ORIGIN-DESTINATION ANALYSIS ON THE LONDON ORBITAL AUTOMATED NUMBER PLATE RECONGITION NETWORK

Charles Fox
Adaptive Behaviour Research Group, University of Sheffield, UK
Peter Billington
Telematics Technology LLP, Sheffield, UK
Dominic Paulo
Mouchel, London, UK
Clive Cooper
Highways Agency, Dorking, UK

ABSTRACT

The UK Highways Agency (HA) has a specific interest in understanding the actual journeys taken by drivers using the London orbital motorway (M25) in order to assist with planning and management of strategic routes. It is thought that the majority of traffic on Public Service Agreement (PSA) Routes, rather than travelling end to end is joining and leaving mid-route. We provide a tracking algorithm to detect journeys on the routes using data from the existing automated number plate recognition (ANPR) network. This network was designed for other purposes and in particular only covers subsets of lanes at sites, requiring inference of traffic behaviour in the non-covered lanes. Furthermore, its cameras translate plate numbers into non-unique hash values which give rise to many spurious matches. A full-lane calibration study was performed, and fused with induction-loop flow data to obtain parameters for such inferences.

Introduction

The UK Highways Agency (HA) has a specific interest in understanding the actual journeys taken by drivers using the London orbital motorway (M25) in order to assist with planning and management of strategic routes. It is thought that the majority of traffic on Public Service Agreement (PSA) Routes, rather than travelling end to end is joining and leaving mid-route, possibly travelling between only one or two junctions.

Motorway Incident Detection and Signalling (MIDAS) data is available which contains count, speed and occupancy information, the analysis of which provides an understanding of the stresses on individual links and the turning movements of traffic at specific junctions. However, this data does not readily provide useful estimates of where vehicles start and end their journeys on the M25.

The HA's goal is the construction of an Origin-Destination (O-D) matrix, in the form of a look-up table having all possible origins along one axis and all possible destinations along the other, each cell in the matrix giving the number or relative proportion of vehicles making the journey between the relevant orthogonal origin/destination pair. The provision of such a table would enable a greater

understanding of origin-destination movements and the operational implications of the real routes taken by traffic around the M25 and on relevant arterials.

The present ANPR infrastructure (National Traffic Control Center, NTCC) on the M25 and its arterials was designed for other purposes and in particular does not provide full lane coverage. Instead, single or pairs of cameras are positioned at sites covering subsets of the lanes. The proportion of traffic detected by these partial lanes varies according to the flow of traffic. For example, a camera in lane 1 of a three-lane motorway will detect most of the traffic during low flows, but only a third of the traffic when the motorway is saturated. Conversely, a pair of cameras on lanes 2 and 3 will detect a low proportion of traffic during low flows but a higher portion nearing saturation. A single camera in lane 2 may show a more complex flow/detection relationship.

The situation is further complicated by the hashing scheme used by the existing infrastructure. The existing cameras use a hashing scheme which represents each number plate by a 24-bit tag (recently upgraded from 18 bit tags). A 24 bit value provides a maximum of 16,777,216 possible tags to represent approximately 30 million vehicles in the UK, so it might be expected that tags matched between any two specific locations would have a high degree of uniqueness.

However, a process known as “character merging” is applied prior to generation of the hash value from the plate characters. In this process, character sets which are commonly mis-read by ANPR systems (such as O, D and 0, or 8, B and 3) are merged (e.g. all Ds are replaced by Os etc) which has the effect of reducing significantly the uniqueness of the hash value. Thus matches between tags at the origin and destination of a route must be processed to removed spurious matches due to hashing clashes (where different plates give rise to the same hash value), as well as to adjust for the flow-affected detection proportion.

In this study, we recorded one day of calibration data from 12 routes of different lane configurations, using higher quality cameras (Telematics Technology LLP, Sheffield, UK) which have a higher detection rate and report raw plate numbers rather than hashes. The calibration cameras were placed to cover all lanes of the calibration routes, and allow us to report how the proportion of traffic in each lane varies as a function of total flow as measured by MIDAS data. These multipliers are used in conjunction with a tracking algorithm to obtain estimates of origin-destination route profiles for a variety of day types.

