to a lesser extent, car) journey are still liable to be recorded as separate "vehicle" trips.

The data received consists of individual records of movements, containing the following items of information for 5 weekdays (all day) in May 2013:

- Anonymised ID
- Day
- Beginning of start hour
- Start time period
- Origin CSRM zone
- Destination CSRM zone
- Flag for external Origin
- Flag for external Destination

After initial investigations, an improved version of the data was obtained, reflecting the following changes:

- For trips which start and end within the same zone, the point in the trip which was furthest from the origin was identified. If this point was outside the origin zone, and above a statistically significant distance for that origin zone, then the trip was broken into two separate trips, to and from that point. (Note that this does not consider the route of the trip (a potential future improvement) and that it added only a very small number of new trips)
- Inclusion of intra-zonal trips which move a significant distance within a zone. Note, these trips should not be considered representative of all internal movements, as the technology is not suited to identification of most short trips
- For trips which display a stopped time of 30 minutes (where sufficient data exists to show the probe as stopped) those trips (which were previously at a 1 hour temporal granularity) were split into 2 separate smaller trips.

This final improvement had a surprisingly small impact on the total number of trips provided, for which there are a number of possible explanations:

- the previous INRIX algorithm already detected some stops under an hour, as stops at the same cell tower were detected
- there may be relatively few probes which provide sufficient off-call data to be sure they have stopped for between 30 and 60 minutes
- there may really be relatively few trips with stops of this duration (e.g. drop-off trips will almost invariably be stopped for an even shorter period).

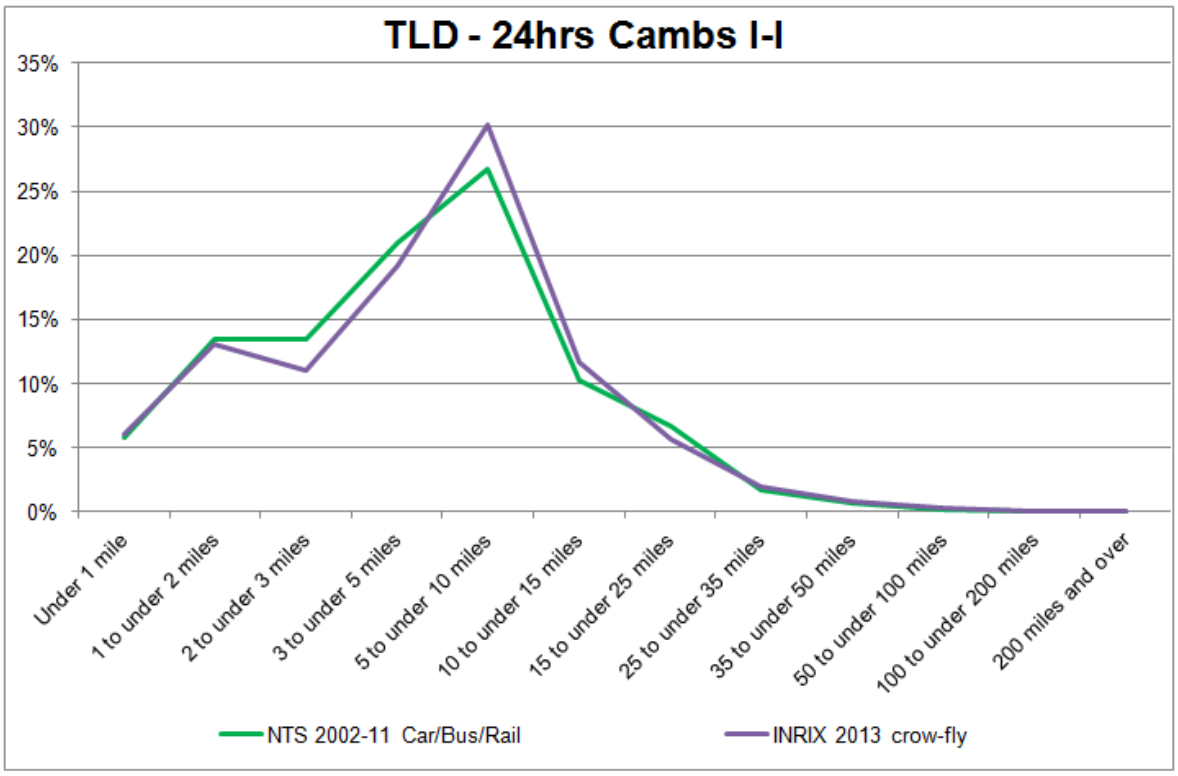
With the intention of removing trips made by “slow modes”, speed thresholds have been applied. Only trips exceeding the average speed threshold have been retained as likely motorised mode trips. Within Cambridge, a trip speed distribution was calculated from the 2006 AM CSRM, based on crow-fly distances. This indicated that an average speed threshold of 12kph would filter out relatively few motorised vehicles[27],

[27] While TomTom journey time data for around 800m of the most congested inbound leg of Histon Road suggested that over short distances some vehicle average speeds will dip below this, it was felt that this would be the exception and that trips picked up in the INRIX data (i.e. between the 14 sectors of Cambridge) will typically be far longer than the 800m congested section surveyed in TomTom, thus giving rise to higher average speeds.

while removing a significant proportion (but still less than half) of the cyclists[28]. Outside Cambridge, as the *Cyclestreets* journey planning website for Cambridge suggests a "Quick" average speed threshold of 24kph, this was used to filter out the slowest trips starting or ending outside Cambridge itself, to distinguish cycles from cars. This is expected to remove the vast majority of the cycles and lose very few motorised vehicles. Without much more sophisticated data processing (following individual routes etc), it was not thought possible to improve on this filtering approach using the available data. Both thresholds were applied by INRIX within the mobile phone data processing algorithms, prior to receipt of the data. The data still contained public transport and goods vehicle trips.

With regard to checks, a 24hr TLD comparison with NTS was undertaken using CSRM zone centroid coordinates of origin and destination (and the crow-fly distance between), for all trips internal[29] to the study area. The resulting distribution was compared against NTS data internal to Cambridgeshire as shown below. A good match was found, which did not reveal any material bias. This comparison was based on unprocessed data as received from INRIX; a further comparison is reported below, using the processed INRIX data at the AM peak level.

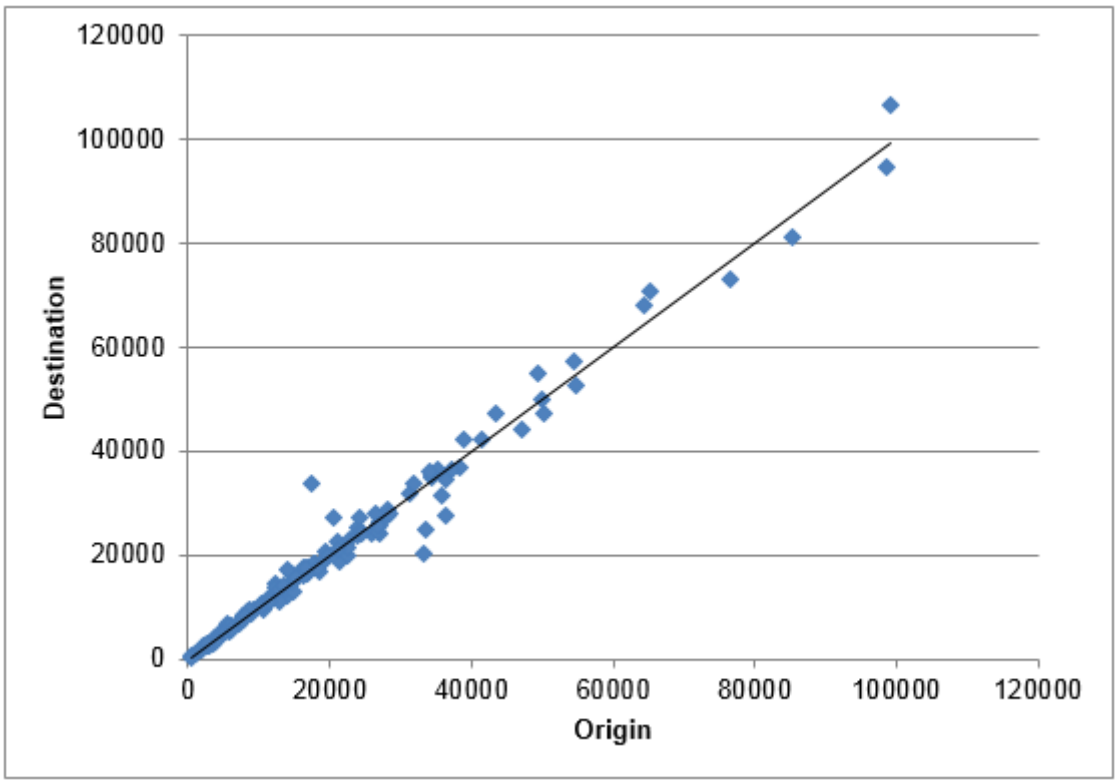
Figure 5-3 Trip Length Distributions (24hr) for INRIX (2013, Study Area) and NTS (2002-11, Cambridgeshire)



A further check on matrix symmetry showed a correlation coefficient between origins and destinations of 0.99, with an average difference of 7% between origins and destinations for internal (I-I) movements. All subsequent work focused on the AM peak period only.

[28] The *Cyclestreets* journey planning website for Cambridge estimated typical speeds in Cambridge as follows: 'quick' equates to 24kph, 'cruising' to 20kph and 'unhurried' to 16kph, suggesting that a 12kph threshold is well below the average cycle speed.
[29] The comparison was wholly internal (unlike Figure 8 for TM data) as the INRIX data does not include external trip ends.

Figure 5-4 Comparison of daily trip symmetry for weekday internal movements



The data received for this study was based solely on analysis of origins and destinations and (except for the aforementioned filtering of slow modes) includes all modes of transport.

In order to correct for the inclusion of public transport trips, the following approach was adopted:

1. From CSRM[30], calculate the ratio of AM peak period highway vehicle trips to mechanised mode person trips
2. Apply 29^2 such factors (one for each sector-sector movement) to the appropriate areas of the AM peak period INRIX sample matrix[31]
3. Expand the result from step 2 to a 2006 AM peak hour vehicle trip matrix using the Prior Matrix to control on a super-sector basis (5x5 sectors).

Thereafter, the HGVs were removed. It was agreed that the INRIX GPS-based HGV matrix that was supplied alongside the mobile phone data should not be used for this process. This was to ensure the greatest consistency with the HGV component of the existing CSRM matrix, so that the analysis of changes to the matrices could be focused on light vehicles alone. Hence, the procedure followed was to remove HGVs using the CSRM Post-ME HGV AM peak matrix. The subtraction of HGVs from All-vehicles on this basis did result in some negative numbers, but when aggregated across the 184 fusion-blocks the resulting Light vehicles matrix did not contain any negative numbers (due to the small proportion of HGV in the total vehicle matrix).

There are further issues relating to the INRIX treatment of external origins and destinations. The INRIX data does not recognise zones beyond the CSRM internal area boundary, and codes them to the boundary zone. However, it also attaches a marker indicating whether movement beyond that zone was detected (at either

[30] CSRM data from the 2011 AM peak period was used – this better reflects the current proportions of public transport trips to subtract from the INRIX data (itself collected in 2013).
[31] In practice this was the matrix of trips after disaggregation to ultimate external trip ends, discussed below.

end of the “trip”). Hence, for a movement between two boundary zones X and Y, there are four possibilities that can be identified:

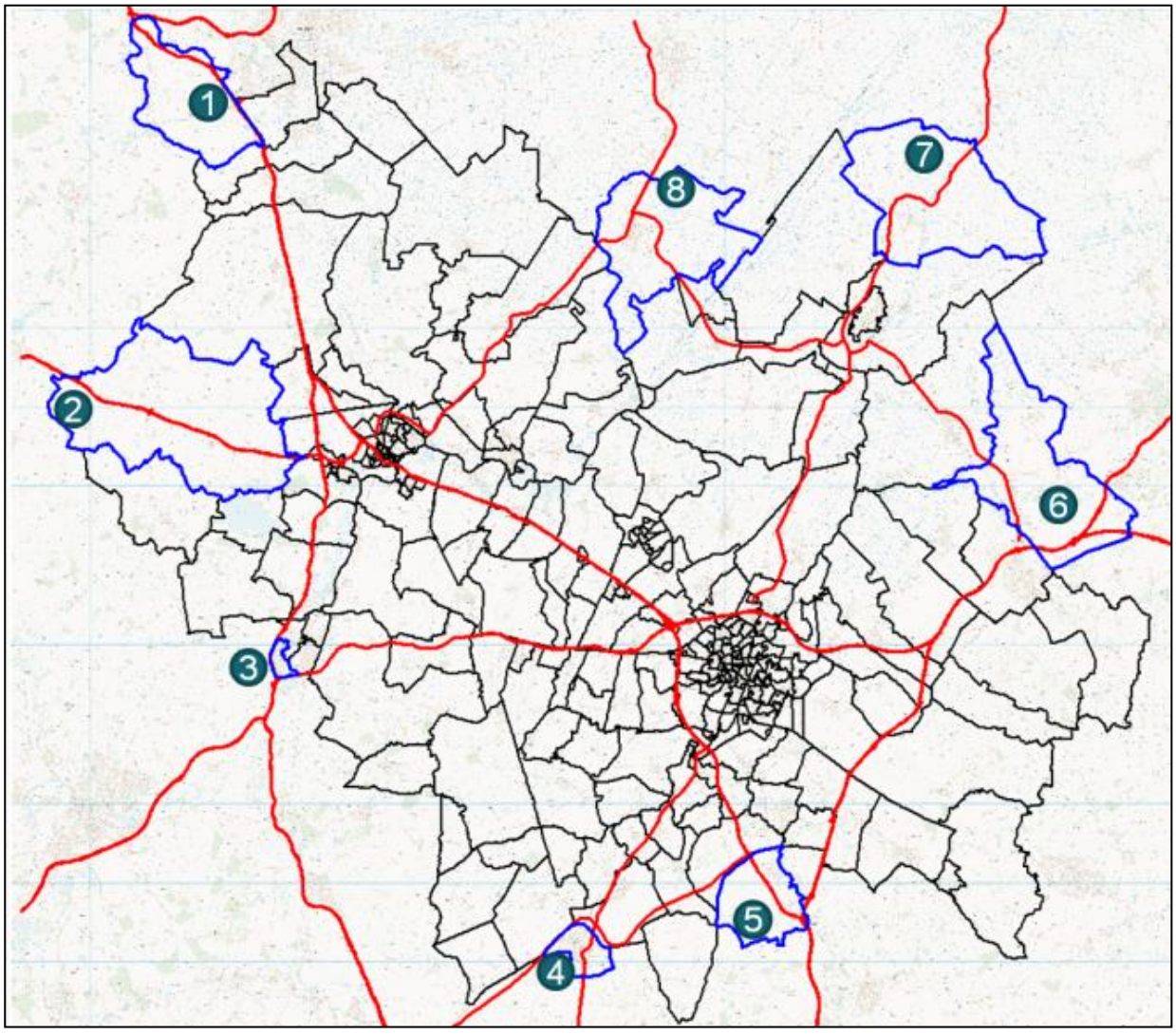
- true movement between X and Y
- movement from X via Y but continuing beyond
- movement starting before X passing via X to Y
- movement starting before X passing via X and Y and continuing beyond.

The last three possibilities need to be appropriately allocated. Two approaches were investigated.

The first, simpler approach is to use select link (SL) matrices from the CSRM at the internal area boundary, to determine the proportion of trips entering (or exiting) through a particular boundary zone which originate (or are destined) for each CSRM external zone. This could be done using a single inbound and a single outbound cordon select link matrix: that would, however, imply the same distribution of external trip-ends for trips using the M11 to the south or the A14 to the NorthWest. This is a significant and undesirable assumption.

The eight areas through which external trips travel, as identified in the INRIX data, are shown in the figure below.

Figure 5-5 Cordon Crossing Points used in Select Link Analyses to Attribute INRIX Trips to External Zones



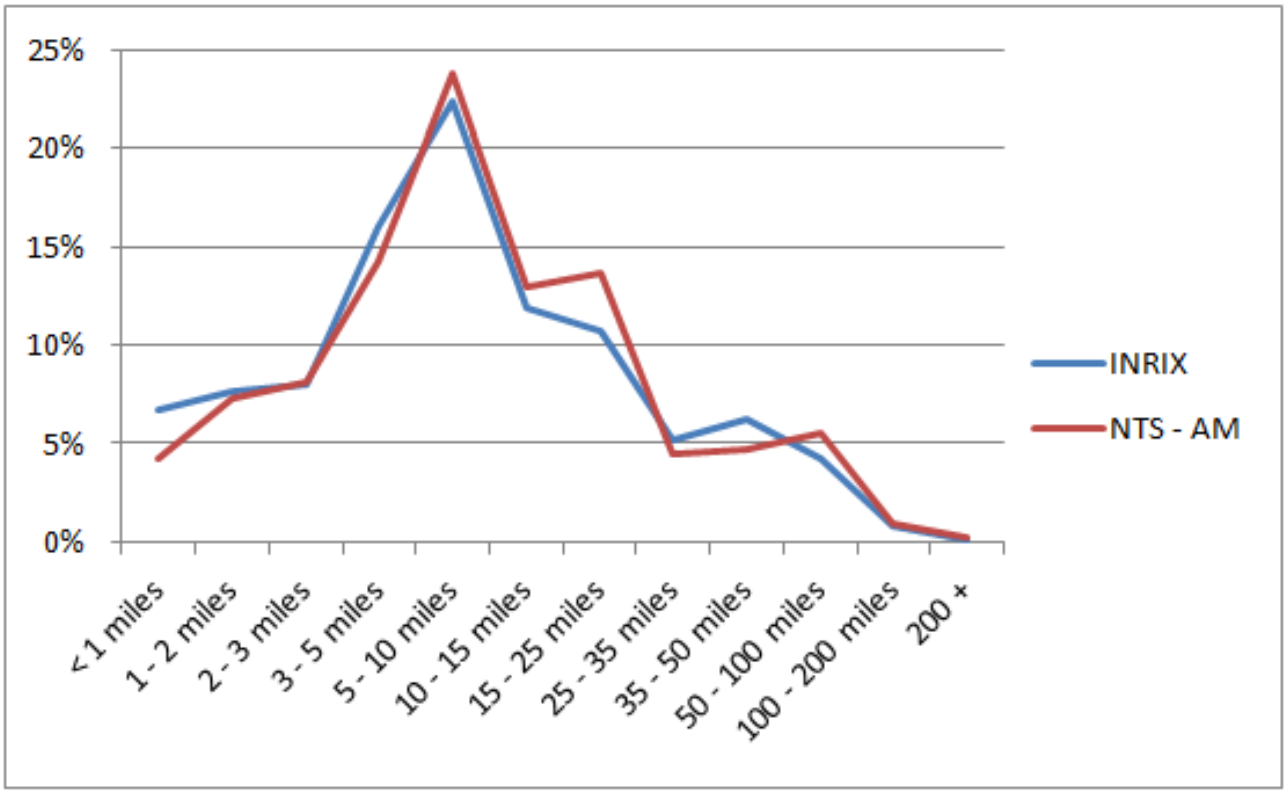
The more comprehensive method of dealing with external trip ends involves building IN and OUT select link matrices for each of the 11 strategic network links cutting the boundary[32]. External-internal trips would take their external origin zone distribution from the appropriate inbound SL matrix. Internal-external trips would take their external destination zone distribution from the appropriate outbound SL matrix. External-external could assume the same external trip end distributions as determined from the internal-external and external-internal SL data but it would be more accurate to deduce external-external trip end distributions, specific to (say) IN from M11 and OUT via A14 from a SL matrix cutting the inbound M11 and outbound A14.

A rigorous treatment of external-external trips would involve 11 IN x 11 OUT, thus 121 SL analyses. In reality, some 104 SLs have been undertaken for external-external trips as a) a SL matrix from link l to link l is not useful and b) trips between three separate pairs of boundary crossings are not relevant to the model (i.e. those between the pairs of crossing points in areas 1, 5 and 6).

It is noted that the INRIX data may contain many trips peripheral to the study area which are not relevant to the CSRM model network. Such trips may be detected on secondary or minor roads crossing the study area boundary, which themselves are not modelled. It may be appropriate to omit such trips to prevent the modelled links in these peripheral areas from becoming overloaded. These trips have been identified as trips passing between neighbouring boundary crossing points that are unlikely to have travelled through the modelled network in this area, for example trips from location 3 to location 4 in the above diagram. Identification of these movements was done by comparing results from select link analyses between the external crossing points in the CSRM and assigned INRIX data. Those movements that had no, or minimal, demand in the CSRM were then removed from the INRIX data.

The final step in the processing of the INRIX data was to compare the resulting trip length distribution with that from the NTS for car journeys in Cambridgeshire, as in the figure below.

Figure 5-6 Trip Length Distributions (AM Peak) for Adjusted INRIX and NTS (2002-11)



[32] Note there many more minor routes which cross the study area boundary but all bar a few are not included in the network model. Of the few minor roads that are included, the INRIX data shows no AM peak trips.

The figure shows the trip length distribution for all movements with an internal trip end (in other words, internal-internal, internal-external and external-internal movements). The trip length adjustments were then applied to the full matrix (i.e. including external-external movements).

5.2. Variance Assumptions

As with the TM data, any form of statistical fusion relies on the variances being reasonably estimated; again the focus is on proportions rather than trip volumes. Ideally, the variance of the uncorrected data should not be viewed in straightforward sampling terms, as much of the sample is built from the same travellers observed over many days. However, in the absence of a better approach it was considered reasonable to assume this. The "removal" of the PT and HGV records and the factoring required to replicate the relevant NTS TLD all affect the data in various ways, including a reduction in the total INRIX sample size. It is not clear how best to deal with this. From the point of view of maintaining consistency of methodology, the change in proportions brought about by the removal of PT and HGVs is viewed as a further bias correction in the same way as the distance-based TLD correction is treated (as explained in the preceding chapter for TM data). For this purpose, the original INRIX sample size is retained. Hence the calculations for the variances are the same as those in the "Variance Assumptions" section in the TM chapter, namely:

$$\theta.\left[\exp\left(\left|\ln(b_{IJ})\right|\right)\right]^2.p_{IJ}.(1-p_{IJ})/N$$

Where Θ is a scalar discussed below, p_{ij} are the corrected proportions, N is the original (i.e. uncorrected) sample size and b_{ij} is the bias adjustment factor defined as:

$$\frac{p_{IJ}(Corrected)}{p_{IJ}(Uncorrected)}$$

The corrected proportions were those post-removal of PT and HGV trips and after correcting the TLD. The uncorrected proportions exclude the effects of these changes, being based directly on the data received from INRIX. The uncorrected proportions therefore exclude those walk and cycle trips which had already been removed by INRIX using speed filters. The above ratio also reflects the impact of the allocation of external trip ends to external zones as described above.

Given that, even after adjusting for public transport and HGVs, there will still remain trips which do not relate to cars or LGVs, as well as potentially some loss of vehicles due to the cycle speed cut-off, an additional blanket factor Θ was applied to all elements in the INRIX variance matrix. This also accounts for the potential underestimate of variance resulting from use of a single combined bias correction, whereas in reality the effect of bias corrections on the variance could be applied incrementally, which would give a greater increase in variance in some cases[33]. A value of Θ of 1.5 was adopted. It is clear that this is a gross simplification and

[33] As an example, consider a case where the different bias corrections are:
b1 - (all mechanised modes to highway) – 0.8
b2 - (highway to light vehicle) – 0.9
b3 - (TLD correction) – 1.1
In this case the overall factor is 0.8 * 0.9 * 1.1 = 0.792 and the calculated effect on the variance would be to increase it by $1/0.792^2 = 1.59$. But if the formula is applied sequentially so that the variances associated with rises as well as falls are considered individually, then the effect on the variance would be to increase it by $1/(0.8^2) * 1/(0.9^2) * 1.1^2 = 2.33$.
In this particular instance the assumed additional Θ factor 1.5 would increase the as calculated variance (1.59) to that which would result from applying the bias corrections incrementally (2.33). While this is coincidental, the choice of individual bias correction factors is realistic and therefore the value of Θ is seen to be reasonable. Note that wherever the bias corrections are all in the same sense – i.e. all >1 or all <1 – then the effect on the variance is the same whether or not the bias corrections are considered incrementally. In such cases the Θ factor can be considered as accounting only for the uncertainties in the removal of PT and HGV trips and the TLD factoring.

there will be value in developing a more sophisticated approach to the estimation of non-sampling variance in mobile phone data data in future studies.

In summary, the key processes applied to the INRIX dataset can be listed as follows:

1. Remove walk and cycle trips using speed thresholds (in this case undertaken by INRIX prior to provision of the data)
2. Attribute all external trip ends to appropriate external zones
3. Remove public transport trips using the proportion of PT to total mechanised trips found from the CSRM
4. Expand the resulting INRIX (vehicle) AM peak data to 2006, using the Prior as a constraint
5. Subtract the CSRM Post-ME HGV AM peak vehicle matrix
6. Compare the TLD of trips with internal trip ends from the resulting light vehicle matrix with that for cars from NTS and, accordingly, apply corrections to the whole matrix
7. Sum uncorrected INRIX AM peak matrices (from item 2 above) from zones to fusion blocks and calculate the proportions matrix p, along with the total sample size N
8. Sum corrected matrices (from item 6 above) from zones to fusion blocks and calculate the corrected proportions matrix and hence the correction factors, b
9. Apply the recommended variance calculation to yield the diagonal of the variance matrix Ω.
10. Calculate the off-diagonal elements of the variance matrix Ω, using the formula given previously in the context of the TM variances.

6. Matrix Fusion Methodology

6.1. Matrix Fusion Theory

The fusion equation for any two alternative matrix estimates t_1 and t_2 is taken directly from Equation 6 in Skrobanski et al. (2012), shown below. (Note that the equation can be used in terms of either trips or trip proportions (summing to one) and the latter of these was used throughout the matrix fusion process[34]).

$$t' = \left(\Omega_{t_1}^{-1} + \Omega_{t_2}^{-1}\right)^{-1}\left(\Omega_{t_1}^{-1} t_1 + \Omega_{t_2}^{-1} t_2\right)$$

where **t′** is the fused matrix, and **Ω** represents the covariance matrix. Suppose fusion were to be undertaken for each and every ij cell in a trip matrix representing k zones; the diagonal elements in **Ω** represent the variances of the k^2 proportions p_{ij}, while the off-diagonal terms represent the covariances between different cells (e.g. p_{ij},p_{rs}). If there are k^2 cells in the trip matrix[35] **t**, there are k^2*k^2 elements in the covariance matrix **Ω**.

Three important facts should be noted as discussed below.

6.1.1. Simple Inverse Variance Weighting

Firstly, if **Ω** contains no "off-diagonal" terms (these are the covariance terms between different matrix cells), then the equation in fact represents the standard result that for any cell the two independent estimates should be linearly combined with weights in inverse proportion to their variances. In fact, we understand that for the work reported in Skrobanski *et al*., the covariance matrices were assumed to be diagonal. In this case, the above equation can be written out, for any matrix cell ij, as:

$$t'_{ij} = \lambda_{ij} t^1_{ij} + (1-\lambda_{ij}) t^2_{ij}$$

Where $\lambda_{ij} = \dfrac{V^2_{ij}}{V^1_{ij} + V^2_{ij}}$, $1-\lambda_{ij} = \dfrac{V^1_{ij}}{V^1_{ij} + V^2_{ij}}$ and *V* are the variance terms, and the variance of the output fused estimate, t'_{ij} is given by:

$$V'_{ij} = 1 \Big/ \left[\frac{1}{V^1_{ij}} + \frac{1}{V^2_{ij}}\right]$$

6.1.2. Use of Full Covariance Matrices

Secondly, an abiding problem is the size of the covariance matrices. With N zones, there are N^2 zone pairs, so that the covariance matrix contains N^4 elements. With our example of N = 325, the full covariance matrix has of the order of 10^{10} cells which, even allowing only for single precision arithmetic, requires more than 40GB of storage. By contrast, restricting the covariance matrix to the diagonal terms reduces the size to the order of 10^5 cells.

[34] In this study at least, the "new data" being fused does not contribute in any useful way to an estimate of total traffic, since there is no clear basis on which either the TM or the INRIX data has been sampled, meaning that the effective sampling fractions cannot be determined. Hence the fusion is undertaken at the sample level (of proportions), rather than the expanded level (of trips), recognising that after fusing, the output proportions will need to be expanded.

[35] In this context the trip matrix **t** is in fact a column vector of k^2 elements, corresponding to a square matrix of k zones.

6.1.3. Inversion of Covariance Matrices

Finally, in the case where the terms in the variance matrix are compatible with a pure multinomial distribution (as would be appropriate if the sampling rates were the only contribution to the variance), it turns out that the full matrix **Ω** (i.e. including off-diagonal terms) is not invertible in the classic sense (otherwise known as a singular matrix), because its rank is one less than the number of rows and columns[36]. The inversion is a requirement of the main equation in Skrobanski *et al*'s paper.

Deterministic techniques such as the Moore-Penrose pseudo-inverse, which can be calculated in MATLAB (pinv function), can take the covariance matrix and facilitate a result which has some of the properties of the inverse. However, these were discounted since they give no guarantee that the result is sufficiently close to the inverse itself, especially if the covariance matrix is singular (otherwise known as ill-conditioned output). And while perturbing the elements of a singular matrix even by a small amount makes it non-singular and thus theoretically possible to invert, this is likely not be very useful in practice since though it will allow computation of the inverse, it is often ill-behaved and inaccurate, having a high condition number[37].

That said, there is a possibility of returning well-behaved inverses and so perturbation techniques were chosen to be implemented in MATLAB in order to compute matrix inverse approximations.

However, to avoid ill-behaved inverses in practice, a set of randomised perturbation tests needed to be undertaken (five such tests were undertaken in each case, akin to Monte-Carlo simulations) but with the possibility that the resulting outputs could be very sensitive to errors in the input variances and covariances (which themselves are not accurately known and are necessarily estimated). It is noted that this negates the main perceived advantage of the full fusion method in its original analytical form but due to the non-invertible nature of covariance matrices where the underlying distribution is multinomial, it was concluded that this was the only realistic implementation of full data fusion.

6.2. Practical Issues

6.2.1. Sample Sizes

The fusion process relies on the combination of sample data, the reliability of which increases as the sample size increases. For this reason, fusion is best applied over aggregate sectors or in this study, 'fusion blocks', the derivation of which is explained in Section 3.2.2.

6.2.2. Vehicle/User Classes

For the purposes of this study, investigating the efficacy of data fusion techniques, the analyses have been confined to light vehicles. Considering a single user class simplifies both the data processing and the comparison of results. While the TM data distinguishes between cars, LGVs and HGVs, the matrices in the CSRM are split only into light and heavy vehicles and the INRIX mobile phone data is not disaggregated at all.

[36] the additional constraint that all proportions must sum to one in effect means that one of the rows in the matrix can be written in terms of the other rows, i.e. is linearly dependent on them. Of course, if at some future date it was considered that the "new" data was sufficiently reliable to give an estimate of the level, as well as the pattern, of daily travel, the constraint would no longer be required and the matrix would, in principle, be invertible. However, given the practical result just noted, it may be questioned whether the additional computation required to deal with the off-diagonal terms would be justified.

[37] The condition number of a matrix is the norm of the matrix times the norm of its inverse and the smaller the number the more accurate the associated matrix inverse. A singular matrix is associated with an infinite condition number. A small perturbation of a singular matrix is non-singular, but the condition number is likely to be large, and hence the associated matrix inverse is likely to be inaccurate, since a small change to a matrix might not change its norm much, but it is likely to change the norm of its inverse a great deal.

WebTAG Guidance[38] is clear that the expectation is that cars, LGV and HGVs should be modelled as separate vehicle types. The fact that the CSRM was developed without a distinction between cars and LGVs has constrained this study to make comparisons on that basis – this is not a recommendation for others.

In the preceding chapters, the means of deriving the data required for fusion of light vehicle data has been set out in turn for each of the input data-sets.

6.2.3. Spatial Coverage

The spatial coverage of the CSRM is shown in Figure 3-1. While the model focuses on trips within the modelled area, all trips passing through the study area are theoretically in scope. Trips between external trip ends that cross the periphery of the modelled area on minor roads are not necessarily reflected but those through trips making strategic journeys across the study are included. In comparison the TM data samples *all* trips wholly or in part travelling within the study area.

The INRIX mobile phone data also samples all trips travelling anywhere within the study area, but crucially (and unlike the CSRM and TM data) there is no knowledge as to the ultimate location of external trip ends. Consequently, a distribution of external trip ends has been developed, as documented in Section 5.1. This is an undoubted shortcoming of the data with unknown impact on the non-sample variances (though an allowance has been made as discussed in Section 5.2). In future studies it should be possible to avoid this shortcoming as data providers are now readily able to provide data coded to external trip end locations.

6.2.4. Post-Processing

In all cases the matrix fusion has been undertaken on the basis of sample proportions with the output from the fusion being an array of proportions – i.e. a distribution of trips requiring expansion. Various methods for expanding the data were tested, as set out in Section 7.1.

As the fusion was undertaken for aggregated 'fusion blocks', even when expanded the output results require disaggregating into individual ij trip matrix cells. The alternative approaches used during the test programme for this study are documented in Section 7.2.

Finally, the efficacy of the fusion was gauged in terms of improvements in the assignment validation of the revised prior matrix, and in terms of the 'work done' by subsequent matrix estimation. The metrics used and the datum for comparison are discussed in subsequent sections.

6.3. Summary of Approach

The following points summarise the steps taken to apply matrix fusion:

1. Assemble the existing "prior" light vehicle AM peak proportion matrices **p** into the defined 'fusion blocks' and derive their associated variance matrices **Ω,** as discussed in Section 3.
2. Separately for TM and INRIX data, assemble proportion matrices for the fusion blocks, and associated variance matrices, as described in the respective Sections above.
3. Carry out a range of tests for bias in the TM and INRIX data, introducing corrections **b** as necessary. Following this, derive updated proportion matrices **p′** and associated corrected variance matrices **Ω′**, again as set out in Sections 4 and 5.
4. Set up procedures to fuse the data according to:

$$p^{updated} = \left(\Omega^{-1} + \Omega'^{-1}\right)^{-1}(\Omega^{-1}p + \Omega'^{-1}p')$$

[38] WebTAG Unit M3-1 Highway Assignment Modelling.

5. Expand and distribute the output proportions to yield an updated trip matrix estimate, covering all ij cells required.
6. Assign the updated estimate (together with the original unadjusted heavy vehicle prior) and compare validation statistics with those achieved during the original CSRM development.
7. Undertake ME for light vehicles only and compare the ME statistics with those obtained for light vehicles during the original CSRM development.
8. Assign the resulting post-ME light vehicles (together with the original post-ME heavy vehicle matrix) and again compare validation statistics with the equivalent statistics from the original CSRM development.

6.4. Implementation

All but step 4 in the above list were undertaken using standard transport planning tools in the form of SATURN modelling software and Excel spreadsheets. Excel was used to:

- aggregate the matrices into fusion blocks;
- calculate variances for the fusion proportions;
- export the resulting proportions and variances for subsequent input to MATLAB;
- undertake the expansion of output proportions (after fusion within MATLAB);
- distribute trips within each fusion block;
- export to SATURN for subsequent assignment.

SATURN was used for the:

- assignment of the revised matrices to the modelled highway network;
- ME process required for each matrix fusion test;
- extraction of validation and ME statistics for the purposes of comparison between tests.

The actual fusion of proportions was undertaken within MATLAB.

MATLAB is a fully-functional language that lends itself easily for use in model development, calibration and validation, statistical and numerical analysis, as well as for building simulations and deriving analytical solutions to optimisation problems. Inputs to MATLAB comprised comma separated variable (CSV) files, with the array format dependent on the particular input. The input files were:

- Array of Prior matrix proportions (184 proportions)
- Covariance matrix for Prior matrix proportions (184x184 covariances)
- Array of proportions from the data to be fused (184 proportions)
- Covariance matrix for the fusion data proportions (184x184 covariances)

Outputs comprised:

- Array of updated matrix proportions (184 proportions)
- Trace of Covariance matrix for the updated proportions (1 value)

While the simple inverse variance weighting method for matrix fusion may be applied using MATLAB, there is no need for this as the weighting may be applied directly within a spreadsheet or some standard matrix

manipulation software, for any practical problem size. Consequently, the MATLAB matrix fusion script in Appendix B is confined to the version used for the full covariance-based fusion.

7. Matrix Fusion Test Programme

To test the fusion process, the idea in principle is to carry out separate fusions of the prior matrix with the new data; and by taking the fused matrices through the ME procedure, to test for resulting improvement in terms of both count validation and the extent of changes brought about by the ME. A programme of tests was conceived at the study outset to investigate the relative strengths of TM and INRIX data when combined with the existing CSRM prior matrix, using a formal fusion process.

It soon became clear that the choice of whether to attempt a full covariance matrix approach in the fusion process, or only to use the diagonal terms (variance) is important. If the difference between the two options is insignificant, it is clearly far better to use the simpler and well used inverse variance weighting approach. Other key issues which emerged as requiring further investigation through the test programme are listed and then explained further, below:

- the choice of expansion approach;
- the use of data from the new source to distribute ‘fused’ trips within the amalgamated fusion blocks;
- Alternative approaches to the treatment of unobserved areas of the prior matrix (apply fusion, retain the original Prior, use the new data source or take their mean);
- Alternative assumptions for the variances of unobserved areas of the prior matrix, where matrix fusion is the test method for those areas (i.e. high and low variance tests alongside a central case).

Inevitably, with so many dimensions to the tests in an already technically involved context, the test programme was necessarily complex, requiring a large number of separate tests.

7.1. Expansion

It is recognised that the data is being fused on proportions and that after fusing, the matrices will need to be expanded. During the test programme it became clear that the quality of the results, in terms of assignment validation of the updated prior matrix, was dependent upon the method of expansion used. Five alternative methods were tested; evolving from the initial assumption that the fused output proportions would best be preserved by employing a global expansion. However, after testing the alternatives with different datasets, the fifth method was seen to yield the best results and was used subsequently for the remainder of the test programme:

1. Global expansion to CSRM 2006AM light vehicle prior matrix total;
2. Global expansion initially as per (1), but adjusted globally by the ratio of observed screenline flows to assigned screenline flows[39]
3. Differential expansion constrained to CSRM 2006AM light vehicle prior matrix total, for each 5x5 super-sector;
4. Differential expansion by super-sector as (3), but with movements between two external trip ends constrained to the original prior matrix values
5. Differential expansion (with different factors by super-sector), with the factors calculated as the ratio of observed screenline flow to assigned screenline flow for each screenline.

[39] The Cambridge Outer Cordon, County N-S Screenline and Huntingdon Radials were used in the calculation of this adjustment factor

It is noteworthy that the universal coverage of data collected from GPS and mobile phones means that they may contain proportionately more trips peripheral to the study area than would be seen in a conventionally derived prior matrix. Conventional matrix development is founded on RSIs at screenlines or cordons focused on the core study area, with calibration counts located on key urban radials and interurban routes; this results in relatively little modelled demand between areas which are peripheral to the study but are still strictly within the study area. In contrast, the 'big data' sources provide data consistently across all areas and it was a recognition of this that led to testing of option #4 in the above list; which aimed to model peripheral trips for greater consistency with the original prior matrix

Option #5 was developed by considering flows across the screenlines which divide the 5 super-sectors. A total of six adjustment factors were calculated by comparing modelled and observed flows at each of the screenlines listed in Table 7-1. In each case a factor was calculated as the ratio of observed to modelled screenline flow in the direction concerned. The factors were then applied to the appropriate super-sectors of the matrix, as shown in Table 7-2.

Table 7-1 Source of 5x5 Super-Sector Screenline Expansion Factors

Expansion Factor ID	Source Screenline	Source Direction
1	Cambridge City Cordon	Inbound
2	Cambridge City Cordon	Outbound
3	Huntingdon Cordon	Inbound
4	Huntingdon Cordon	Outbound
5	Huntingdon Town Centre	Trip weighted average of both directions
6	County N-S Screenline	Trip weighted average of both directions

Table 7-2 Application of Expansion Factors to Updated Prior

Expansion Factor ID	Super-Sector Destination				
Super-Sector Origin	**A**	**B**	**H**	**E**	**X**
A	1	1	2	2	2
B	1	1	2	2	2
H	1	1	5	4	4
E	1	1	3	6	6
X	1	1	3	6	6

7.2. Distribution

The fusion process outputs 184 proportions and irrespective of how these are expanded, the resulting 184 trip volumes need further disaggregation to yield the full 325x325 demand matrix. The default assumption is

to take the requisite distribution from the original CSRM prior matrix, but clearly the distribution from the new dataset, whether TM or INRIX data, could be used instead. All three methods of distribution were tested.

7.3. 'Unobserved' Areas of the Matrix

As explained in Section 3.3.2 the unobserved parts of the prior matrix (i.e. those not founded on RSI surveys but relying on synthetic data from other models) present special difficulties in the context of data fusion, as the variances for these values are unknown. As previously explained, the default assumption is for an 'RSI Equivalent' variance to be calculated as if the relevant values were based on observed data, with these equivalent variances then doubled. Fusion tests were undertaken with either:

- 'high' variances (namely 4 times the 'RSI Equivalent' variance), or;
- 'low' variances (1/2 the 'RSI Equivalent' variance assumed for all synthetic-sourced values, except those from the EERM which remain at twice the 'RSI Equivalent' variance).

7.4. Matrix Estimation

It is important to note that the efficacy of the fusion was gauged in terms of improvements in the assignment validation of the revised prior matrix, and in terms of the 'work done' by subsequent matrix estimation – the precise metrics used are defined in the following section. As introduced in section 3.1.1, for the purposes of the matrix fusion tests undertaken, the assignment comparisons were made using the version of the prior, before any application of matrix estimation. As the CSRM used an initial pre-ME, the 'A1 correction', to correct for known errors, before a subsequent full ME, the full changes which the results documented herein attribute to matrix estimation, appear unusually large.

It is also important to avoid confusion with the link fusion methodology; as Link Fusion is an alternative to ME itself, some of the link fusion comparisons were made with the version of the prior immediately before the full ME, i.e. after the 'A1 correction'. Thus the datum for making comparisons differs between the link and matrix fusion tests.