Ontology

Each physical NTCC camera is positioned over one lane of a road, which may be a motorway, arterial or slip-road. Typically there are one, two or occasionally three cameras covering lanes of the same road, facing in the same direction. Such a group of cameras is called a site. A site thus defines a road having a number of lanes, a direction of travel, and a subset of those lanes which are covered by cameras. An outstation is a physical box which collects data from one or more sites (e.g. two sites facing in opposite directions) and relays the data to NTCC. A lane configuration describes which subset of lanes is covered for an *n*-lane road. The most common lane configurations for 2-lane roads in the NTCC network are lanes (1) and (1,2). The most common lane configurations for 3-lane roads are lanes (2),

(1,2) and (1,3). An origin-destination configuration is an ordered pair of lane configurations corresponding to on origin and a destination. A route consists of an original site and a destination site. A calibration route consists of origin and destination locations corresponding to sites but uses data collected from all lanes using Telematics calibration cameras rather than by NTCC site cameras. A profile is a graph showing the number of journeys on a route as a function of time of day.

A link is a short (100m) section of a road in a direction, and refers to all lanes of the road. Flow data (MIDAS) measures the number of vehicles per 5-minute interval at each link.

Each calendar day in the study was given a day type from 0-11. For example, 0 is working Mondays, 1 is working Tuesdays, 8 is holidays and 9 is the first day of school holidays. The present study finds profiles for 108 routes on and around the M25, using eight weeks (8Gb) of NTCC data.

The diamond-shaped icons in Figure 2 represent NTCC ANPR camera locations. The triple-segmented orange lines between camera locations in Figure 2 represent the routes over which profiles have been evaluated.

Matching algorithms

Both the calibration study and the 8Gb NTCC profiling require matching vehicles to be identified at the origin and destination. In the calibration study we have access to (noisy observations of) the plate numbers themselves, given by the Telematics cameras. In the profiling we have noisy 20-bit plate hashes. Both data sets are vulnerable to containing large numbers of false positive matches, which we filter using a journey time tracking algorithm.

Running for one route on one day's data, the algorithm begins by retrieving all candidate matches using a large SQL query, such shown in Figure 1. (The bulk plate log contains times, sites and hashes of ANPR detections).

```
SELECT spl_d.hashPlate,
    spl_o.plateTimestamp_GMT AS timeOrig,
    spl_d.plateTimestamp_GMT AS  timeDest,
       spl_o.siteId AS siteOrig,
       spl_d.siteId AS siteDest,
       spl_d.plateTimestamp_GMT-spl_o.plateTimestamp_GMT
    AS travelTimeSecs
FROM BulkPlateLog AS spl_d INNER JOIN
                  BulkPlateLog AS spl_o
                  ON spl_d.hashPlate = spl_o.hashPlate
WHERE spl_d.plateTimestamp_GMT>spl_o.plateTimestamp_GMT
AND             spl_o.siteId = '30320545'
AND             spl_d.siteId = '30320549'
AND spl_o.plateTimestamp_GMT > '2009-06-30 22:00:00'
AND spl_o.plateTimestamp_GMT < '2009-07-02'
AND spl_d.plateTimestamp_GMT > '2009-06-30 22:00:00'
AND spl_d.plateTimestamp_GMT < '2009-07-02'
ORDER BY  spl_d.plateTimestamp_GMT
```

Figure 1: SQL query used to recover all possible matches for a route on a date.