7.5. Summary of Tests

Given the number of different options implicit in the foregoing sections, a large number of tests needed to be carried out. Following some initial tests (1-17, 20 and 21) to determine the mathematical practicalities of the fusion process, the tests as listed in Table 7-3 were undertaken.

The tests are identified by a 'Test ID' taken from the final successful attempt at each test and are listed in chronological order. In all cases fusion was undertaken using both simple inverse variance weighting and full covariance based approaches. However, given the initial results discussed below, only the simple inverse variance weighting results were then fully processed through SATURN assignment and ME processes.

The tests vary according to

a) the data set with which the prior is to be fused (i.e. TM, INRIX or both)

b) the assumptions for calculating the variances for both the EERM-based prior cells and the 'unobserved' prior cells (see below)

c) the derivation of the output values *between* fusion blocks ('inter-fusion block' movements), for the unobserved parts of the prior matrix (i.e. using matrix fusion or some combination of the input datasets)

d) the source of the distribution of matrix cell values *within* a fusion block (the 'intra-fusion block' distribution) (i.e. Prior, TM or INRIX)

e) the means of expansion to the observed flows (i.e. the five alternatives listed in section 7.1).

Regarding the 'High' and 'Low' variance tests for the unobserved part of the prior matrix, these tests comprised Tests 35 and 36 respectively in the following table and were complemented by a variety of alternatives to the treatment of unobserved areas of the matrix. For each of the following tests the fusion block proportions for these unobserved areas were taken from:

- The new dataset (Tests 39 and 45);
- The existing prior matrix (Tests 40 and 42);
- The mean of the new data and the existing prior matrix (Tests 41 and 43).

Test 45 is a special test in that not only are the unobserved parts of the matrix constrained to the new dataset (in this case, the INRIX mobile phone data), but *all* fusion blocks across the entire matrix are similarly constrained to the INRIX data. Thus no fusion takes place and instead the existing prior matrix is simply constrained to the inter-fusion block distribution given by the INRIX data. This much simpler use of the new dataset is what practitioners would probably attempt to do in the majority of cases (where there is an absence of any knowledge of variances and no intention to combine data on a statistical basis). It is akin to current best practice in the use of observed RSI data – i.e. that the observed data should be used at a sector level to constrain more spatially detailed synthetic data.

Table 7-3 Programme of Matrix Fusion Tests

ID		Prior Matrix Variance Assumptions for Fusion Blocks[40]		Processing Assumptions		
Test ID	Data to be fused with CSRM Prior	Variance of EERM-sourced prior data	Variance of other unobserved prior data	Source of constraints for inter-fusion block unobserved movements	Source of Intra-Fusion Block Distribution for use in output matrix	Expansion method for 184 Fusion Block proportions from MATLAB[41]
18	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	3
19	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	4
22	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	3
23	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	4
24	TM & INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	3
25	TM & INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	4
26	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	1
27	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	2
28	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	5
29	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	1
30	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	5
31	TM & INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	TM	1
33	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	INRIX	1
34	INRIX	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	INRIX	5
35	TM	4*'RSI-Eq.'	4*'RSI-Eq.'	Matrix Fusion	TM	5
36	TM	2*'RSI-Eq.'	½*'RSI-Eq.'	Matrix Fusion	TM	5
39	TM	N/A	N/A	TM	TM	5
40	TM	N/A	N/A	Prior	TM	5
41	TM	N/A	N/A	(TM+Prior)/2	TM	5
42	INRIX	N/A	N/A	Prior	INRIX	5
43	INRIX	N/A	N/A	(INRIX+Prior)/2	INRIX	5
44	TM	2*'RSI-Eq.'	2*'RSI-Eq.'	Matrix Fusion	Prior	5
45	No Fusion	N/A	N/A	INRIX	Prior	5

[40] Where 'RSI-Eq.' refers to the 'RSI-Equivalent' estimated variance introduced in Section 3.3.2.
[41] Numbers refer to the numbered alternatives for expansion, as listed in Section 7.1, above.

8. Matrix Fusion Results

8.1. Metrics for Assessment

Several metrics for improvement of the prior matrix were used during the research which can broadly be categorised as follows:

- addressing the statistical uncertainty of the output fused proportions;
- addressing the quality of assignment validation of the updated prior, for instance against observed screenline counts;
- addressing the scale of the changes imposed by the subsequent ME process to satisfy a comprehensive set of observed counts; and
- addressing the quality of assignment validation of the post-ME matrix, in terms of counts and journey times.

The first of the above metrics would ideally compare the aggregate variance of the output fused proportions from each test with those from the input prior matrix. It was not readily practicable to devise a formal test of whether the fusion technique had reduced the output variance to any particular level of statistical confidence. The use of the 'F' Test was considered, which can be used to assess the statistical significance of differences in two variances. However, the problem here cannot be posed appropriately.. However, the trace[42] of the output covariance matrix was used as a proxy for this.

The second set of metrics relate to the output fused assigned flows and comprise the percentage of the:

- 13 available count-sets within 7.5% of the observed count
- 11 calibration screenlines within 5% of count and GEH<4
- 19 individual validation counts within DMRB count validation criteria.
- The average absolute difference between assigned and observed flow over the 13 count-sets.
- Finally, the % root mean square error (%RMSE) = SQRT[sum(observed-modelled counts)^2]/(mean observed count), on the 11 calibration screenlines.

The third set of metrics were based on Table 5 of WebTAG Unit M3.1 and comprise:

- R^2 coefficient for linear regression of Prior and Post ME matrix cell values
- R^2 coefficient for linear regression of Prior and Post ME matrix trip end values
- Change in the mean and standard deviation of the trip length distribution
- Percentage of fusion blocks changed in ME by <5%.

The tests were assessed primarily using the first two of these metrics, with work required by subsequent ME, the third metric, considered as a supplementary measure. The fourth metric, the outputs from ME, was not considered a good basis on which to measure the success of the fusion process (and results were in any case so similar in terms of flow and journey time validation that they failed to help in the comparisons).

An important additional criterion for the success of the fusion process is that the recommended technique needs to be easily disseminated and applied by practitioners.

[42] Sum of the elements of the leading diagonal of the matrix concerned. This was proposed as a suitable measure of uncertainty in the *Skrobanski et al.* paper.

8.2. Simple inverse variance weighting versus full covariance fusion

Initial tests using both the full covariance matrices and the simple inverse variance weighting approach were undertaken. These showed that each perturbation undertaken using the full covariance approach tended to yield a more reliable fused output (i.e. the trace of the output covariance matrix was lower than that achieved using the simpler approach). However, this was not so in all cases. Additionally, it was found that without close attention to the condition numbers of the covariance matrices there is no guarantee of a stable output (as opposed to one unreasonably sensitive to errors in the input assumptions).

In all the tests shown in Table 7-3 both the simple variance weighting method and the full covariance approach yielded a significant reduction in the trace of the output covariance matrix:

- Using INRIX mobile phone data – c. 85% reduction
- Using TM GPS data – c.91% reduction
- Using both datasets – c.94% reduction

These tests all showed a fractionally greater reduction in uncertainty using the covariance approach, of between zero and 0.1% depending on which perturbation of the covariance approach was considered. However for many pertubations the difference from the result using the simple inverse variance weighting was negligible.

To investigate this further, a Hotelling T^2 test was used to check if the output proportions from the two methods were statistically different at the 5% confidence level. Generally this was the case, but not in every case.

Given the very marginal improvements provided by the full covariance approach (and in some cases no statistically significant difference), it was concluded that the considerable complexity of the full covariance approach could not be justified compared to the inverse variance weighting approach. The practical problems of implementing the full covariance approach and interpreting its outputs are significant and it is unreasonable to expect practitioners to be able to use it. Subsequent analyses were therefore confined to the simpler approach (inverse variance weighting[43]).

8.3. Alternative variance assumptions and output processing methods

The results from the inverse variance weighting fusion tests are presented in Table 8-1.

The table lists all the aforementioned metrics together, i.e. updated prior matrix assignment calibration/validation statistics and then ME statistics[44]. The statistics are all expressed such that a greater ratio or percentage represents a better result. Colour coding has been added to highlight when the statistics are better (green) or worse (red) than the equivalent statistic derived from the original CSRM prior matrix. Additionally, the greatest improvements have been underlined, though in some cases these results are not significantly better than several others.

Tests 39-43 did not implement data fusion over the full matrix; instead, for unobserved parts of the prior matrix, simpler assumptions were used to determine the proportions for the relevant fusion blocks, as explained below. Similarly, Test 45 did not use data fusion at all, instead relying on the INRIX data to constrain the inter-fusion block proportions.

[43] In other words, no off-diagonal elements are assumed in the covariance matrix.
[44] Note that most of the ME statistics fall well short of those recommended in WebTAG but this is because the statistics reflect the impact of a two-step calibration process, including the initial ‘A1 Correction’ ME introduced to correct for an apparent localised error during the original matrix development phase.

All the best prior assignment validation statistics were obtained using expansion method (5) – namely the application of a set of expansion factors varying between different areas of the matrix, calibrated to meet observed screenline flows. This is not a surprising finding but is useful to inform the expansion of prior matrices in other studies, particularly if it then results in any subsequent ME having a reduced impact.

Table 8-1 Results of Matrix Fusion Tests

ID			Processing Assumptions			Quality of Fused Output (new prior)					SATME2 Results				
Run ID	Data to be fused with CSRM Prior	Prior Matrix Variance Tests for Unobserved Areas	Source of constraints for inter-fusion block unobserved movements	Source of Intra-Fusion Block Distribution for use in output matrix	Expansion method for 184 Fusion Block proportions from MATLAB	% of all Output Fused assignment count-sets within 7.5% of count	% of all Output Fused assignment calibration screenlines within 5% of count and GEH<4	% of Output Fused assigned flows within DMRB flow criteria of individual validation counts	1 - Average absolute difference between observed and modelled flow across 13countsets)	1 - %RMSE for calibration screenlines	ME2 cellwise regression R^2	ME2 tripend regression R^2	ME2 absolute change in mean trip length	ME2 absolute change in trip length StdDev	% of 'fusion blocks' changed in ME2 by <5%
No. of comparisons						*13*	*11*	*19*	*13*	*11*	*325*[2]	*650*	*1*	*1*	*184*
CSRM						**15%**	**9%**	**68%**	**81%**	**80%**	**0.90**	**0.90**	**88%**	**88%**	**24%**
18	TM	-	MF	Prior	3	31%	18%	58%	82%	81%	0.89	0.92	90%	91%	24%
19	TM	-	MF	Prior	4	31%	18%	53%	82%	80%	0.92	0.92	90%	89%	20%
22	INRIX	-	MF	Prior	3	8%	9%	53%	83%	83%	0.85	0.87	83%	82%	23%
23	INRIX	-	MF	Prior	4	8%	9%	68%	84%	83%	0.91	0.90	89%	88%	23%
24	TM & INRIX	-	MF	Prior	3	31%	27%	47%	84%	82%	0.87	0.90	86%	86%	23%
25	TM & INRIX	-	MF	Prior	4	31%	9%	53%	83%	81%	0.92	0.91	89%	88%	22%
26	TM	-	MF	TM	1	15%	9%	47%	68%	54%	0.90	0.89	93%	95%	17%
27	TM	-	MF	TM	2	15%	18%	21%	74%	69%	0.86	0.88	80%	84%	19%
28	TM	-	MF	TM	5	54%	36%	26%	88%	86%	0.88	0.88	94%	95%	19%
29	INRIX	-	MF	TM	1	46%	36%	53%	88%	86%	0.88	0.86	99%	98%	20%
30	INRIX	-	MF	TM	5	46%	36%	53%	90%	90%	0.87	0.83	98%	99%	22%
31	TM & INRIX	-	MF	TM	1	15%	9%	37%	73%	64%	0.90	0.88	95%	94%	21%
33	INRIX	-	MF	INRIX	1	8%	9%	37%	82%	83%	0.80	0.85	86%	85%	15%
34	INRIX	-	MF	INRIX	5	31%	45%	37%	88%	87%	0.80	0.84	86%	85%	17%
35	TM	High	MF	TM	5	46%	27%	26%	87%	85%	0.88	0.88	94%	95%	20%
36	TM	Low	MF	TM	5	54%	36%	26%	88%	86%	0.88	0.88	94%	95%	19%
39	TM	No Fusion	TM	TM	5	46%	18%	16%	84%	81%	0.87	0.89	90%	92%	21%
40	TM		Prior	TM	5	54%	27%	26%	88%	87%	0.86	0.82	94%	91%	24%
41	TM		TM Mean	TM	5	54%	36%	21%	87%	85%	0.87	0.86	92%	92%	22%
42	INRIX		Prior	INRIX	5	38%	45%	42%	87%	87%	0.83	0.86	88%	88%	17%
43	INRIX		INRIX Mean	INRIX	5	31%	45%	47%	87%	87%	0.85	0.86	90%	88%	21%
44	TM	-	MF	Prior	5	31%	18%	47%	85%	85%	0.86	0.91	87%	87%	26%
45	No Fusion		INRIX[45]	Prior	5	62%	36%	47%	86%	81%	0.87	0.81	97%	93%	27%

Note: Counts were arranged into 11 screenlines of calibration counts, plus sets of 'Cambridge' and 'non-Cambridge' validation counts, making a total of 13 count-sets.

[45] In this particular test, the INRIX data was used to constrain inter-fusion block movements over the entire matrix and not only its unobserved areas.

Arguably the best results come from use of INRIX data, but using a TM distribution to distribute trips within each fusion-block (Test ID 30). Such a scenario is unlikely to be implemented in practice as the modeller is unlikely to go to the time and expense of obtaining and processing two new datasets for use in this way. In general, the spatially less detailed INRIX data was not found to be a good source for the distribution within the fusion blocks; comparing Test 34 with Test 30 almost all metrics are superior using a TM distribution, in an otherwise identical test. However, where a TM distribution is used in both cases (IDs 28 and 30) more of the metrics are better for the INRIX data than the TM data, in terms of both prior assignment validation and ME impacts, consistent with INRIX being a richer, less biased data source.

One of the foci of the tests was how to treat the fusion of those parts of the matrix not observed in the original prior, for which variances are unknown and have to be estimated. The later tests (Tests 39-43) investigate this and demonstrate that with either new data source there is no advantage in undertaking a statistical fusion of the unobserved fusion-blocks, compared to taking the output proportions directly from the original prior or (particularly in the case of the INRIX data) from the mean of the prior and new data source.

Where fusion has been undertaken for these unobserved fusion blocks, varying the assumed prior matrix variances (Tests 35 and 36 versus 28) made little difference to the results. What difference there was suggests that the testbed model Prior matrix is a marginally better starting point than the TM data, such that reducing the Prior variance (and therefore using a higher proportion of the Prior in the fusion) gives a marginally better result. However, the results are so similar that it is not possible to draw any conclusion from this, and Test 41 (i.e. using the mean of Prior and TM constraints with no formal fusion of the unobserved fusion-blocks) performed virtually as well and is far more defensible.

The final Test 45 is an interesting ‘control’ in that it represents what practitioners are most likely to do in the absence of prescriptive guidance. In this test no fusion is applied but the INRIX data is used to constrain inter-fusion block movements across the entire matrix. The resulting updated Prior matrix assignment performs relatively very well at approximating the sums of the available count-sets to within 7.5%. It also performs well on the some of the measures for the changes introduced by subsequent ME. Crucially, in terms of performance against independent validation counts, it is beaten by several of the other tests, including the aforementioned ‘winner’, Test 30, which also surpasses Test 45 when looking at average absolute errors across screenlines and calibration count-sets.

9. Matrix Fusion Conclusions

Based upon the results presented in the preceding chapter, the following comparisons are made.

9.1. Alternative Matrix Fusion Methods

The full matrix-based approach using covariances is extremely difficult to deal with from the outset. In general terms, for a model with S sectors between which movement data is to be fused, the number of elements in the covariance matrix equals S^4. For models with many sectors this could create significant storage problems before even attempting to invert the matrix, as required by the matrix fusion equation.

Even when the problem is small enough to be manageable, the use of proportions in the fusion, which seems sensible in all other regards, presents a further technical problem. As the proportions sum to one, the terms ae not entirely independent and the rank of the covariance matrix is one less than its dimension; by definition it is un-invertible. The derivation of the Moore-Penrose pseudo-inverse, computed using singular value decomposition, has provided a work-around, but there is no guarantee of a stable solution even then, with the need to monitor the condition number of the output to avoid alighting on a solution which is unreasonably sensitive to small changes in the inputs.

The practical problems alone suggest that use of the full covariance matrix (as is implied by the matrix fusion equation shown below) is all but impossible for realistic sized problems and project teams with conventional transport modelling skills.

$$t' = \left(\Omega_{t_1}^{-1} + \Omega_{t_2}^{-1}\right)^{-1}\left(\Omega_{t_1}^{-1} t_1 + \Omega_{t_2}^{-1} t_2\right)$$

The fact that the use of covariances appears to improve the results only very marginally means that a clear recommendation is that they should in practice be ignored. Under this assumption, the matrix fusion method reduces to a straightforward inverse variance weighting.

9.2. Alternative Data Sources

9.2.1. Inter-Fusion Block Distribution

9.2.1.1. TrafficMaster or INRIX

Comparing the benefits of fusing the Prior matrix with either the TM GPS data or the INRIX mobile phone data, Tests 18/19 and 22/23 respectively, are directly comparable but the results are not particularly informative. The INRIX data performs best (Test 23) in terms of the validation of the updated Prior against independent counts and the scale of errors against the calibration screenlines; while the Prior updated with TM data shows marginally less change required during the subsequent ME process.

The improvement offered by the INRIX data is more significant in Test 29 compared to Test 26. Similarly, in Test 30 compared to Test 28 the INRIX data is on the whole seen to give a better result. (Test 28 may have more assigned model counts within 7.5% of the observed count-sets, but the average absolute error is in fact greater.) Consequently, using this particular testbed model and data-sets, it appears that the INRIX data is the better dataset to input to the matrix fusion.

9.2.1.2. TrafficMaster *and* INRIX

Comparing Test 24 with Test 18 (TM only) and Test 22 (INRIX only), the performances against the various comparative metrics are inconclusive. The same is true for Test 25 compared with Test 19 (TM only) and Test 23 (INRIX only). Finally, comparing Test 31 with Test 26 (TM only) and Test 29 (INRIX only), the test

using both data sources to input into the matrix fusion never out-performs both the tests using one new data-source only, except in the very final metric, where it is only marginally better.

On the basis of these results there appears to be little or no benefit in processing and using both new datasets to fuse with the Prior matrix. Mathematically, using both new sources *will* serve to reduce the uncertainty in the final fused result, but this does not appear to have made a material improvement to the quality of the updated Prior matrix in the case of this testbed model. However, it cannot be concluded that there is never any value in using two new data sources, particularly given that none of the above examples use the preferred method of expansion (method 5).

9.2.2. Intra-Fusion Block Distribution

There are no directly comparable results from which to draw any conclusions over use of the Prior matrix intra-fusion block distribution compared to using INRIX data. However, looking at Tests 28 and 44, the TM distributed data generally performs better in terms of the assigned count calibration metrics, although the Prior distribution of Test 44 is materially better in terms of the assigned count validation metric. Similarly, against most measures of ME changes, the TM distributed test performs better than the Prior distributed test, except for the final metric, on which Test 44 is one of the best performing. These differences are inconclusive.

Tests 30 and 34 facilitate comparison of the TM distributed data with equivalent INRIX distributed data. It is clear that on virtually all measures the TM distributed test performs better. This is in line with expectation given the spatial detail of the GPS-based TM data compared to the spatially less precise mobile phone data from INRIX.

9.3. Alternative Processing Assumptions

9.3.1. Method of Expansion

In the above results it is not possible to make a direct comparison between all five of the tested expansion methods. The closest such comparison is between Tests 18 and 19 and Tests 26 to 28. Of these, Test 28 (using the expansion method 5) is better than all the others on all bar one of the Prior matrix count comparison metrics. Consequently, the remaining tests all assumed this approach which may be summarised by:

> Differential expansion (with different factors by super-sector), with the factors calculated as the ratio of observed screenline flow to assigned screenline flow for each screenline.

9.3.2. Treatment of unobserved fusion-blocks

Test 39 to 43 test some alternatives to the fusing of data for unobserved area of the matrix where the variances are not known and have to be estimated. Tests 39 to 41 are directly comparable to Test 28; all use TM data to fuse with the Prior in the observed parts of the matrix and use the same expansion method and means of intra-fusion-block distribution, but in terms of the treatment of the inter-fusion block constraint for the unobserved parts of the matrix:

> Test 28 – matrix fusion as per the rest of the matrix
>
> Test 39 – constrain directly to TM data (no fusion)
>
> Test 40 – constrain directly to the Prior (no fusion and no material change from the original Prior)
>
> Test 41 – constrain to the mean of the TM data and the Prior (no fusion).

Except for the final metric measuring changes due to subsequent ME, where Test 40 in particular shows a small improvement on the matrix fusion approach of Test 28, on all other metrics, none of the alternative

tests are material improvements on Test 28. However, with no robust method to calculate the variances for these fusion-blocks there is some risk that this result would not be seen in another example with different estimated variances. It would be better for these statistically 'unknown' areas of the matrix if the adopted method did not rely on variance estimations and thus did not use fusion in the statistical sense. On this basis, the best of the alternative tests appears to be either of Tests 40 and 41 – both are superior to Test 39 in terms of the primary criteria of fit between observed flows and assigned flows from the updated Prior. It is difficult to choose between these two and it is likely, again, that the 'best' is context specific. Either way, the results suggest that constraining the inter-fusion-block distribution to the Prior, or to the mean of TM data and the Prior, would be the best alternative to matrix fusion for the unobserved areas of the matrix.

Similar comparisons can be made for the INRIX sourced tests:

Test 34 – matrix fusion as per the rest of the matrix

Test 42 – constrain directly to the Prior (no fusion and no material change from the original Prior)

Test 43 – constrain to the mean of the INRIX data and the Prior (no fusion).

Unlike the above TM-based comparisons, these tests show that on every metric the alternative Tests 42 and 43 are equal to, or better than, the INRIX-based fusion for the inter-fusion block distribution of the unobserved areas of the matrix. It would be interesting to see how these results might differ had the tests used the TM data for the intra-fusion-block distribution, which has already been demonstrated to be better for this purpose. Of the two alternatives, Test 43 is probably the better performing with this testbed model.

Taking all the above results together it would be prudent to base the inter-fusion-block distribution for the unobserved areas of the matrix on the mean of the original Prior and whichever 'new' dataset is being used. The reasons are:

- Statistical fusion requires knowledge of variances which we do not know reliably for these areas of the matrix
- In this particular set of results the mean appears to be the best of the alternatives
- Using the mean reduces the sensitivity of the results to errors in any one dataset (and is equivalent to statistical fusion where the variances are equal).

9.4. Summary

The conclusions drawn from the preparatory research and the above comparisons are:

- The values of the different data sources to be fused are best expressed as proportions which sum to one, to facilitate subsequent expansion after the fusion has been completed.
- The fusion should be undertaken for aggregated areas of the matrix, defined such that the sample of prior matrix data available for each area is statistically significant.
- Use of the new data sources substantially reduces the statistical uncertainty of existing prior matrix data and can improve the resulting prior matrix validation and reduce the impact of subsequent ME.
- The full covariance approach to matrix fusion offers very little benefit over more conventional inverse variance weighting and cannot be justified given its considerably greater complexity and need for specialist interpretation.
- Particular attention should be paid to the expansion method, making best use of calibration data to inform the expansion, so as to reduce the subsequent impacts of any ME processes. In this instance an approach based on sector-based expansion to screenline totals (method 5) works best.

- In this case, the INRIX mobile phone data was marginally the better choice for use in matrix fusion for the inter-fusion block movements, reflecting its big sample size, network coverage and relative lack of bias.
- The distribution chosen to distribute the intra-fusion block output proportions also has an impact; the TM GPS data was found to be the best source for this data, reflecting its spatial accuracy.
- For those areas of the prior matrix which are unobserved (no RSI data) the most defensible approach is to assume the new and existing sources are equally reliable and take the output 'fused' proportions as the mean of the input proportions.

10. Link Fusion Methodology

10.1. Link Fusion Theory

10.1.1. Context

What has been termed here as “Link Fusion” is the extension of the “Matrix Fusion” method presented in the first half of this report, to include the fusion of link count data with origin-destination matrix data. Count data is readily collected by manual or automatic means and is well understood in terms of its statistical confidence. It should therefore be an excellent source for improving the initial matrix. However it is not in a form which is readily combined with matrix data.

As the counts are already correctly scaled, the Link Fusion can proceed using the full initial Prior matrix, *D*, rather than an equivalent matrix of unscaled proportions, as was the case with Matrix Fusion. A further difference from the Matrix Fusion, as first introduced in section 3.1.1, is that the Link Fusion tests and comparisons should be made using the version of the prior *after* the ‘A1 correction’, to facilitate direct comparisons between Link Fusion and the ME process used in the development of the original CSRM.

Use of counts to improve a Prior matrix is in practice currently implemented through some form of Matrix Estimation. A popular approach in the UK is the use of the SATME2 module of the SATURN traffic modelling suite. For further context, see Appendix C.

It should be noted that in SATME2:

- no consideration is taken of the reliability of either the prior matrix or the observed link counts
- the aim is to get as close a fit as possible (subject to convergence) to the counts
- the matrix is adjusted at the ij (cell) level, which has implications for the comparison with the Link Fusion method.

10.1.2. “Link Fusion”

The method developed by Skrobanski et al aims to combine the estimates of matrices and counts according to statistical principles. Unless the counts are considerably more accurate than the matrices, it will not be expected to obtain such a good fit to the counts as is achieved by SATME2. However, it should avoid the “mechanistic” problems with ME which can lead to considerable distortions of the prior matrix (and which the checks in WebTAG Unit M3.1 are designed to control). It follows from this that while both methods aim at updating the trip matrix on the basis of counts, the Guidance relating to the two methods can be expected to be very different.

Using the same notation as above, the Skrobanski method[46] can be described as follows:

Both the prior matrix D_{ij} and the observed counts V_a^{obs} have associated covariance matrices Ω^D and Ω^V. In theory these could be fully populated with covariances but in practice they are assumed to be diagonal, reflecting the findings of the Matrix Fusion research and facilitating the requisite matrix inversion that otherwise would be computationally problematic. The fusion is achieved by means of minimising the (Least squares) criterion:

$$\left(D' - D\right)^T .\left(\Omega^D\right)^{-1} .\left(D' - D\right) + \left(p.D' - V^{obs}\right)^T .\left(\Omega^V\right)^{-1} .\left(p.D' - V^{obs}\right)$$

where the matrices D_{ij} are interpreted as one-dimension vectors, and *p* is a (PIJA) matrix of the form a * ij.

[46] NB A number of different methods are described in the paper, but we are here confining to the “analytical” WLS method

The solution is shown to be:

$$D' = \left[\left(\Omega^D\right)^{-1} + p^T.\left(\Omega^V\right)^{-1}.p\right]^{-1}\left[\left(\Omega^D\right)^{-1}.D + p^T.\left(\Omega^V\right)^{-1}.V^{obs}\right]$$

The two Ω matrices, being diagonal, are readily inverted. The inversion of the first term in square brackets can be simplified by use of the Sherman–Morrison–Woodbury formula whereby:

$$M = \left[\left(\Omega^D\right)^{-1} + p^T.\left(\Omega^V\right)^{-1}.p\right]^{-1} = \Omega^D - \Omega^D.p^T.\left(\Omega^V + p.\Omega^D.p^T\right)^{-1}.p.\Omega^D$$

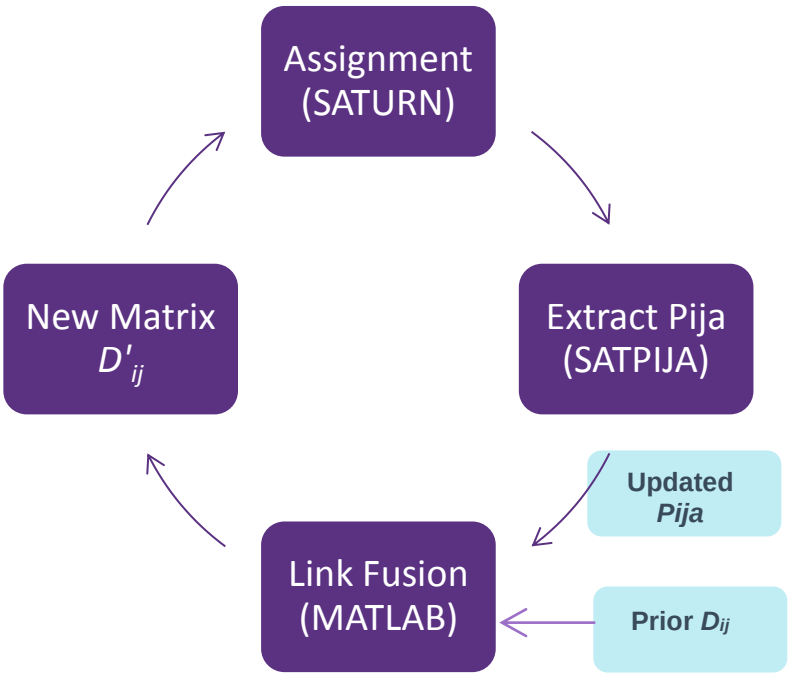
The matrix *M* can also be shown to be the covariance matrix of the revised estimate *D'*.

10.2. Revised *Pija* Proportions

As with SATME2, the updating of the trip matrix can be expected to lead to revised *Pija* proportions *p*. This does not appear to be discussed in the Skrobanski paper, though in principle a comparable iterative scheme could be applied. However in practice, once we go into the next round, there is a new "prior" matrix *D'* from the first round which yields a changed set of *Pija* proportions (if it doesn't the process can stop). The variance matrix for this is supplied by *M* from the first round. But not being diagonal, the inversion of this huge matrix is challenging if not infeasible. For the purposes of this research it was therefore not considered that the original Pija proportions might change as a result of the revised matrix.

However, there is a precedent in the recommended method of application of SATME2, which suggests that iterations of the Link Fusion process may still be possible. This could potentially achieve a 'better' end result due to the *Pija* proportions having greater consistency with the updated Prior, *D'*. In SATME2 the recommendation is to assign the initial updated Prior, D'_1, in order to extract revised *Pija* proportions and input these *Pija* proportions to the second loop of estimation, using the *initial* input *D,* rather than D'_1. Likewise, a third iteration would start with *Pija* proportions from the assignment of D'_2 but the same initial input *D*. This is the process followed for five iterations in the matrix estimation of the test-bed CSRM model, against which the Link Fusion is compared in Section 14. The figure below illustrates the process.

Figure 10-1 Schematic of Iterative Link Fusion Process as Might be Implemented using SATURN



Further research should be undertaken into the potential benefits of an iterative approach to Link Fusion.

Skrobanski et al suggest that uncertainty relating to *p* should be treated as an issue of bias. Some discussion is provided in §2.7 of the paper, but no substantial advice is given. There are two aspects here. If it is assumed that the assignment routeing is "correct", then the only issue is that of "convergence" between the matrix and the *Pija* proportions (as just discussed). More serious is the likelihood that the routeing is in some way inconsistent[47]. While this should certainly be noted, we have no proposals for dealing with this.

10.3. Spatial Detail

In theory the Link Fusion may be undertaken at any level of spatial detail. If applied to IJ sectors rather than ij cells (where $D_{IJ} = \sum_{i \in I} \sum_{j \in J} D_{ij}$) then the covariance matrices would need to be calculated on an IJ basis as well as the prior matrix, *D*. This is compatible with the approach adopted in this study for the Matrix Fusion tests. However, additionally the *Pija* proportions *p* would need aggregating such that:

$$p_{IJa} = \frac{\sum_{i \in I} \sum_{j \in J} p_{ija} D_{ij}}{D_{IJ}}$$

Before proceeding with the tests reported below, it was necessary to agree the level of spatial detail at which the Link Fusion was to be undertaken. Note that in the context of Matrix Fusion the prior cells had to be aggregated to provide volumes of statistical significance, and thereafter the fusion was undertaken on an aggregated basis using 184 'fusion blocks'. The same argument could be seen as valid in the context of Link Fusion. However, as conventional matrix estimation (such as SATME2) is applied on a ij cell basis, then applying Link Fusion on a more aggregate basis than this would impose additional constraints on the process and thereby potentially reduce any benefit that might otherwise be observed in using the Link Fusion method as opposed to conventional matrix estimation.

The choice of level of spatial detail has implications for the Link Fusion data formats (and for the data processing within any iterative form of the Link Fusion and assignment process, should that be relevant). In this context, the practical differences between applying the Link Fusion at a cell or sector level are summarised below. The formats of data already used in Matrix Fusion or readily obtained within the standard matrix estimation processes[48] are highlighted in bold. The SATURN assignment process requires matrices of the form *D'ij* and the *Pija* factors are subsequently generated in the form p_{ija}. These are constraints of the SATURN software.

Table 10-1 Implications of Applying Link Fusion at Aggregated 'IJ' Sector Level

Spatial Level of Application	ij Cell	IJ Sector or 'Fusion Block'
Trips, D	$D_{ij} = p_{ij\|IJ}.\ D_{IJ}$	$\boldsymbol{D_{IJ}} = \sum_{i \in I} \sum_{j \in J} D_{ij}$
PIJA Factors, *p*	$\boldsymbol{p_{ija}}$	$\boldsymbol{p_{IJa}} = \frac{\sum_{i \in I} \sum_{j \in J} p_{ija} D_{ij}}{D_{IJ}}$
Covariances, *Ω* *(see section "Variances" below for more discussion)*	$\Omega_{ij} = (p_{ij\|IJ})^2.\ \Omega_{IJ}$	$\boldsymbol{\Omega_{IJ}}$[49]
Additional processing required before each Link Fusion iteration (where relevant)	-	p_{IJa} factors to be calculated from $\boldsymbol{p_{ija}}$ *factors*
Additional processing required before each assignment iteration (where relevant)	-	D'_{ij} values to be calculated from D'_{IJ} values

[47] as well as possible miscoding, this could arise because of inappropriate generalised cost specification, or because the link cost curves are poorly specified, or even (more fundamentally) because Wardrop's condition is not compatible with behaviour.
[48] In this case referring to p_{ija} factors being generated directly from SATPIJA.
[49] See derivation of Prior matrix variances in section 3.3.

The process is thus simpler to implement on an ij basis, as well as more robust in terms of demonstrating Link Fusion to be a viable alternative to ME. Link Fusion was therefore undertaken on an ij cell basis.

10.4. Grouped Counts

In the original CSRM matrix development, counts were grouped in accordance with Guidance (then WebTAG 3.19) in order to ensure better consistency in the reliability of counts used across the model area. When using matrix estimation (as in SATME2), the aim is to reproduce the observed flows. In this context, it makes sense to aggregate in order to improve reliability, as well as allowing for possible errors in modelled route choice. However, the Link Fusion takes account of the statistical reliability of individual sites that might have been grouped. Consequently, it is our view that there is no advantage in grouping counts for Link Fusion and therefore the counts were not grouped[50].

10.5. Practical Considerations

It is clear that the mathematically complete interpretation of the above Link Fusion equations involve the inversion of very large covariance matrices. In the context of the initial Matrix Fusion work it was concluded that in practice the off-diagonals of the input covariance matrices (in this case Ω^D and Ω^V) could be ignored without much impact on the fused results; that is, simple inverse variance weighting was found to be a very good approximation to the 'complete' result.

In the context of Link Fusion, the same assumption can be made and indeed is essential for a practicable method. However, with reference to the Link Fusion equation below, the first square-bracketed term (which is *M*, the output covariance matrix) will have many significant off-diagonal elements as a result of the *Pija* proportions, *p* and p^T, which 'spread' the significant diagonal terms of Ω^V into off-diagonals areas of *M*.

$$D' = \left[\left(\Omega^D\right)^{-1} + p^T.\left(\Omega^V\right)^{-1}.p\right]^{-1}\left[\left(\Omega^D\right)^{-1}.D + p^T.\left(\Omega^V\right)^{-1}.V^{obs}\right]$$

$$M = [(\Omega^D)^{-1} + p^T.(\Omega^V)^{-1}.p]^{-1}$$

Early tests confirmed that ignoring these off-diagonal elements in *M* gave nonsense results; the output *D'* was far too large as it did not include the negative contributions of the off-diagonal covariance terms of *M*. This means that although we can ignore covariances in Ω^D and Ω^V the resulting method is rather more complex than the 'simple' inverse variance weighting possible with Matrix Fusion. Fortunately the Sherman-Morrison-Woodbury formula allows us to simplify M as:

$$M = \left[\left(\Omega^D\right)^{-1} + p^T.\left(\Omega^V\right)^{-1}.p\right]^{-1} = \Omega^D - \Omega^D.p^T.\left(\Omega^V + p.\Omega^D.p^T\right)^{-1}.p.\Omega^D$$

The term requiring inversion is (Ω^V+$p.\Omega^D.p^T$) i.e. a matrix of dimension 'a by a' for a problem with 'a' counts. This is much more manageable[51]. Note that in subsequent notes on the implementation of Link Fusion in MATLAB, this matrix is termed the 'Quadratic Matrix'.

[50] N.B. If it were decided to group sites, then the p matrix should be redefined by adding across links a ∈ A. Strictly, we should ensure that no double-counting occurs in the routeing (no used path should include more than one link a within a grouped A); the CSRM counts were grouped in a manner to avoid this happening. Otherwise, the effect of grouping is merely on the dimension of the flow vector V^{obs} and the associated *p* matrix. There are no theoretical implications.

[51] Indeed, in this form of *M* it would be possible to account properly for all the covariances in Ω^D and Ω^V. However, it was confirmed through tests that the off-diagonals of Ω^D and Ω^V should not be included in *M* if they are not also included in the second square-bracketed term in the equation for D', and as has already been stated, including them is impracticable.

10.6. Implementation

10.6.1. Introduction

The CSRM test-bed model of the Cambridge sub-region, as introduced at length in the context of Matrix Fusion, was also used for the Link Fusion tests. In common with the Matrix Fusion tests, analyses were focused on the light vehicle type, of the Base year morning peak hour model, with 325 zones. Unlike the Matrix Fusion tests, the Link Fusion work took the Post A1 Correction matrix as its start point, thereby facilitating direct comparison with the main matrix estimation undertaken during the CSRM model development (see discussion in Section 3.1.1).

Each test was assessed for its impact on the quality of calibration and validation of the post-fusion matrix, and how that compared with the performance of the equivalent matrix derived through the ME undertaken during the CSRM model development.

During these tests the heavy vehicle matrix was assumed unchanged – ie the same post ME heavy vehicle matrix was used in all comparisons.

The count set used for Link Fusion was based on that input to the ME, comprising 174 light vehicle counts. That said, some changes were made, as explained below in the section on Count Data.

10.6.2. MATLAB Processing

Inputs to MATLAB comprised comma separated variable (CSV) files, with the array format dependent on the particular input. The input files were:

- Prior matrix array, *D* (105,625 values)
- Diagonal of the covariance matrix for the Prior, Ω^D, as a column array (105,625 variances)
- Observed count array, V^{obs} (174 counts)
- Diagonal of the covariance matrix for the observed counts, Ω^V, (174 variances)
- Assignment *Pija* matrix $\left(p^T\right)_{ij.a}$ from SATURN[52] (105,625 rows by 174 columns)

Outputs comprised:

- Array of updated matrix values, D' (105,625 values)
- Trace of Covariance matrix for the updated values (1 value)

To avoid storing the interim result *M* (which using double floating point precision would require about 83GB of storage space), it is possible to carry out the calculations to update the demand matrix *D'* on an element-wise basis, though this is more complicated and may take longer to run in MATLAB, given that it is optimised to run matrix computations. The theory behind this process is presented below.

10.6.3. Element-wise Calculation of D'

The calculation for *D'* consists of a square matrix (*W*, say) of dimensions $105{,}625^2$ post-multiplied by a column vector (*H*, say) of dimension 105,625. Using z to denote an "i-j" element, the calculation can be written:

$$D'_z = \sum_{z'} W_{zz'}.H_{z'}$$

[52] The SATPIJA function was used to extract these proportions for the Light vehicle user class from the 'Post A1 Correction' assignment which was input into the main run of SATME2 during the CSRM model development.

This makes it clear that the calculation can be done separately for each row of *W*. Thus, provided *W* can be calculated row by row, it is not necessary to store the full matrix.

Consider first the left hand term $W = \left[\Omega^D - \Omega^D.p^T.\left(\Omega^V + p.\Omega^D.p^T\right)^{-1}.p.\Omega^D\right]$.The product $p.\Omega^D.p^T$ will have 174^2 elements. Because the matrix Ω^D is assumed diagonal, we can write the general term:

$\Omega^D_{zz'} = \Omega^D_{zz'}.\delta_{zz'}$ (where $\delta_{zz'}$ is the Kronecker delta)

If we use a,b for the counts, the formula for the (a,b) cell is $\sum_z\sum_{z'} p_{za}.\ p_{z'b}.(\Omega^D_{zz'}.\delta_{zz'}) = \sum_z p_{za}.\ p_{zb}.\Omega^D_{zz'}$. After adding on the elements of Ω^V, the resulting inverse (*Y*, say) will also have 174^2 elements.

The product $\Omega^D.p^T$, which pre-multiplies *Y*, is a 105,625 * 174 matrix with the formula for the (z,a) cell given by:

$$\sum_{z'} p_{z'a}.\ (\Omega^D_{zz'}.\delta_{zz}) = \sum_z p_{za}.\ \Omega^D_{zz'}$$

Correspondingly, as $p.\Omega^D = (\Omega^D.\ p^T)^T$ the (b,z′) cell of the transpose can be written as $p_{z'b}.\ \Omega^D_{z'z'}$

So the (z, z′) term of the product is given as: $\Omega^D_{zz}\ .\ \Omega^D_{z'z'}.\sum_a\sum_b Y_{ab}.p_{za}.\ p_{z'b}$. Each such term is then subtracted from the term $\Omega^D_{zz}.\delta_{zz'}$ to give the matrix $W_{zz'}$.