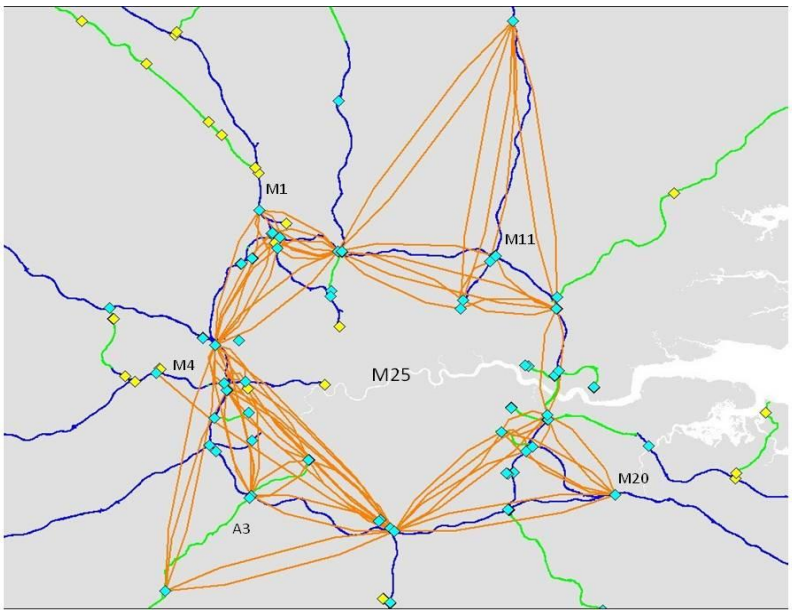


Fig.2 Map of the M25 London orbital, showing routes used.

Careful use is made of SQL clustered indexing to allow such queries to run in a few seconds over the 8Gb data set. The query returns all pairs of origin-destination plate detections having the same plate number or hash, which occur on the date (beginning at 22:00 the previous day), and where the destination detection occurs after the origin detection. This gives rise to a set of candidate journey durations, *y(t)* (where *t* ranges over the discrete set *T* of times at which matches occur). As the day progresses, the number and journey durations of spurious matches increases, as shown in Figure 3.

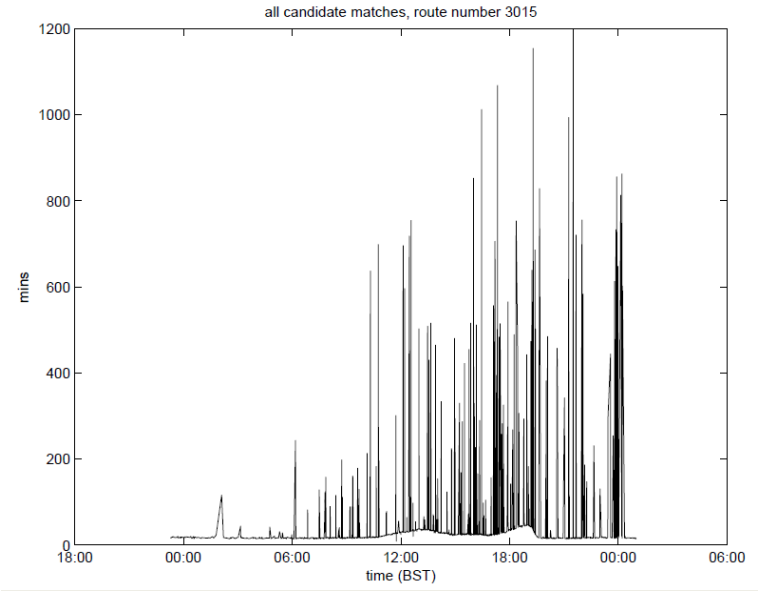


Figure 3: All matches, including many spurious ones, returned by the SQL query.