Hence, while the size of the (left hand bracket) matrix W is still $105{,}625^2$, its calculation is less onerous. Each term can be calculated separately, and in particular it is clear that a full row (z) can be calculated independently of any other row.

Each row now needs to be post-multiplied by the vector $\left[\left(\Omega^D\right)^{-1}.D + p^T.\left(\Omega^V\right)^{-1}.V^{obs}\right]$with 105,625 elements, to give the term D'_z.

The pseudo code required to implement the Link Fusion is presented in Appendix D.

11. Link Fusion Tests Programme

11.1. Introduction

Before introducing the tests undertaken, this section describes the post processing of the outputs from the MATLAB Link Fusion process, for analysis in the SATURN traffic modelling suite. The post-processing discussed in the context of the Matrix Fusion tests was relatively involved as the output from the fusion was an updated prior matrix of proportions at sector-level. Compared with this, the Link Fusion post-processing is comparatively straightforward as:

- Being based on count data and prior matrix values (rather than proportions), no subsequent expansion is required
- Being in a cellular ij format rather than sector-level IJ format, no subsequent distribution is required
- Being an alternative method to matrix estimation by maximum entropy, the outputs may be compared directly with post-ME outputs from the original CSRM model development.

This last point is key to understanding the subsequent sections on the metrics for comparison, and the results. Link Fusion is not about improving the estimate of a prior matrix before subsequent ME; it is an alternative to ME itself. Note that all the Link Fusion tests involve a single application of Link Fusion, compared to the five iterations of ME used in the comparator CSRM model (see Section 10.2).

11.2. Post-Processing

The output from the MATLAB Link Fusion process is a 106,525 element array, D', representing the best estimate of the origin-destination Light vehicle movements for use in the CSRM SATURN base year AM peak assignment. The steps to process this are as follows:

- convert the array into 325*325 zone trip matrix as a .csv file
- convert the .csv to a .ufm file[53]
- apply purpose splits to convert this light vehicle matrix into the eight constituent light vehicle user classes, UC1-8 (using proportions taken from the original CSRM model)
- re-introduce the HGV user classes (UC9-10) by copying from the post SATME2 matrix from the original CSRM model
- assign the resulting 10 user class matrix to the CSRM network
- No expansion or distribution required.

11.3. Test Specification

11.3.1. Initial Tests

A considerable number of test runs were undertaken in MATLAB using the test-bed model, some of which were flawed in one way or another and which are not reported here. Understanding these test results led to numerous conclusions, as discussed later in this report, and refinements to the processing of the count data, the prior matrix and their variances, as discussed in the sections on input data below.

[53] An unformatted matrix file for use in the SATURN suite.

The initial tests also led to the creation of spreadsheet versions of the Link Fusion method to manually understand and check outputs. The most valuable of these was a simple example with six ij cells and six counts which was used to validate the MATLAB code for the element-wise calculations introduced above and reproduced in Appendix D2.

A similar spreadsheet example was used to confirm the validity of the approach using full covariance input matrices, Ω^D and Ω^V. This was necessarily simple (just two ij cells of demand and two counts in a network of five two-way links) given the matrix inversions required, but it provided comfort that the method does indeed work, even if it is impracticable to use full covariance matrices for real-world problems.

Finally, the initial tests led to the realisation that even when ignoring the off-diagonal terms in Ω^D and Ω^V, it is not possible to do this for the output covariances in M. As explained earlier, it is essential that these contributions are included in the calculation of D', even if there is insufficient space to store M, in which case the calculations need to be implemented on an element-wise basis.

11.3.2. Main Tests

The results of the initial Link Fusion tests suggested that the counts were making little difference, because their variances were high compared to the prior matrix variances. Given that the count variances are relatively easily derived, this in turn called into question the means of calculating the prior matrix variances. During the earlier Matrix Fusion work these were calculated at an IJ sector level[54] and so for use in the initial Link Fusion tests they were disaggregated to the ij cell level, as explained in detail in section 13.2. A final test, referenced in the brief for the final commission on this project, was to test the impact of disaggregating the IJ sector level variances on a different basis, such that the resulting Ω^D_{ij} variances were larger and the counts had greater scope to affect the fused output, D'_{ij}. The method of disaggregating the IJ sector level variances was therefore one of the key considerations in the definition of the main set of Link Fusion tests. These methods are discussed in section 13.2

A second consideration is the version of the prior matrix fed into the Link Fusion; specifically, whether the Link Fusion was applied to the version of the prior, D_{ij}, pre- or post-correction for errors in the A1 corridor in the original CSRM model development. As described in section 10.6.1 the intention was to use the post A1 correction version, to facilitate direct comparison with the SATME2 main matrix estimation results. Indeed, the prior variances, Ω^D_{ij}, were calculated assuming a post-correction prior matrix, so the tests relating to a pre-correction version of D_{ij} are not strictly correct and are only included herein for completeness. The test specifications are summarised below.

Table 11-1 Link Fusion Test Specifications

ID[55]	Means of disaggregating Ω^D_{IJ} to Ω^D_{ij}[56]	Input demand matrix, D_{ij} (trips)	
1	In proportion to $(p_{ij\|IJ})^2$	Pre A1-correction	77222
2		Post A1-correction	73877
3	In proportion to $p_{ij\|IJ}$	Pre A1-correction	77222
4		Post A1-correction	73877

[54] To be precise, the earlier IJ level variances were calculated at a 'fusion-block' level, though this doesn't affect the issue at hand.
[55] These tests were referred to as LF5 to LF8 in the final sets of working files in early 2017. LF6 was a successful validation check (to within 5 sf) of the Test LF4 undertaken much earlier in the study during 2015, undertaken to ensure continuity through the research.
[56] See section 13.2 for full details.

12. Link Count Data

12.1. Adjustments to Link Fusion Inputs

In the original CSRM main matrix estimation 174 light vehicle counts were used in the AM peak matrix estimation. In an attempt to improve results from the initial tests, these counts and their associated *Pija* files were adjusted in the following ways:

- 'Balancing' for inconsistent counts
- Greater consistency with the assignment *Pija* factors
- Dealing with inconsistencies between 'Demand' and 'Actual' assigned flows.

The following sections consider each of these in turn.

12.1.1. Count 'Balancing'

The count set was interrogated in detail and a handful of inconsistencies between upstream and downstream counts at junctions were identified. Any such counts were 'balanced' (i.e. the average between the upstream and downstream was taken) and these revised values subsequently input as the observed count-set, V^{obs}. The variance calculations (see section 12.3 below) for the affected counts were left unchanged.

12.1.2. Adjustment to *Pija* Factors

Initially the files of *Pija* factors for use in the link fusion process were taken from SATURN by using the KPP parameter within the SATPIJA module. In theory, when *Pija* is multiplied by the base matrix, it should produce the link demand flow. In most cases, this condition was met, but at some link locations the two values were found to be inconsistent. Possible inconsistencies were identified in the creation of the KPP file which it was thought might contribute to issues in the link fusion process. These inconsistencies included:

- The link fusion process only uses one user class. As such, routeing of the eight separate light vehicle user classes is not reflected.
- In producing the KPP, SATURN will run through a further iteration, so the assignment may differ slightly from that of the original prior matrix.

To account for these differences between the *Pija* file created by SATURN and the 'true' *Pija* file, a demand-weighted average select link based methodology was developed for the combined light vehicle user classes:

- Select link matrices were generated at each of the 174 count locations
- The total light vehicle total (user classes one to eight) was calculated from the output select link matrices for each of the 174 count locations
- These totals were divided through by the prior matrix light vehicle total (at an ij level) to produce the *Pija* matrices for each count location.

This methodology largely addressed the issue of *Pija* multiplied by the demand matrix not equalling the link demand flow. It should be noted that minor inconsistencies still remained; this is likely because the revised methodology still requires a further assignment iteration of SATURN to complete the select link analyses.

12.1.3. Demand uplift for counts

The final inconsistency in the link fusion inputs is down to the difference between demand and actual flows within SATURN. Link fusion is based on multiplying a *Pija* matrix by a demand matrix to give a link *demand* flow for comparison against a link count. The link count, however, is represented by an *actual* flow within

SATURN. To reflect this, the 174 link counts were uplifted to represent a demand flow rather than an actual flow using the ratio between the assigned demand and actual flows on each link in the prior assignment[57].

There is a further slight complication in the case of the CSRM - and indeed any congested assignment model where a pre-peak assignment has been undertaken in order to reflect the effects of traffic queued up from before the modelled peak hour. Where queues are carried from the previous modelled period, '*PASSQ*'[58], the demand matrix *D* should avoid double counting these. Thus the uplift factor for the desired count is calculated as:

$$V^{obs}{}_{Revised} = (V^{obs} - PASSQ) * \frac{Demand\ Flow}{Actual\ Flow}$$ for each link count.

12.2. Discussion of Count Adjustments

All of the above adjustments would generally have a small effect on the target count for use in the Link Fusion. They were investigated and the above solutions implemented in response to difficulties with the initial tests. In the light of experience it is possible that the final test results would be little changed if all these adjustments were to be removed, though this has not been tested. Note that in no cases were the variance calculations adjusted in response to these changes, though this may be worthwhile should the adjustments be significant.

While small, these adjustments are all valid in the sense that they should improve the overall result – in terms of the comparison of assigned actual flows with observed counts – and in some cases may be necessary to get a reasonable fit. This is especially true of the 'demand versus actual' adjustment which may be significant downstream of any highly congested location. The elimination of inconsistencies between counts, as in section 12.1.1, may in any case be viewed as best practice in matrix estimation – clearly any algorithm will fail to meet two mutually inconsistent targets. Finally, the recalculation of the *Pija* factors is a time-consuming process, possibly for relatively little gain, but it is worth flagging that in any practical implementation such issues would need to be highlighted to practitioners (who typically work with multiple car user classes despite the fact that the count data will only quantify the total number of cars).

12.3. Variance Assumptions

The derivation of an appropriate variance matrix for the observed counts presents further problems. As with the matrix fusion process, the Skrobanski paper is more or less silent on this aspect. While Appendix C of that paper reports an analysis of coefficients of variation, it is unclear whether these were used directly to provide the relevant variances.

For the CSRM test-bed model much count data was available which was processed in a manner consistent with Guidance, with the statistics associated with each factor duly calculated and recorded. This was one clear advantage of using the CSRM.

Guidance relating to the processing of count data is summarised in Appendix E.

During the development of the CSRM, both ATC and MCC counts were sourced from various surveys at different times and factored in line with the guidance cited above. The Local Model Validation Report for the CSRM describes (para 2.27) the:

> *"system of ratios … used to move these individual counts to an October 2006 average weekday, taking a mean of the ratios required to convert from:*
>
> *Each day in a week to the average for that week;*

[57] Note that the same process is automatically carried out within the SATME2 module of SATURN during matrix estimation.
[58] 'PASSQ' is used here as it is the name of the relevant function within SATURN.

Each monthly average weekday to the October average weekday in that year, and

Each yearly average October weekday to the average October weekday in 2006."

The ratios were calculated using relatively rich long-term ATC monitoring data for four sites on the A14 (to create factors for application to trunk road counts) and four sites on radial roads in Cambridge (to create factors for application elsewhere). Each ratio was calculated as the mean of many ratios observed at the monitoring count sites, together with an associated standard error in each. These were used in the Link Fusion work to derive the variances of the counts used.

In summary, if R_k represents a set of k ratios to convert the observed count, y, to the factored count x for the required annual average hour of the week (ie using the factors as listed above) then:

$$RSE(x) = \sqrt{\sum_k RSE(R_k)^2 + RSE(y)^2}$$

Where RSE is the relative standard error, which for the input count y is calculated as:

$$RSE(y) = \frac{\sigma_y}{\bar{y} * \sqrt{N_y}}$$

Where the mean observed value and standard deviation are $\bar{y}$ and σ_y respectively, calculated over N observations. Thus the required variances associated with the factored counts input to the Link Fusion, x, are calculated as:

$$\sigma_x{}^2 = (\bar{x} * RSE(x))^2$$

13. Matrix Demand Data

13.1. Prior Matrix

The matrix data to be fused with the link count data was essentially that from the CSRM introduced in section 3 of this report. This comprised the AM peak hour prior matrix summed over the eight light-vehicle user classes. However unlike the Matrix Fusion work which involved aggregating the prior matrix cells to create more statistically significant 'fusion-blocks', the prior was used in its original (325*325 cell) format. Two versions were used, as already introduced:

- Pre-A1 correction (comprising 77,222 pcu/hr);
- Post-A1 correction (comprising 73,876 pcu/hr).

13.2. Variance Assumptions

In principle, the variance matrix for the prior matrix *D* can be taken from the earlier process described in the context of Matrix Fusion. However, depending upon the level of spatial detail at which the Link Fusion is undertaken, complication may arise from the fact that the Matrix Fusion was done at a more aggregate level than the zonal ij. In practice, the Matrix Fusion process used estimates of aggregate cells which may be written $D_{IJ} = \sum_{i \in I} \sum_{j \in J} D_{ij}$. The disaggregation to zonal cells was done on the basis of the original prior matrix, by applying the appropriate proportions. The variances, Ω_{IJ}, may similarly be disaggregated.

However, if the proportion of an IJ fusion block in an ij cell is $p_{ij|IJ}$, then assuming *p* is an independent scalar the variances Ω_{IJ} and Ω_{ij} are related thus[59]:

$$\Omega_{ij} = p_{ij|IJ}^2 . \Omega_{IJ}$$

With the fusion blocks being aggregations of many cells, the individual values of $p_{ij|IJ}$ are small, and those of $p_{ij|IJ}^2$ are even smaller. In the belief that the poor initial test results may have been caused by these very small prior matrix variances, tests were also undertaken on the assumption that:

$$\Omega_{ij} = p_{ij|IJ} . \Omega_{IJ}$$

While having no mathematical basis, this form has the intuitive property that the trace of the resulting Ω_{ij} will equal that of the matrix Ω_{IJ}, on which it is based. It will also give rise to a greater contribution from the counts in the Link Fusion process.

Thus, given D_{IJ}, Ω_{IJ} from the Matrix Fusion, and the proportions $p_{ij|IJ}$ from the original prior matrix, the relevant quantities for the prior in the Link Fusion are derived as follows:

$$D_{ij} = p_{ij|IJ} . D_{IJ}$$

and $$\Omega_{ij} = p_{ij|IJ}^2 . \Omega_{IJ}$$

or $$\Omega_{ij} = p_{ij|IJ} . \Omega_{IJ}$$

Strictly, of course, the individual values of $p_{ij|IJ}$ should not be considered independent, as the proportions sum to 1, but as long as there are a reasonable number of constituent ij cells, this is probably not a major problem.

[59] As $Var(a.x) = a^2 . Var(x)$

13.2.1. Treatment of Bias from the 'A1 Correction'

As previously explained, the most appropriate prior matrix to use to compare the effects of Link Fusion with that of ME by maximum entropy is the Post A1-Correction version. However, the variances, Ω_{IJ}, were calculated for the Pre A1-Correction version of the prior matrix. A correction to the variances is therefore required, in order to reflect the biases introduced by the factoring in the A1-Correction process.

The method chosen was that described above to correct for adjustments made to the variances of the Trafficmaster GPS data (see Figure 4.2). If the bias introduced in the A1-Correction is b_{ij} then the correction to the calculated variance is:

$$\left[\exp\left(\left|\ln(b_{ij})\right|\right)\right]^2 \qquad \text{where} \qquad b_{ij} = \frac{T_{ij}^{Post\ A1\ Correction}}{T_{ij}^{Pre\ A1\ Correction}}$$

This correction factor was applied as a 325 * 325 matrix to the variances, Ω_{ij}, calculated above. The resulting corrected matrix of variances was converted into a column vector of 105,625 variances for use in the MATLAB code presented in Appendix D.

14. Link Fusion Results

14.1. Metrics for Assessment

Several alternative metrics were used to compare results from the Link Fusion tests, with each other and with the equivalent main matrix estimation results derived during the original CSRM model development:

- Trace of the covariance matrix, *M,* of the fused output, *D'*
- Sum of absolute changes between the input and output demand estimates*, D* and *D'*
- Sum of absolute changes between the input and output covariance matrices, Ω^D and *M*
- Sum of absolute differences between the observed counts, V^{obs}, and their modelled value in the initial assignment model, *p.D*
- Sum of absolute differences between the observed counts, V^{obs}, and their estimated modelled value after the Link Fusion, *p.D'*
- Change to prior matrix as assessed using WebTAG thresholds for acceptable changes during Matrix Estimation and through comparison of sector-to-sector flows.

And, after assignment of the resulting output, D', to the traffic model:

- Number of calibration counts which pass DMRB criteria
- Average GEH of calibration counts
- Number of independent validation counts which pass DMRB criteria
- Average GEH of independent validation counts

Finally some graphical evidence is presented which illustrates changes in assigned flows on the network and changes in the trip length distribution.

14.2. Initial Tests

The results from the initial tests were often counter-intuitive, with *D'* being far greater than expected, primarily owing to the omission of the negative covariance terms in *M*. In general, the results from these tests are not reported, but the table below is included as it illustrates the importance of including the off-diagonal terms of *M* in the calculation of *D'*. The comparison is based on a simplified example using just four counts

Table 14-1 Effect of Off-Diagonal Terms in M

Matrix	Full Matrix Total	Affected 1670 cells	Details
D	73,877	2,330	Prior post A1 correction
D' with off-diagonal terms of M	73,567	2,021	Less than 1% decrease
D' without off-diagonal terms of M	116,159	44,613	57% increase

The initial results did reveal that the link fusion methodology was capable of producing negative trips in the output (due to the inversion of the term included in the equation for *M*). Although potentially a problem, especially if outputs are used 'blindly' in downstream assignment and appraisal processes, these negative

values make only a small contribution to the overall results[60] in this study. However, it is possible that in other cases the effect of negative numbers could be significant and therefore it is recommended that any software that may be written to implement the Link Fusion method should check for negative numbers and, ideally, constrain them. The challenge would be to apply the constraint without compromising the quality of the fused output and there is no guarantee that this can be achieved.

14.3. Main Test Results

The results for the tests introduced in section 11.3.2 are presented below, first in terms of the mathematical metrics, then in terms of the changes made to the prior matrix, *D*, and finally in terms of comparisons between observed counts and estimates from the assignment models.

14.3.1. Mathematical Metrics

The key results are presented in the following table. As explained above, Tests 1 and 3 are of less interest as the variances used for these are not entirely consistent[61] and they cannot be directly compared with the main ME process from the original CSRM, unlike Tests 2 and 4. The metrics presented in the table are:

- Input prior matrix total, *D*
- Trace of the covariance matrix, Ω^{D}, of the input prior matrix, *D*
- Output fused matrix total, *D'*
- Trace of the covariance matrix, *M* or $\Omega^{D'}$, of the fused output, *D'*
- Sum of absolute changes between the input and output demand estimates, *D* and *D'*
- Sum of absolute changes between the input and output covariance matrices, Ω^{D} and *M*
- Sum of absolute differences between the observed counts, V^{obs}, and their modelled value in the initial assignment model, *p.D*
- Sum of absolute differences between the observed counts, V^{obs}, and their estimated modelled value after the Link Fusion, *p.D'*

Table 14-2 Link Fusion Results - Mathematical Metrics

ID	Pre/Post A1 Correction	Variance Multiplier	D	Trace (Ω^{D})	D'	Trace ($\Omega^{D'}$)	∑abs(∆D)	∑abs($\Delta\Omega^{D}$)	∑abs(V^{obs}-p.D)	∑abs(V^{obs}-p.D')
1	Pre	p^2	77222	56010	72901	48001	5735	8009	35805	18510
2	Post	p^2	73877	56010	72478	48001	3663	8009	23797	18430
3	Pre	P	77222	1509840	73731	1378528	17108	131313	35805	4503
4	Post	p	73877	1509840	73786	1378528	15424	131313	23797	4487

In all cases, the matrix total is reduced by the Link Fusion *(D'<D),* but more importantly the trace of the covariance matrix is reduced *(Trace($\Omega^{D'}$) < Trace(Ω^{D}))*. Thus, the confidence in the fused result is greater than the confidence in the prior. It can also be seen from the final two columns that as an approximation to

[60] Considering the results from the main test #2, the sum of the 22 negative values was 321.5. This represents 0.4% of the matrix total, spread over 0.02% of its cells.
[61] The variances used were those calculated for the Post A1 Correction, but the results are included nonetheless, for completeness.

the observed counts, V^{obs}, the output matrix improves on the 'fit' offered by the input matrix. It can be seen that the initial fit ($V^{obs} - p.D$) is better for Test 2 and 4 – which reflects the impact of the A1 Correction.

Considering the impact of the multiplier used in the disaggregation of the prior matrix variances, as this rises so too does Ω^{D} meaning that V^{obs} assumes more importance in the Link Fusion and the difference ($V^{obs} - p.D'$) is reduced. But the input and output variances (Ω^{D} and $\Omega^{D'}$) are greater. As p<1, the p^2 multiplier gives the lower input prior variances and commensurately lesser contribution from the counts, such that the error ($V^{obs} - p.D'$) is not reduced as much as when using p as the multiplier.

There is a balance to be made between the reduction in this error ($V^{obs} - p.D'$) and the confidence in the output solution. However, this is likely to be case specific and it would require more research to confirm if a rule of thumb could be used to determine where the balance point lies. Given that the p^2 multiplier has a basis in mathematics and yields a greater *relative* reduction in the variance ($Trace(\Omega^{D'}) / Trace(\Omega^{D})$ then we believe this is a better theoretical basis for determining the variances to use in the Link Fusion. However, in this example, the resulting error (between observed counts and $p.D'$ flows) after application of Link Fusion is substantially lower in Test 4.

In theory, this relative shortcoming of Test 2 could be overcome by applying the method on an iterative basis as described in section 10.2; this might yield a similarly relatively low error in the end, while retaining a higher confidence than possible with a method using greater prior matrix variances such as Test 4. Ideally the effects of different variance assumptions, and the use of Link Fusion on an iterative basis, would be investigated in greater detail using a number of different models, before firm conclusions are drawn.

It should be noted that the issue of applying the disaggregating factor, p, to the variances Ω^{D} is something that would not generally be encountered in practice; it is an artefact of the earlier stages of this study. However, the investigation of this has been helpful as it facilitates a more general comparison between cases with relatively higher or lower prior matrix variances.

14.3.2. Changes Made to the Prior Matrix

The remainder of the Link Fusion results focus on the two internally consistent tests based on the Post A1-Correction matrix. Using a tool developed by Highways England for the purposes of quantifying changes imposed on a prior matrix during matrix estimation, the Test 2 and Test 4 outputs have been compared with the input prior matrix. The tool assesses the changes against the criteria set out in WebTAG; changes in excess of the specified thresholds would not normally be deemed acceptable, if introduced through ME. However, there are two reasons why these thresholds may not be relevant for changes due to Link Fusion:

- While ME takes no account of the relative statistical confidence of the different inputs, Link Fusion is driven by such considerations; if the prior is unreliable but the counts are known to be highly accurate, it is perfectly valid for the counts to 'distort' the prior matrix significantly.
- There are many professionals in the field who have never agreed with the prescriptive application of these thresholds. The thresholds are in any case arbitrary and there is no reason to believe that Link Fusion changes should be within them.

However, as a benchmark to facilitate comparison between the changes imposed by ME and those due to the Link Fusion tests, the performance against the WebTAG thresholds is reported alongside the various regression parameters in the table below.

Table 14-3 Regression Comparison of Link Fusion Changes

Comparisons against Prior 'Post-A1 Correction'	Matrix Estimation (CSRM)		Link Fusion Test 2		Link Fusion Test 4	
	Value	WebTAG "Pass"	Value	WebTAG "Pass"	Value	WebTAG "Pass"
Regression of Origin Totals						
Slope	1.005	✔	0.976	✘	0.980	✘
Intercept	-0.568	✔	1.252	✔	4.177	✔
R2	0.958	✘	0.993	✔	0.962	✘
Regression of Destination Totals						
Slope	0.970	✘	0.974	✘	0.974	✘
Intercept	7.330	✔	1.533	✔	5.557	✔
R2	0.932	✘	0.993	✔	0.975	✘
Regression of OD Cell Totals						
Slope	0.996	✔	0.977	✘	0.983	✔
Intercept	0.004	✔	0.003	✔	0.011	✔
R2	0.945	✘	0.936	✘	0.866	✘

Looking first at the comparisons of trip ends – both origins and destinations - the R^2 statistics for Test 2 indicate a matrix which is far closer to the distribution of input prior matrix trip ends than in either the original ME process for the CSRM or in Test 4. Comparing the set of all OD matrix cells, the metrics suggest least change is imposed by the original ME process. In all cases, Test 4 with its inherent assumption that the prior is relatively less reliable, shows greater movement from the input prior matrix, as would be expected.

The following tables provide a more informative way to compare the changes made to the prior matrix. The five sectors referred to are those defined for the CSRM and shown in Figure 3-3:

- A/B Cambridge
- C Huntingdon
- E Other areas within Cambridge sub-region
- X Areas outside Cambridge sub-region.

The first three tables show the changes introduced by ME during the original CSRM model development.

Table 14-4 Prior (post A1 correction), distribution of demand across 5 sectors

PCU/hr	A	B	H	E	X	Total
A	950	1,225	40	1,248	513	3,976
B	1,224	2,520	65	2,347	1,265	7,421
H	56	40	970	1,237	348	2,651
E	3,574	4,078	2,703	23,846	7,537	41,737
X	1,721	2,465	804	6,917	6,183	18,091
Total	7,527	10,328	4,581	35,595	15,846	73,877

Table 14-5 Post ME, distribution of demand across 5 sectors

PCU/hr	A	B	H	E	X	Total
A	373	1,033	57	1,582	617	3,662
B	888	1,420	73	2,184	1,001	5,567
H	26	21	1,121	1,368	294	2,829
E	3,651	5,060	3,647	24,518	7,441	44,317
X	1,571	2,302	824	7,038	5,944	17,679
Total	6,509	9,835	5,723	36,690	15,297	74,054

Table 14-6 % Change between Prior and Post ME

PCU/hr	A	B	H	E	X	Total
A	-61%	-16%	44%	27%	20%	-8%
B	-27%	-44%	12%	-7%	-21%	-25%
H	-54%	-48%	16%	11%	-16%	7%
E	2%	24%	35%	3%	-1%	6%
X	-9%	-7%	3%	2%	-4%	-2%
Total	-14%	-5%	25%	3%	-3%	0%

The ME processes used in the development of the original CSRM created inter-sector changes of up to 54% and intra-sector changes of up to 61%. Overall the matrix total changed by only 0.2%.

Table 14-7 Post Link Fusion Test 2, distribution of demand across 5 sectors

PCU/hr	A	B	H	E	X	Total
A	953	1,231	40	1,261	521	4,005
B	1,230	2,522	65	2,342	1,250	7,410
H	55	40	964	1,194	287	2,539
E	3,579	4,085	2,673	23,710	7,019	41,065
X	1,738	2,476	831	6,382	6,032	17,460
Total	7,555	10,353	4,572	34,889	15,109	72,478

Table 14-8 % Change between Prior and Post Link Fusion Test 2

PCU/hr	A	B	H	E	X	Total
A	0%	0%	0%	1%	2%	1%
B	0%	0%	0%	0%	-1%	0%
H	-3%	-1%	-1%	-3%	-17%	-4%
E	0%	0%	-1%	-1%	-7%	-2%
X	1%	0%	3%	-8%	-2%	-3%
Total	0%	0%	0%	-2%	-5%	-2%

The changes introduced by Link Fusion Test 2 are far smaller than those seen in the original ME with only one inter- or intra-sector movement changing by more than 10%. The matrix total changes by 2% - which is much more than with ME but still a small change by the standards of matrix development for transport models.

Table 14-9 Post Link Fusion Test 4, distribution of demand across 5 sectors

PCU/hr	A	B	H	E	X	Total
A	989	1,385	41	1,445	671	4,532
B	1,330	2,687	67	2,316	1,040	7,439
H	51	37	1,089	1,430	274	2,882
E	3,726	4,137	2,847	23,708	6,638	41,055
X	1,840	2,587	895	6,361	6,196	17,879
Total	7,936	10,832	4,939	35,259	14,820	73,786

Table 14-10 % Change between Prior and Post Link Fusion Test 4

PCU/hr	A	B	H	E	X	Total
A	4%	13%	4%	16%	31%	14%
B	9%	7%	2%	-1%	-18%	0%
H	-9%	-8%	12%	16%	-21%	9%
E	4%	1%	5%	-1%	-12%	-2%
X	7%	5%	11%	-8%	0%	-1%
Total	5%	5%	8%	-1%	-6%	0%

The changes introduced by Link Fusion Test 4 are also far smaller than those seen in the original ME with a maximum inter- / intra-sector movement change of 31% and all others at 21% or below. The matrix total changes by 0% - which is consistent with the change as a result of ME. However, compared to the changes resulting from Test 2, the output matrix is altered considerably more in Test 4. This reflects the relatively greater uncertainty assumed for the prior in Test 4.

14.3.3. Assignment Model Metrics

The following table presents the quality of the resulting calibration performance over the 174 counts used, compared to that achieved in the main ME of the original CSRM.

Table 14-11 Calibration Count Performance

Summary of 174 calibration counts				
	Prior	**Original CSRM**	**Test 2**	**Test 4**
Total Passing Link Flow or GEH criteria	106	144	121	150
Average GEH	5.67	2.90	4.99	2.60

The reduction in error against the calibration count set mentioned above ($V^{obs} - p.D'$) is borne out by the SATURN assignment results. 121 calibration counts 'pass' the DMRB criteria with Test 2, compared to 106 with the Prior assignment and 144 for the equivalent post-ME2 assignment from the CSRM. However, with Test 4, 150 calibration counts 'pass' – ie an improvement on matrix estimation.

The calibration results for Test 4 relative to Test 2 are to be expected; they prove only that as we increase the relative contribution of the input counts compared to the Prior, we obtain a result which better approximates the counts. However, while Test 4 may be compared directly with Test 2, the direct comparison with the ME undertaken in the original CSRM is somewhat compromised. In the CSRM there were many trip end constraints used for Cambridge City in the original ME process (as introduced in section 3.1), so the algorithm was not focused solely on the 174 counts used in the Link Fusion tests; had there not been any trip end constraints it is possible that more than 144 of the calibration counts in the CSRM would have met the flow or GEH criteria.

Crucially, in Test 4 we also see an improvement in the number of independent *validation* counts passing, as below.

Table 14-12 Validation Count Performance

Summary of 65 validation counts				
	Prior	**Original CSRM**	**Test 2**	**Test 4**
Total Passing Link Flow or GEH criteria	46	53	48	55
Average GEH	4.6	3.6	4.2	3.5

That Test 4 can improve on the number of calibration counts *and* validation counts which meet the DMRB acceptability criteria is significant. This is clear evidence that sometimes Link Fusion may be a superior approach to matrix calibration than existing maximum entropy ME methods.

14.3.4. Graphical Comparisons

The following plots compare the prior and post matrices (ie after ME or Link Fusion), as assigned flow differences. In each plot the bandwidths are proportional to the change in assigned flow. The scales are identical to facilitate direct comparison between plots. The chapter concludes with a comparison between trip length distributions for the CSRM Prior, post-ME and Link Fusion test outputs.

Figure 14-1 Difference Plot (Prior post A1 correction compared to Post ME, green is increase, blue is decrease)

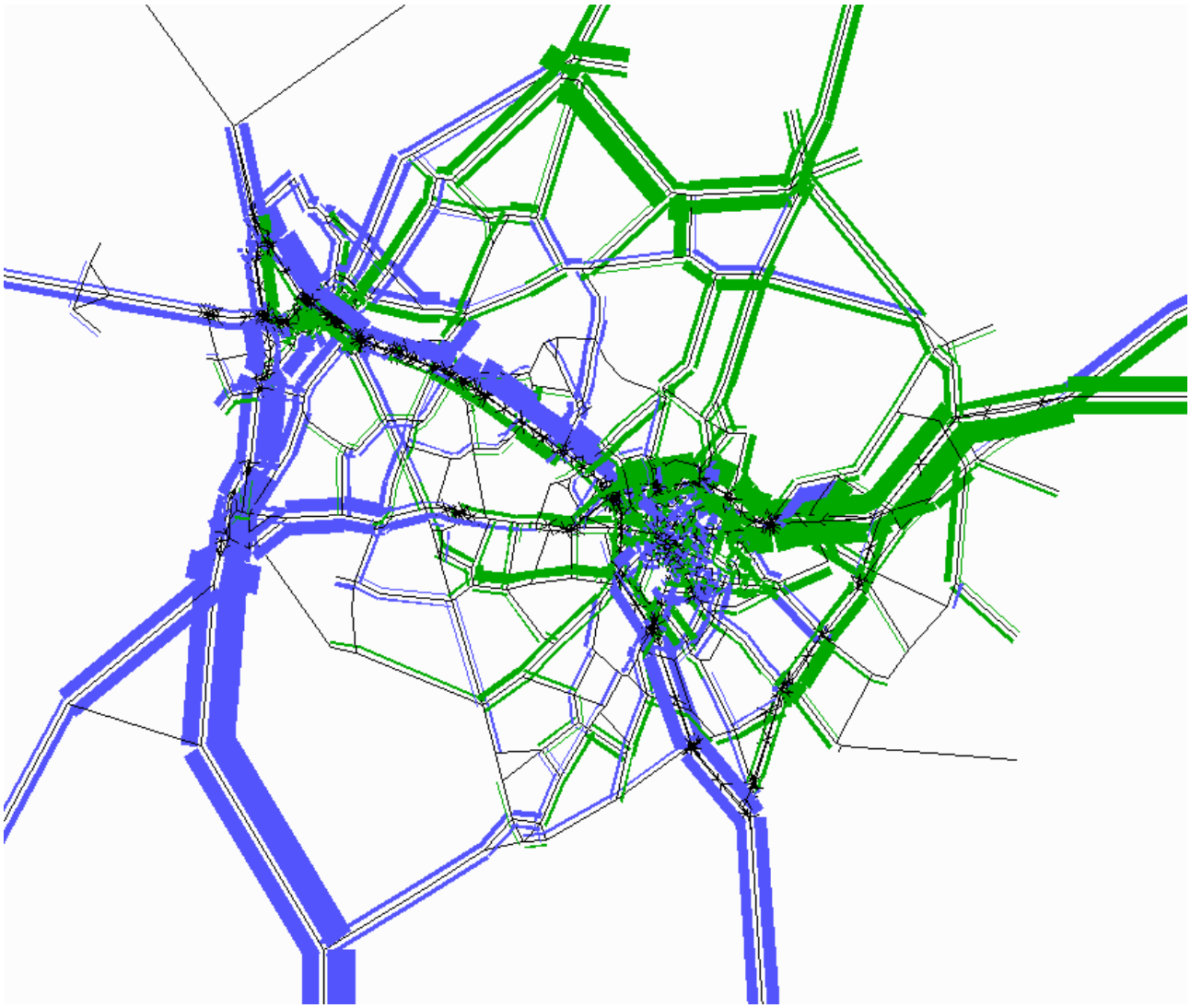
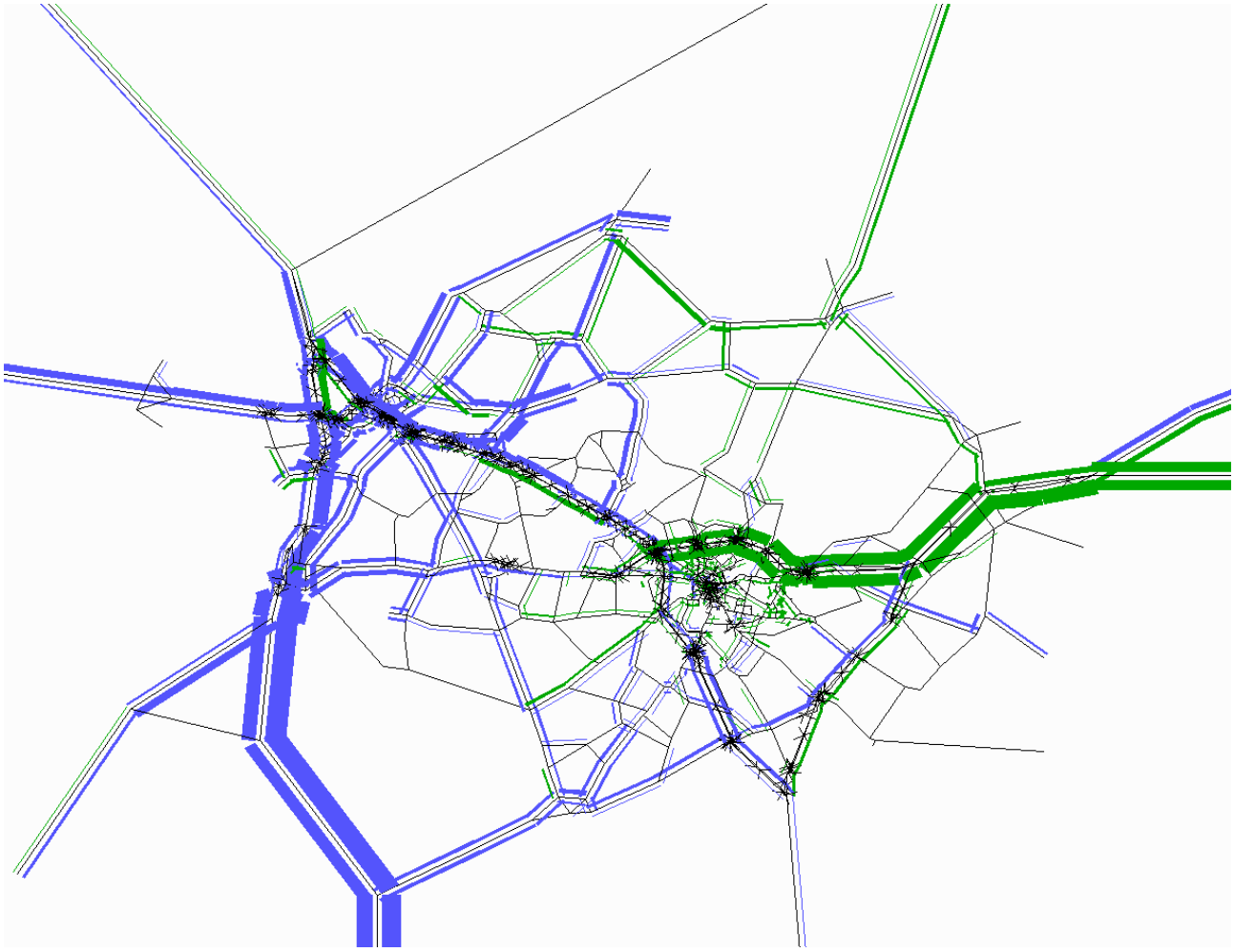
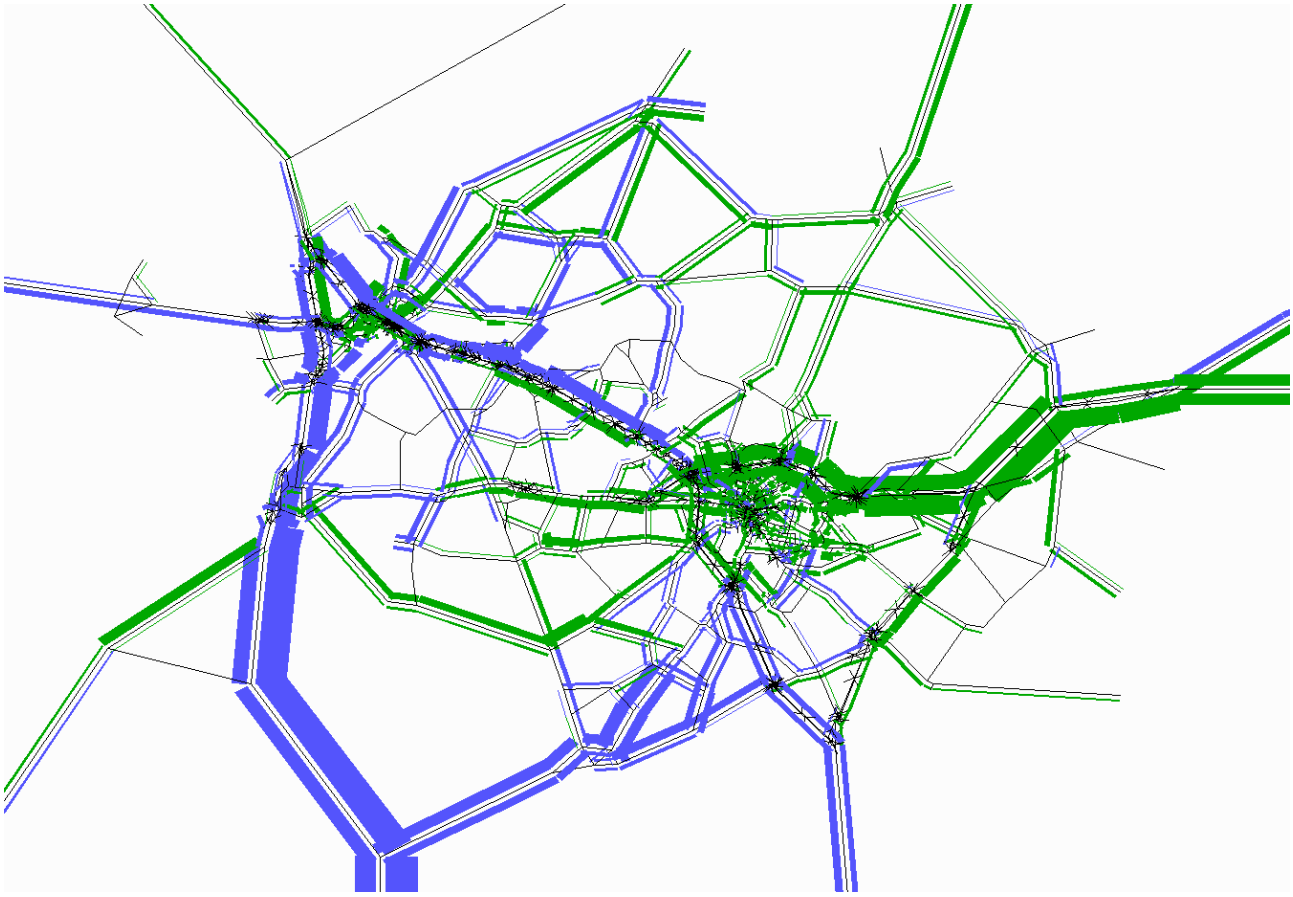