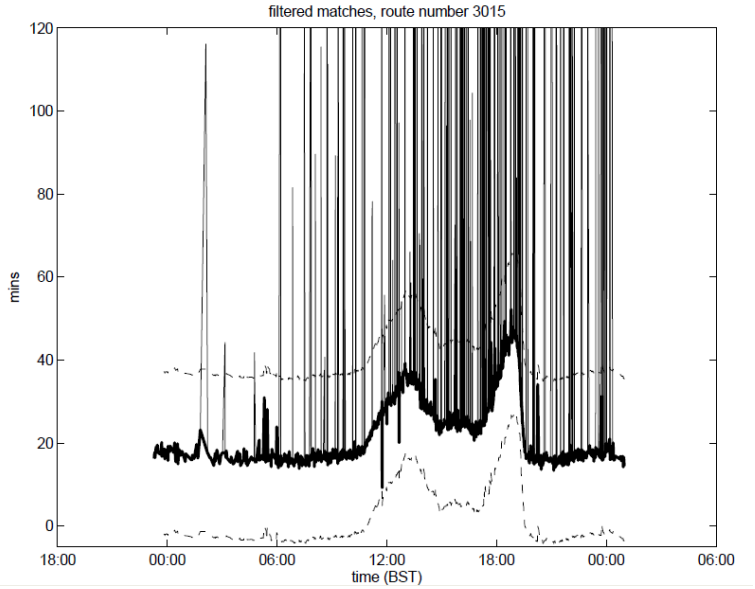


Figure 4: Accepted matches after filtering out the spurious ones.

A form of tracking is then used to sequentially filter out the spurious matches, whose output is shown in Figure 4. The thick line shows the accepted matches; dotted lines show the acceptance window. (Note change of *y*-axis scale from previous figure.)

The algorithm maintains an exponentially weighted moving average journey time throughout the day,

$$\hat{x}(t) = \lambda y(t') + (1-\lambda)\hat{x}(t-1), t \in T'$$

$$\hat{x}(t) = \hat{x}(t-1), t \notin T'$$

where *t'* ranges over all accepted match times, *T'*. This quantity is a point estimate of a hidden variable *x(t)*, the population duration at each match time, which is likely to change considerably during rush hours and quiet periods. The algorithm also maintains an acceptance window *σ(t)*. At each candidate match *y(t)*, we assume that

$$t \in T' \Leftarrow |y(t) - \hat{x}(t-1)| \leq \sigma(t)$$

This process may be viewed as a form of online EM algorithm (Dempster et al., 1977) which, for each point in sequence, alternates between making a hard classification (E step; assigning *t* to *T'* or $\bar{T}'$) and a parameter update (M step; updating $\hat{x}$).

As there are few spurious matches at the start of a day, we make an initial estimate of journey time at midnight from the mean of the first ten matches,

$$x(0) = \frac{1}{N}\sum_{1}^{10} y(t).$$

To estimate a useful value of σ, it is noted that the histogram of journey times typically consists of a relatively dense Gaussian-like region of good matches, followed by a much sparser and less populated region of high-duration spurious matches. Taking the median of this distribution gives a point somewhere in the Gaussian, and we take its value as an initial window width (so σ is half of this median).

To deal with unusual data cases (e.g. broken cameras, lane and road closures) we also impose hard maxima and minima on the tracked duration and window, $0 < \hat{x} < 180$mins and $15\text{mins} < \sigma < 30$mins.

Iterative lostness detection and re-tracking

The above algorithm proved sufficient for use on the calibration data, as it contained relatively few spurious matches due to its reports of raw plate numbers. The NTCC data, consisting of 24-bit hashes with character merging, gave rise to many more spurious matches which on occasion could result in the catastrophic loss of tracking, i.e.,

$$|\hat{x}(t) - x(t)| \gg \sigma(t).$$

as illustrated in Figure 5. During lostness, the acceptance corridor moves away from the bulk the candidate matches, resulting in a reduced density of matches over the day. Due to the hard limits imposed, tracking is always regained at some later stage in our data, so lostness has only been noted to occur during the rush-hour and similar rapid peaks, when the journey time increases faster than the size of the acceptance window.

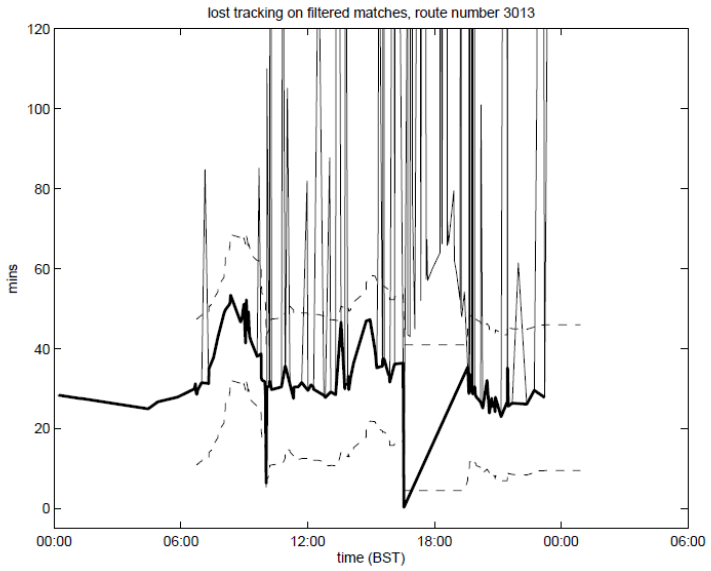


Figure 5: Loss of tracking.