Figure 14-2 Difference Plot (Prior post A1 correction compared to Test 2, green is increase, blue is decrease)



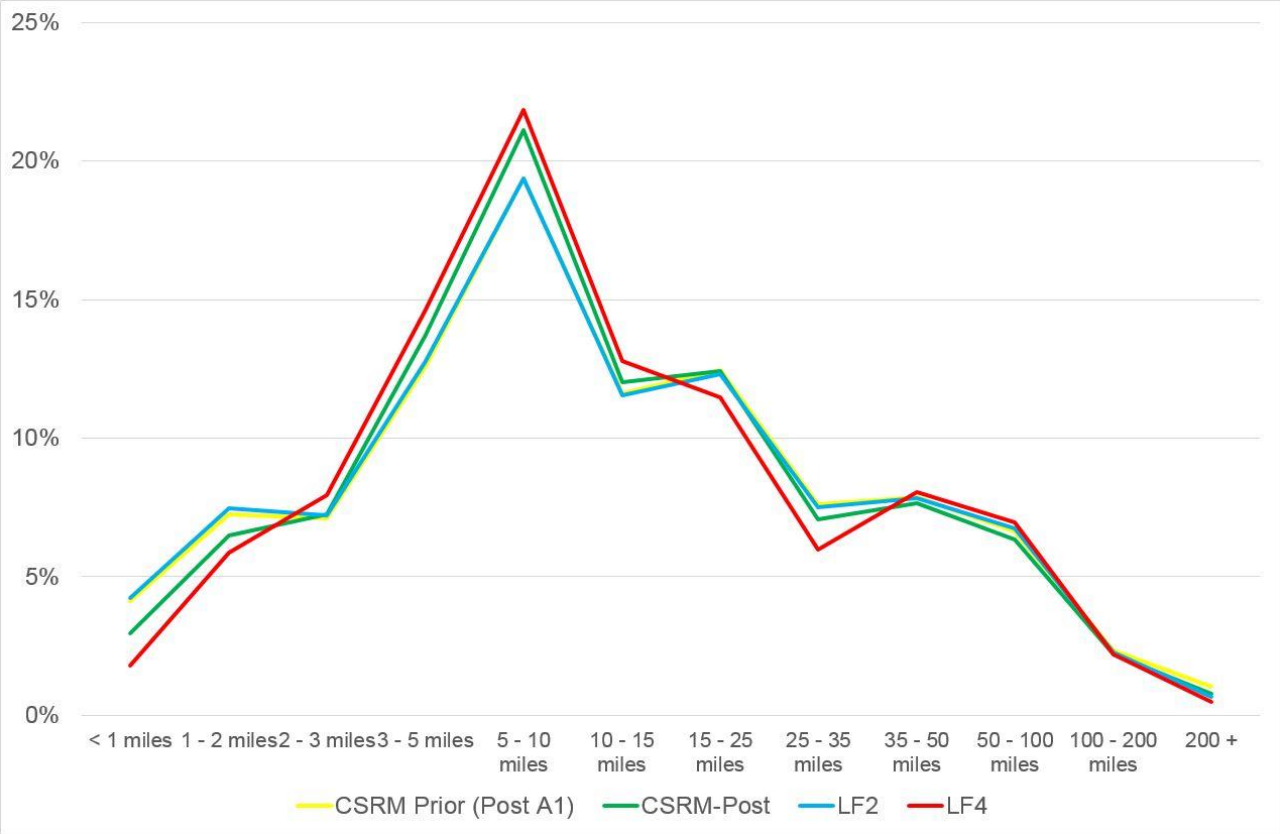
The flow changes brought about by the Link Fusion Test 2 are far less than those seen in the original post-ME assignment. Comparing Figure 14-2 with Figure 14-1, the changes are less widespread and (without detailed analysis) appear to be equal or less great on all links. This is highly significant and completely in line with expectation given the sector-level changes reported above.

Figure 14-3 Difference Plot (Prior post A1 correction compared to Test 4, green is increase, blue is decrease)



The observed flow differences are marginally greater than those observed above for Test 2, but still less than the changes due to the original ME. This is consistent with the above sector-based analyses – Test 4 inevitably results in more change as it assumes a lower confidence in the prior.

Figure 14-4 Link Fusion Trip Length Distribution Comparison



The trip length distributions are all very similar but it is noteworthy that the Link Fusion Test 2 distribution is closer to the original prior distribution than is the distribution from the original ME calibration outputs (CSRM-post) and Link Fusion Test 4. This is consistent with the observations made from the above sector-based changes and flow-difference plots.

15. Link Fusion Conclusions

15.1. Introduction

In contrast to the Matrix Fusion conclusions, the Link Fusion conclusions have potentially significant implications for the future development of trip matrices from mixed matrix and count data. As long term traffic counters provide all the data required for accurate calculation of variances, in this sense, a count dataset is a better starting point than the 'big data' matrix datasets considered earlier in this study. That said, an isolated count is not a rich data source and even a wide spread of counts across a survey area relies on the existence of some form of assignment model with which to interpret the link counts in terms of origin to destination movements.

Where a reasonable assignment model exists, with which to provide the necessary '*Pija'* factors, then the research work undertaken suggests that Link Fusion does 'work' and is probably realistic to implement in most cases.

15.2. Link Fusion Implementation

The following conclusions relate to the implementation of the Link Fusion method:

- as with Matrix Fusion, there may be numerous practical problems with estimating the variances of the prior matrix (Ω^D), due to use of synthetic data of unknown reliability and of non-sampling errors
- in practice, the covariance matrices of the prior matrix (Ω^D) and the input count-set (Ω^V) need to be assumed diagonal
- the output covariance matrix, *M*, used in the last step to derive the output fused matrix *D*' can be expected to have significant off-diagonal elements
- the trace of *M* is a useful indicator of the degree of uncertainty of the output, an indicator which is not available for an equivalent entropy maximising approach
- on larger problem sizes, memory constraints are inevitable and it is likely that sparse matrix and/or geographic segmentation techniques would be required to keep the problem-size manageable. We have not investigated these but note that a) the matrices are not necessarily sparse and b) the techniques are heuristic in nature and not easily automated for general application
- temporary memory storage issues can be overcome through the use of element-wise code, although at a cost of longer runtimes. However, both storage space and runtime can be optimised if the problem size deems it necessary, by implementation of the method as a bespoke software making use of sufficiently powerful hardware via parallel processing
- in one of the practical tests undertaken using the CSRM, the Link Fusion approach resulted in a output with better performance than entropy maximising ME, in terms of both calibration and independent validation results[62]
- in terms of changes that the method imposes on the prior matrix, the changes are greater where the input counts are deemed to be more reliable[63]

[62] Further, the Link Fusion calculations were applied only once, while the comparator CSRM matrix estimation was undertaken using five iterations of the SATME2 software.
[63] Using the CSRM testbed model and its observed count-set, the changes imposed on individual OD cell values were found to be least with the original ME process, but at sector level the changes were consistently

- the initial results demonstrated that the Link Fusion methodology is capable of producing negative trips in the output (due to the inversion of the term included in the equation for *M*).

15.3. Discussion

The current work with the CSRM Cambridge model employs a mix of element-wise and block(matrix)-wise code which works on a standard laptop without needing to use sparse matrix techniques. In general there is balance between use of element-wise code (requiring little memory) and block(matrix)-wise code which is likely to be faster but may require the application of heuristic techniques as the problem size increases, depending on memory constraints. This balance should be the subject of further investigation.

It is important to note that the Link Fusion method should have a definite theoretical advantage over ME in that it takes account of input reliability and provides error estimates for the outputs too. In particular, it does not "force" the matrix to reproduce the counts which are themselves, of course, subject to error.

Our work has shown that Link Fusion does improve the Prior matrix estimate and that in one case this was more successful than ME using the SATME2 software. While the concept has been proved it is not possible to conclude whether the technique will *generally* improve on ME, without many further tests[64]. It should then be possible to select the best technique to maximise the value of the (potentially expensive) data collated for the study in hand.

Even if it can be shown that Link Fusion is generally (or in certain circumstances) an improvement on conventional entropy maximising ME, further work would be required to confirm if the improvement is sufficient to justify the additional technical complexity/costs of Link Fusion software development.

The possibility of outputting negative numbers is potentially a problem. It is noteworthy that in Skrobanski *et al*[1] the assumption is that negative numbers will not arise[65]. Our work suggests this is incorrect, and that the Skrobanski solution to the optimisation problem may need to be re-visited. It is recommended that any software that may be written to implement the Link Fusion method should check for negative numbers and, ideally, constrain them. The challenge would be to apply the constraint without compromising the quality of the fused output and there is no guarantee that this can be achieved.

The possibility of iterating the solution, using improved assignments to provide updated '*Pija*' factors at each iteration is also worthy of further research. Such iteration is certainly found to be helpful in the context of ME and there is no reason to suggest Link Fusion should be different in this respect. However, with potentially long run-times this may not be a practical solution for the largest problem sizes.

In summary, the Link Fusion method 'works', and *may* be the best method available, from a theoretical standpoint, but the practical difficulties are significant and further research would be required before concluding that a bespoke software implementation is justified and, if it were, how to optimise the run-times for such software.

---

less great with Link Fusion. However, without considering other models and/or count-sets it is not possible to draw any conclusions from this observation.

[64] there is some evidence to suggest that on balance entropy maximising techniques may outperform Least Squares techniques such as Link Fusion - based on four performance tests using Monte Carlo simulation, a Mean Square Error objective measure, and considering normal data and three types of messy data – missing values, outliers and multi-collinearity: http://www.researchgate.net/publication/236108549_Comparing_Generalized_Maximum_Entropy_and_Partial_Least_Squares_Methods_for_Structural_Equation_Models

[65] from the foot of page 5, "The properties of the OD Fusion problem mean that the constraint on trips being positive applies automatically and we drop further explicit reference to it."

15.4. Implications for Iteration

The prior matrix variances used in the Link Fusion were disaggregated from those variances initially calculated for the Matrix Fusion research at a more aggregate level. Different methods to disaggregate them were considered. Thus, due to the evolution of the study, Link Fusion was undertaken using variances calculated in two different ways. As documented in section 14.3.1 this gave different results, with greater changes in the output matrix when the prior matrix variances were relatively high compared to the count variances. These greater changes were associated with a reduced 'error' in the output, i.e. a reduced discrepancy in flows at the link level ($p.D'\text{-}v^{obs}$), but with less confidence in the result (i.e. a greater trace for the output covariance matrix, *M*). This relationship is intuitive but it is not straightforward to extrapolate its implications where iteration is being considered.

Should multiple runs of Link Fusion be undertaken, for instance by updating the *Pija* file at each iteration, as suggested in section 10.2, then it would be expected that the error ($p.D'\text{-}v^{obs}$) would reduce progressively, at least to a point, though it may not in practice be convergent. It is not clear how the output covariance matrix *M* would change over the iterations though it is reasonable to assume that the lowest uncertainty will be associated with those cases where the input count and prior matrix variances are smallest. Given that the count variances are likely to be well understood, the issue is what to assume for the prior matrix variances (though sometime these too may be reasonably well understood and there may be little scope for different variance assumptions). Opting for a relatively greater prior matrix variance will yield a result with less error ($p.D'\text{-}v^{obs}$) but greater uncertainty in the output. There is a balance to be had but, where resources allow, opting for the least input variances possible and then iterating the process if necessary should give the best overall result.

With entropy maximising methods this is not an issue; no consideration is given to uncertainty and all effort goes into minimising the error ($p.D'\text{-}v^{obs}$).

15.5. Further Research

We recommend that the following points are investigated further.

1. Systematic comparisons of Link Fusion with SATME2, over a variety of examples, are required to ascertain whether one generally outperforms the other in terms of the various metrics introduced herein. Examples should be selected with reference to the criteria set out earlier in this report, but also should avoid any complicating issues in existing SATME2 work – such as trip end constraints or link flow inequalities. If Link Fusion is superior, is it superior enough to justify software development?
2. Iterating *M* and *D'* could be used to either reduce the error ($p.D'\text{-}v^{obs}$) or the uncertainty of the output, or both. However, for most practical problem sizes, it is impractical to feed the full covariance matrix *M* back into the Link Fusion process. How acceptable might it be to iterate while dropping the covariance terms in *M* from the calculations in the subsequent iteration?
3. Whether or not the above point yields useful results it is worth attempting to iterate on the basis of *Pija* rather than *M* and *D'* – or indeed to consider nested loops of both forms of iteration. Iteration of *Pija* is considered in this report but has not been tested.
4. Where different prior matrix variance assumptions are reasonable (i.e. 'greater' or 'lesser' variances), which is better in terms of optimising the balance between error and uncertainty, while attempting to minimise the number of iterations where relevant? The answer will be case specific, but it may be possible to develop some rules of thumb.
5. Where is the balance point between fast matrix-wise calculations with big space requirements and slower element wise calculations that will always work but may be slow? Is it possible to derive

general advice on the use of sparse matrix techniques, the selection of low value thresholds and the application of partitioning in cases where memory runs out?

6. Can any negative output values of *D'* be identified and removed without significantly undermining the confidence in the output? If not, can the likelihood and scale of such negative values be predicted?
7. In terms of the practical issues of developing the inputs to Link Fusion, do weighted average *Pija* factors[66] and demand uplift factors[67] make a material difference? If so, is it also worth adjusting the count variances accordingly? Likewise, are there better approximations for the treatment of non-sampled variances than those suggested herein and used in this study?

[66] weighted across assignment user classes
[67] to reflect differences between assigned demand and the modelled actual flows that might result, in the context of upstream bottlenecks in the assignment model

Appendices

Appendix A. Initial Review of Candidate Test Bed Models

A.1. Long List

Table A-1 Long List of Models Considered

Model	Scheme
Wakefield Area Motorway Model	M1 J39 to J42 Managed Motorway
East of England Regional Model	-
A14 Local Model	Kettering
Lower Thames Crossing Model	M25 Junction 30
Cambridge Sub-Regional Model (CSRM)	A14 Improvements
Tyne and Wear Model	-
A30 Junction Model	A30 junction study
A5-M1 Local Area Model	A5/M1 Link
M3/M4 Model	M3/M4 Managed Motorways
A11 Local Model	A11
A453 Transport Model	A453 Multi-Modal Study
M1 Junction 19 Model	M1 Junction 19
A21 Tonbridge to Pembury traffic model	A21 Tonbridge to Pembury
Bedford DC Model	A421
Stevenage Hitchin Model	-
PRISM West Midlands Model	-
Norwich Area Transport Study (NATS) Model	-
South Yorkshire Strategic Transport Model+ (SYSTM+)	-

The South Bristol Link (SBL) model was also proposed as having a sound methodology for matrix build, and passing the calibration and validation criteria stipulated by WebTAG. However, the local authority that owns the model indicated they could not approve its use whilst the SBL scheme was subject to the planning process.

A.2. Initial Review of Long List

Up-to-date validation reports for all of these models were obtained and then reviewed at a high level. Two conclusions became apparent early in this review:

- Few models were found that did not use roadside interview data at a zonal level. Only the Tyne-and-Wear and Cambridge (CSRM) models genuinely meet this requirement. The M3/M4 model did group RSI data into larger sectors, but only to merge 90% of the zone-zone data with 10% of sector-sector data;
- Few models were found that applied matrix estimation at a short screenline level. Only the Lower Thames Crossing Model used only short screenlines in matrix estimation; both the Tyne and Wear and CSRM made some use of short screenlines, but used individual sites as well (CSRM used all automatic counts individually, and grouped only manual counts into screenlines). No other models reported making any use of short screenlines. It was noted that while guidance advising the use of short screenlines was available in WebTAG Unit 3.19 (now M3-1), this unit was relatively new and post dates the development of many of these models; the use of short screenlines was not formerly common practice.

While a model developed using reasonably good practice was sought (as comparing mobile data against a notably poor model would give an unrealistic picture), an ideal “perfect” model was not expected. The object of the study is to compare the inclusion of GPS and mobile data with reasonable “normal” practice. Unfortunately, neither of the above recommendations (use of RSI data only at a sectoral level and use of short screenlines in matrix estimation) appears to represent recent normal practice. It would, still, however, be preferable to have found a model that follows emerging best practice if possible.

Where the documentation failed to state the use of short screenlines, it was assumed that these were not used; no models were wholly explicit in saying that short screenlines were not used, but this would not be expected where the model development preceded changes in guidance . Similarly, it was generally possible to infer from the matrix build methodology that RSI data must have been used at a zonal level, because there was no described methodology for using at any other level, but again, this was rarely totally explicit.

One indication that use of observed data at an aggregate level might have been overlooked during the review, was that an initial review of the CSRM LMVR gave no indication that the RSI data had not been used at a zonal level; however a more detailed matrix development note that we were able to obtain was clear that the data was aggregated.

Some other points also became clear following the review:

- Very few of these models used recent RSI data (within six years of the review). It is noted that, due to the recession, there has been less funding available for model development in recent years.
- Many of these models have had long, involved matrix-build histories, with models built using matrices from earlier versions of the model or merging matrices with those from more strategic models. Several models used previously-estimated matrices in their prior matrices. Such models should be avoided if possible, due to both the difficulty of tracing any data back to its source and the probable low-quality of the prior matrices. This is a particular issue given the focus of the search on small, scheme-specific models; almost all of these models took external to external demand (post-estimation) from a parent strategic model. In the review we found that in a few cases the latest LMVR did not document the complete matrix-building process properly and several earlier LMVRs would have been required as well to fully understand the process.
- Relatively few models reported changes brought about by matrix estimation fully as recommended in the current WebTAG Unit M3-1 (we note that this guidance and the preceding Unit 3.19 is quite new). Four models did not report these matrix changes at all. Of those that did report changes in

some way, it was evident that all failed to meet the WebTAG criteria, rendering the requirement that matrix estimation change the prior matrix significantly, a non-issue.

- Most models appeared to have several validation screenlines with reasonably comprehensive coverage. These were not checked exhaustively for holes, but casual inspection suggested that most did contain at least minor holes. Some models had screenlines with no holes, but almost all will have had minor non-modelled roads as holes in at least some screenlines.

Appendix B. Matrix Fusion MATLAB Code

B.1. Introduction

The MATLAB code provided below is annotated with comments. The code represents that required for a matrix-based implementation using the full covariance matrices.

Note that in the following pages, lines beginning with the % modifier in the source code denote the paragraph as a comment.

Note that in MATLAB, the * operator denotes matrix multiplication, the .* operator denotes element-wise multiplication.

Note that source code lines are ended with the terminator ;.

% The column vector t_1 is represented by `column_vector1`

% The covariance matrix Ω_{t_1} is represented by `covariance_matrix1`

% The column vector t_2 is represented by `column_vector2`

% The covariance matrix Ω_{t_2} is represented by `covariance_matrix2`

% The column vector t' is represented by `fused_column_vector`

% The covariance matrix Ω' is represented by `fused_covariance_matrix`

B.2. MATLAB Code for Matrix Fusion using full Covariance Matrices

```
function [fused_column_vector, fused_covariance_matrix] =
fusion_with_covariance(column_vector1,covariance_matrix1,
column_vector2,covariance_matrix2)

% Get rank of covariance_matrix1 in two independent ways as a cross-check.

size1 = max(size(covariance_matrix1))
determinant1 = det(covariance_matrix1);
[U1,S1,V1] = svd(covariance_matrix1);
tol1 = size1 * eps(max(max(S1)));
rank1 = sum(sum(S1 > tol1))
rank1 = rank(covariance_matrix1)

% If covariance_matrix1 is not of full rank then use singular value decomposition (SVD) to compute the
pseudo-inverse in place of using normal matrix inversion, else covariance_matrix1 is of full rank so use
normal matrix inversion.

if rank1 < size1
   dependents1 = [];
   vectors1 = [];
   for vector = 1:size(S1,1)
      if S1(vector,vector) <= tol1
          dependents1 = [dependents1 V1(:,vector)];
          vectors1 = [vectors1 vector];
      end
   end
   matrix1 = pinv(covariance_matrix1);
else
   matrix1 = inv(covariance_matrix1);
end

% Get rank of covariance_matrix2 in two independent ways as a cross-check.

size1 = max(size(covariance_matrix2))
determinant2 = det(covariance_matrix2);
[U2,S2,V2] = svd(covariance_matrix2);
tol2 = size2 * eps(max(max(S2)));
rank2 = sum(sum(S2 > tol2))
rank2 = rank(covariance_matrix2)

% If covariance_matrix2 is not of full rank then use singular value decomposition (SVD) to compute the
pseudo-inverse in place of using normal matrix inversion, else covariance_matrix2 is of full rank so use
normal matrix inversion.

if rank2 < size2
   dependents2 = [];
   vectors2 = [];
   for vector = 1:size(S2,1)
      if S2(vector,vector) <= tol2
          dependents2 = [dependents2 V2(:,vector)];
          vectors2 = [vectors2 vector];
      end
   end
   matrix2 = pinv(covariance_matrix2);
else
   matrix2 = inv(covariance_matrix2);
end
```

```
matrix3 = matrix1 + matrix2;
```

% If `matrix3` is not of full rank then use singular value decomposition (SVD) to compute the pseudo-inverse in place of using normal matrix inversion, else `matrix3` is of full rank so use normal matrix inversion.

```
% Get rank of matrix3 in two independent ways.
size3 = max(size(matrix3))
determinant3 = det(matrix3);
[U3,S3,V3] = svd(matrix3);
tol3 = size3 * eps(max(max(S3)));
rank3 = sum(sum(S3 > tol3))
rank3 = rank(matrix3)
if rank3 < size3
   matrix4 = pinv(matrix3);
else
   matrix4 = inv(matrix3);
end
```

% The covariance matrix Ω' is represented by `fused_covariance_matrix`

```
fused_covariance_matrix = matrix4;
```

% Calculate Eqn 6 of the OD Data Fusion paper: $t' = \left(\Omega_{t_1}^{-1} + \Omega_{t_2}^{-1}\right)^{-1}\left(\Omega_{t_1}^{-1} t_1 + \Omega_{t_2}^{-1} t_2\right)$

```
vector3 = matrix1 * column_vector1;
vector4 = matrix2 * column_vector2;
vector5 = vector3 + vector4;
```

% The column vector t' is represented by `fused_column_vector`

```
fused_column_vector = matrix4 * vector5;
```

Appendix C. SATME2 Matrix Estimation

C.1. SATURN Assignment

Considering first the assignment process, SATURN aims to find a user equilibrium according to Wardrop's First Principle. Ignoring, for simplicity, the question of user class, this can be expressed as follows:

With demand represented by the matrix Dij, the equilibrium loadings may be obtained by minimising the objective function Z = $\sum_a \int_0^{V_a} c_a(V_a)dV_a$ subject to the following conditions:

$D_{ij} = \Sigma_r R_{rij}$	total demand
$V_a = \Sigma_{ij} \Sigma_r R_{rij} . \delta_{ra\|ij}$	link flows
$C_{rij} = \Sigma_a c_a . \delta_{ra\|ij}$	path costs
$R_{rij} > 0$ iff $C_{rij} = \min_r C_{rij}$	all used paths have minimum cost

where r represents a path, a is a link, and i and j are zones. The (0,1) matrix $\delta_{ra|ij}$ denotes whether route r contains link a. R_{rij} is the volume of demand between i and j allocated to path r. c_a is the link cost, and is a function of the link flow.