To deal with this issue, up to five iterations of a lostness detection and restoration process are performed.

Assuming that lostness only occurs at peaks, and that peaks only occur between 7am and 7pm, we build a histogram of matches over this time period, having 2-hour bins. During lostness there are zero or very few (false) matches, so we test for bins in the histogram where the number of matches falls below ten per cent of its median. If this condition occurs, we re-run the tracking, using a widened acceptance window, $\sigma \leftarrow 2\sigma$, for the duration of the lost bin. In most cases this is sufficient to restore tracking (at the cost of accepting a small number of false positives); in more difficult cases it restores tracking for a period, gets lost again, and further iterations are needed to complete the tracking. Figure 6 shows an iteratively recovered version of Figure 5.

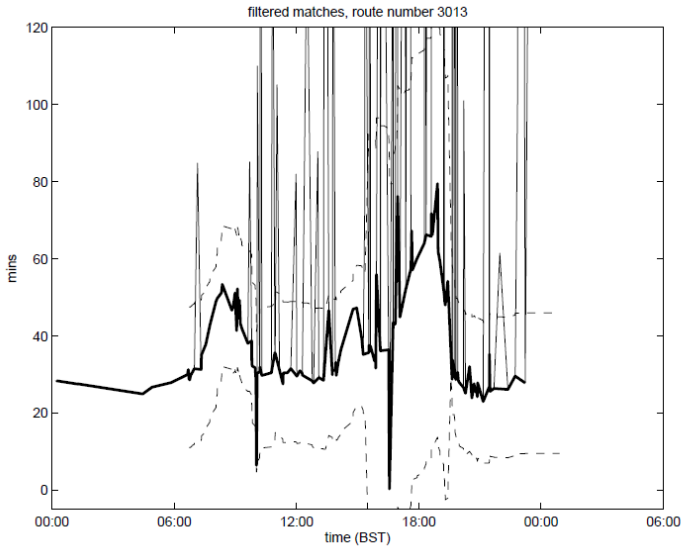


Figure 6: Recovery by detecting lostness then widening the acceptance window.

Calibration Data

The aim was to find how the proportion of routes detected by the NTCC cameras varies as a function of the lane configurations and of the flow. To do this, ANPR cameras were placed on all lanes of 12 calibration routes - selected to represent most typical routes - for one day. In particular these included motorway to motorway, dual carriageway to dual carriageway, dual carriageway to motorway and vice versa. (The calibration routes were all between Kingston, Guildford and the two M25 junctions near to them.)

The matches found by the subset of cameras in each origin-destination configuration of each calibration route were compared to the matches found using all cameras in that route. This enables estimation of the proportion of detections given by each origin-destination configuration, as a function of MIDAS flow. For example, one origin-destination configuration is the case of an origin on a 3-lane motorway with cameras on lanes 1 and 3 and a destination on a dual carriageway with a camera in lane 1. As MIDAS flow (measured at the origin at the journey start time) changes, so does the proportion of journeys along this route that are picked up by the particular cameras in the configuration.