The solution is obtained by an iterative procedure (the details are not material), and under congested conditions it can be expected that multiple paths will be chosen between most ij pairs.

Note that the SATURN "PIJA" factors represent the proportion of ij trips using link a. Hence PIJA = $p_{ija} = \sum_r \frac{R_{rij}}{D_{ij}} . \delta_{ra|ij}$. Note that this corresponds with the "assignment matrix" referred to in Skrobanski et al (2012)[68].

C.2. SATURN ME2

As described in the SATURN manual, the aim of the SATURN matrix estimation is to derive an updated matrix D'_{ij} using observed counts V_a^{obs}, with the property that it is "close" to the original matrix D_{ij} and also that

$\Sigma_{ij} D'_{ij} p_{ija} = V_a^{obs}$

This is achieved by specifying an "entropy" criterion to be maximised with respect to the items of the updated matrix D′ij:

max $\sum_{ij} \left(D'_{ij} . \ln(D'_{ij}/D_{ij}) - D'_{ij} + D_{ij} \right)$

for which the derived solution can be written as $D'_{ij} = D_{ij} \prod_a X_a^{p_{ija}}$ where X_a is a "balancing factor" for each observed link count, to be estimated.

[68] Skrobanski G, Logie M, Black I, Fearon J, Dong Y, Gilliam C (2012), Further Developments In OD Data Fusion Methodologies, European Transport Conference

Note that the set of PIJA values is dependent on the assigned matrix D_{ij}, and is likely to change as a result of the updated matrix. As shown in Figure 13.2 of the SATURN Manual, an iterative approach is followed "*whereby an assignment is used to derive the route choice/PIJA factors which are in turn used to estimate a revised trip matrix. This is then reassigned and the process continued until stable values are found*."

Existing Guidance (WebTAG M3.1) does not offer advice on this point, though it would be useful to know what stopping criterion might be used. In the original CSRM model development five[69] iterations were undertaken. It was considered that as the iterative process is not necessarily convergent there is rarely much benefit in going through more than a few iterations.

[69] In the CSRM, up to six iterations were undertaken, with results taken from the iteration which produced the best post-ME validation. For the AM peak model variant this meant the fifth, and as a consequence all results and analyses continue to be based on the output from the fifth iteration of ME.

Appendix D. Link Fusion MATLAB Code

D.1. Introduction

Two sets of MATLAB code are provided in the sections below, each annotated with comments. The two sets relate to:

1. The Matrix form of the implementation, using the standard matrix functions in MATLAB. This is presented for information only, as the method is impracticable for any realistically sized practical problem due to computer storage constraints.

2. The Element-wise form explained within the main text of this report, to implement the Link Fusion without needing to store the output covariance matrix, M.

Note that in the following pages, lines beginning with the % modifier in the source code denote the paragraph as a comment.

Note that in MATLAB, the * operator denotes matrix multiplication, the .* operator denotes element-wise multiplication.

Note that source code lines are ended with the terminator ;.

% The ij by 1 column vector D_{ij} is represented by `column_vector1`

% The ij by 1 column vector formed from the main diagonal of the covariance matrix $\Omega_{ij.ij}^{D_{ij}}$ is represented by `covariance_vector1`

% The a by 1 column vector V_a^{obs} is represented by `column_vector2`

% The a by 1 column vector formed from the main diagonal of the covariance matrix $\Omega_{a.a}^{V}$ is represented by `covariance_vector2`

% The ij by a matrix $\left(p^T\right)_{ij.a}$ is represented by `pija_transpose`

D.2. MATLAB Code for the Matrix Form

```
Function link_fusion_with_variance_matrix_form(column_vector1,
covariance_vector1, column_vector2, covariance_vector2, pija_transpose)
```

% For the full link fusion data set, ij = `zone_pairs` = 325*325 = 105,625 and a = `count_pairs` = 174

```
zone_pairs = size(column_vector1,1)
count_pairs = size(column_vector2,1)
```

% The a by a matrix $\Omega_{a.a}{}^{V} + p_{a.ij}.\Omega_{ij.ij}{}^{D_{ij}}.\left(p^{T}\right)_{ij.a}$ is represented by `quadratic_matrix`

% Note that here the function B = diag(A) assigned vector A of length L to the main diagonal of a zero square matrix B of dimension L by L. Therefore, `quadratic_matrix` has dimensions given by diag(a by 1) + (ij by a)' * diag(ij by 1) * (ij by a) = (a by a) + (a by ij) * (ij by ij) * (ij by a) = (a by a) + (a by a) = a by a

% Note that for the full link fusion data set, MATLAB has insufficient memory to compute `quadratic_matrix` using matrix form because it must store `diag(covariance_vector1)`, which has dimension ij by ij.= 11,156,640,625 8-byte floating point values = 89,253,125,000 bytes of storage.

```
quadratic_matrix = zeros(count_pairs);
quadratic_matrix = diag(covariance_vector2) + pija_transpose' *
diag(covariance_vector1) * pija_transpose;
```

% The ij by 1 column vector formed from the main diagonal of the inverse of the covariance matrix $\left(\Omega_{ij.ij}{}^{D_{ij}}\right)^{-1}$ is represented by `covariance_vector1_inv`

% The a by 1 column vector formed from the main diagonal of the inverse of the covariance matrix $\left(\Omega_{a.a}{}^{V}\right)^{-1}$ is represented by `covariance_vector2_inv`

```
covariance_vector1_inv = covariance_vector1;
covariance_vector2_inv = covariance_vector2;
for z = 1:zone_pairs
    if covariance_vector1(z) == 0
        covariance_vector1_inv(z) = 0;
    else
        covariance_vector1_inv(z) = 1 ./ covariance_vector1(z);
    end
end
for a = 1:count_pairs
    if covariance_vector2(a) == 0
        covariance_vector2_inv(a) = 0;
    else
        covariance_vector2_inv(a) = 1 ./ covariance_vector2(a);
    end
end
```

% The a by a matrix $\left(\Omega_{a.a}{}^{V} + p_{a.ij}.\Omega_{ij.ij}{}^{D_{ij}}.\left(p^{T}\right)_{ij.a}\right)^{-1}$ is represented by `quadratic_matrix_inv`, provided that `quadratic_matrix` is invertible (which it is for the full link fusion data set)

```
   quadratic_matrix_inv = inv(quadratic_matrix);
```

% The ij by 1 column vector $\left(\Omega_{ij.ij}{}^{D_{ij}}\right)^{-1}.D_{ij} + \left(p^T\right)_{ij.a}.\left(\Omega_{a.a}{}^{V}\right)^{-1}.V_a{}^{obs}$ is represented by `weighted_sum_vector`

% Note that `weighted_sum_vector` has dimensions given by (ij by 1) .* (ij by 1) + (ij by a) * ((a by 1) .* (a by 1)) = (ij by 1) + (ij by a) * (a by 1) = (ij by 1) + (ij by 1) = ij by 1

```
weighted_sum_vector = covariance_vector1_inv .* column_vector1 + pija_transpose
* (covariance_vector2_inv .* column_vector2);
```

% The ij by 1 column vector formed from the main diagonal of $M_{ij.ij}$ is represented by `fused_variance_column_vector`

% Note that here the inner functions B = diag(A) assign vector A of length L to the main diagonal of a zero square matrix B of dimension L by L, whereas the outer function D = diag(C) assigns the main diagonal of square matrix C of dimension L by L to a vector D of length L. Therefore, `fused_variance_column_vector` has dimensions given by diag(diag(ij by 1) – diag(ij by 1) * (ij by a) * (a by a) * (ij by a)' * diag(ij by 1)) = diag((ij by ij) - (ij by ij) * (ij by a) * (a by a) * (a by ij)) = diag((ij by ij) - (ij by ij)) = diag(ij by ij) = ij by 1

% Note that for the full link fusion data set, MATLAB has insufficient memory to compute `fused_variance_column_vector` using matrix form because it must store `diag(covariance_vector1)`, which has dimension ij by ij.= 11,156,640,625 8-byte floating point values = 89,253,125,000 bytes of storage.

```
fused_variance_column_vector = zeros(zone_pairs,1);
fused_variance_column_vector = diag(diag(covariance_vector1) -
diag(covariance_vector1) * pija_transpose * quadratic_matrix_inv *
pija_transpose' * diag(covariance_vector1));
```

% The ij by 1 column vector D'_{ij} is represented by `fused_proportion_column_vector`

% Note that here the function B = diag(A) assign vector A of length L to the main diagonal of a zero square matrix B of dimension L by L. Therefore, `fused_proportion_column_vector` has dimensions given by diag(ij by 1) * (ij by 1) = (ij by ij) * (ij by 1) = ij by 1

% Note that for the full link fusion data set, MATLAB has insufficient memory to compute `fused_proportion_column_vector` using matrix form because it must store `diag(fused_variance_column_vector)`, which has dimension ij by ij.= 11,156,640,625 8-byte floating point values = 89,253,125,000 bytes of storage.

```
fused_proportion_column_vector = zeros(zone_pairs,1);
fused_proportion_column_vector = diag(fused_variance_column_vector) *
weighted_sum_vector;
```

D.3. MATLAB Code for the Element-wise Form

```
Function link_fusion_with_variance_component_form(column_vector1,
covariance_vector1, column_vector2, covariance_vector2, pija_transpose)
```

% For the full link fusion data set, ij = `zone_pairs` = 325*325 = 105,625 and a = `count_pairs` = 174

```
zone_pairs = size(column_vector1,1)
count_pairs = size(column_vector2,1)
```

% The a by a matrix $\Omega_{a.a}{}^{V} + p_{a.ij}.\Omega_{ij.ij}{}^{D_{ij}}.\left(p^{T}\right)_{ij.a}$ is represented by `quadratic_matrix`

% Note that here `quadratic_matrix` is computed element-wise by using an outer loop ranging over each row x of `quadratic_matrix` from 1 to a, and an inner loop ranging over each column y of `quadratic_matrix` from 1 to a. For each row x of `quadratic_matrix` (i.e. each column x of `pija_transpose`), the element-wise multiplication of column x of $\left(p^{T}\right)_{ij.a}$ with the ij by 1 column vector formed by the main diagonal of $\Omega_{ij.ij}{}^{D_{ij}}$ is calculated to give an ij by 1 column vector V, and then for each column y of `quadratic_matrix` (i.e. each column y of `pija_transpose`), the matrix multiplication of the 1 by ij transpose of column vector V with column y of $\left(p^{T}\right)_{ij.a}$ is calculated to give element (x,y) of `quadratic_matrix`,with element x of the a by 1 column vector formed by the main diagonal of $\Omega_{a.a}{}^{V}$ added to the main diagonal of `quadratic_matrix` (i.e. where x = y). Therefore, since both x and y range between 1 and a `quadratic_matrix` has dimensions a by a

```
quadratic_matrix = zeros(count_pairs);
for x = 1:count_pairs
    for y = 1:count_pairs
        quadratic_matrix(x,y) = ((pija_transpose(:,x) .* covariance_vector1).')
* pija_transpose(:,y);
        if x == y
            quadratic_matrix(x,y) = quadratic_matrix(x,y) +
covariance_vector2(y);
        end
    end
end
```

% The ij by 1 column vector formed from the main diagonal of the inverse of the covariance matrix $\left(\Omega_{ij.ij}{}^{D_{ij}}\right)^{-1}$ is represented by `covariance_vector1_inv`

% The a by 1 column vector formed from the main diagonal of the inverse of the covariance matrix $\left(\Omega_{a.a}{}^{V}\right)^{-1}$ is represented by `covariance_vector2_inv`

```
covariance_vector1_inv = covariance_vector1;
covariance_vector2_inv = covariance_vector2;
for z = 1:zone_pairs
    if covariance_vector1(z) == 0
        covariance_vector1_inv(z) = 0;
    else
        covariance_vector1_inv(z) = 1 ./ covariance_vector1(z);
    end
end
for a = 1:count_pairs
    if covariance_vector2(a) == 0
```

```
        covariance_vector2_inv(a) = 0;
    else
        covariance_vector2_inv(a) = 1 ./ covariance_vector2(a);
    end
end
```

% The a by a matrix $\left(\Omega_{a.a}{}^{V} + p_{a.ij}.\Omega_{ij.ij}{}^{D_{ij}}.\left(p^{T}\right)_{ij.a}\right)^{-1}$ is represented by `quadratic_matrix_inv`, provided that `quadratic_matrix` is invertible (which it is for the full link fusion data set)

```
   quadratic_matrix_inv = inv(quadratic_matrix);
```

% The ij by 1 column vector $\left(\Omega_{ij.ij}{}^{D_{ij}}\right)^{-1}.D_{ij} + \left(p^{T}\right)_{ij.a}.\left(\Omega_{a.a}{}^{V}\right)^{-1}.V_{a}^{obs}$ is represented by `weighted_sum_vector`

% Note that `weighted_sum_vector` has dimensions given by (ij by 1) .* (ij by 1) + (ij by a) * ((a by 1) .* (a by 1)) = (ij by 1) + (ij by a) * (a by 1) = (ij by 1) + (ij by 1) = ij by 1

```
weighted_sum_vector = covariance_vector1_inv .* column_vector1 + pija_transpose
* (covariance_vector2_inv .* column_vector2);
```

% The ij by 1 column vector formed from the main diagonal of $M_{ij.ij}$ is represented by `fused_variance_column_vector`

% The ij by 1 column vector D'_{ij} is represented by `fused_proportion_column_vector`

```
fused_proportion_column_vector = zeros(zone_pairs,1);
fused_covariance_matrix_row = zeros(1,zone_pairs);
fused_variance_column_vector = zeros(zone_pairs,1);
fused_variance_count = 0;
fused_variance_maximum = 0;
fused_variance_average = 0;
fused_variance_variance = 0;
fused_covariance_count = 0;
fused_covariance_maximum = 0;
fused_covariance_average = 0;
fused_covariance_variance = 0;
```

% Note that here `fused_variance_column_vector` is computed element-wise by using an outer loop ranging over each element z of `fused_variance_column_vector` (i.e. each row z of `pija_transpose`) from 1 to ij, and an inner loop ranging over each element z_prime of `fused_variance_column_vector` (i.e. each column z_prime of the transpose of `pija_transpose`) from 1 to ij. For each pairing of z and z_prime, both ranging from 1 to ij, the matrix multiplication of row z of the ij by a matrix $\left(p^{T}\right)_{ij.a}$ with the a by a matrix $\left(\Omega_{a.a}{}^{V} + p_{a.ij}.\Omega_{ij.ij}{}^{D_{ij}}.\left(p^{T}\right)_{ij.a}\right)^{-1}$, and then with column z_prime of the transpose of ij by a matrix $\left(p^{T}\right)_{ij.a}$ (i.e. an a by ij matrix), is calculated to give `matrix_element` (dimension 1 by 1). Then for each pairing of z and z_prime, the element-wise multiplication of the zth element of the ij by 1 column vector $\Omega_{ij.ij}{}^{D_{ij}}$ with the negation of `matrix_element` and then with the z_primeth element of the ij by 1 column vector $\Omega_{ij.ij}{}^{D_{ij}}$ is calculated, with element z of the ij by 1 column vector formed by the main diagonal of $\Omega_{ij.ij}{}^{D_{ij}}$ then added where z = z_prime, to give the z_primeth element of

`fused_covariance_matrix_row`. The zth element of `fused_variance_column_vector` is given by the z_primeth element of `fused_covariance_matrix_row`. The zth element of `fused_proportion_column_vector` is given by the matrix multiplication of 1 by ij row vector `fused_covariance_matrix_row` with ij by 1 column vector `weighted_sum_vector`, which accounts for the off-diagonal covariance elements of $M_{ij.ij}$

```
for z = 1:zone_pairs
    for z_prime = 1:zone_pairs
        matrix_element = pija_transpose(z,:) * quadratic_matrix_inv *
(pija_transpose(z_prime,:).');
        fused_covariance_matrix_row(z_prime) = covariance_vector1(z).* (-
matrix_element) .* covariance_vector1(z_prime);
        if z_prime == z
            fused_covariance_matrix_row(z_prime) =
fused_covariance_matrix_row(z_prime) + covariance_vector1(z);
            fused_variance_column_vector(z) =
fused_covariance_matrix_row(z_prime);
        end
    end
    fused_proportion_column_vector(z) = fused_covariance_matrix_row *
weighted_sum_vector;
end
```

Appendix E. Guidance on the Processing of Count Data

Current WebTAG guidance (Unit M1.2) notes the following (para 3.3.5):

> "*Traffic counts may be obtained by automatic means (Automatic Traffic Counts, ATCs) or manually (Manual Classified Counts, MCCs). …In selecting the appropriate type of count …, these factors need to be considered:*
>
> *the accuracy of the data;*
>
> *the choice of survey locations;*
>
> *the need for information by vehicle type; and*
>
> *a recognition of the costs of these data*."

Based on paras 6.2.5 and 6.3.7 of the Traffic Appraisal Manual[70], accuracy figures are given (M1.2 para 3.3.32), but these relate to **measurement** only. Para 3.3.35 notes:

> "*It is normal practice for MCCs to be carried out on a single day but ATCs should be conducted for at least two full weeks. ATCs carried out for two-weeks or longer will capture some day to day variability.*"

Guidance on the factoring of observed count data such that it represents the particular time/day/year required by the traffic model is provided in WebTAG Unit M1.2, notably:

> *(para 3.3.38) "This section deals with the factoring of traffic count data from one base to another. Every factor has an associated reliability and the result of factoring is always to increase the confidence interval of the result. Factoring should therefore be kept to a minimum and the factor with the lowest coefficient of variation should always be chosen where a choice of factors is available."*
>
> *(3.3.40) "On some occasions more than one factor will have to be cumulatively applied in stages. DMRB volume 12, section 1, part 1, Appendix D13 gives the formulae and contains several worked examples on how this can be done and the resulting reliability of the combined results calculated."*

Further points are listed in the now superseded WebTAG 3.19 (para 4.3.2) as follows:

> *"A number of other adjustments are required, as follows:*
>
> *conversion from the day of survey to an average weekday;*
>
> *adjustment to account for diversions on the day that the interviews were conducted;*
>
> *adjustment to account for day to day variability; and*
>
> *conversion from the month of survey to a specific neutral month.*
>
> *Controlling to two-week ATCs will address the first and second of these adjustments and the third to an extent. The fourth adjustment will require factors to be developed from long-term ATCs. It should be noted that the application of each of these factors will add further uncertainty and widen the confidence intervals by unknown amounts."*

To these factors, a further factor to convert to the particular model year may also be required.

[70] DMRB Volume 12 Section 1 Part 1.

Acknowledgements

Thanks are due to John Bates Services, Denvil Coombe Practice, AECOM, Mott MacDonald, Katalysis, and Minnerva for their inputs into the project. The advice and support of Highways England's Project Sponsor, Roger Himlin was also much appreciated.

Andrew Merrall
Atkins
3100 Century Way
Thorpe Park
Leeds
LS15 8ZB