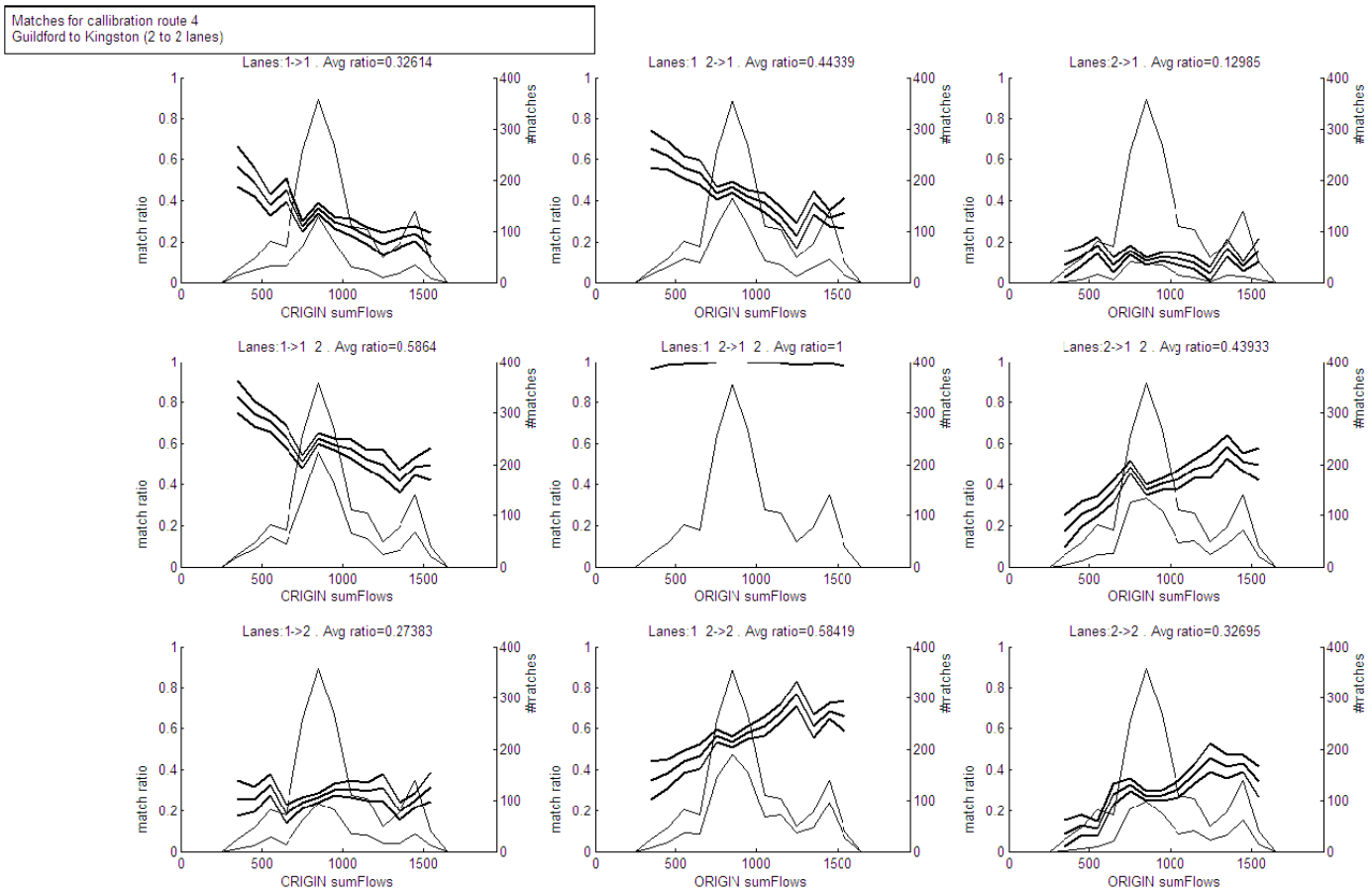


Figure 7: Detection ratios by flow for a single calibration route.

Figure 7 shows the complete set of origin-destination configurations for the simplest case, a dual-carriageway to dual-carriageway route. The thick lines show mean belief and one standard deviation confidence intervals. The thin lines show the number of matches in each flow bin, which affect the confidence intervals. There are nine configurations as both the origin and destination may have cameras in lane one, two or both. The plots show the number of matches detected by all cameras (upper thin plot) and by the configuration cameras (lower thin plot) as a function of the MIDAS flow, in a histogram with bins of 100 vehicles per hour. They also show an estimate of the ratio between these matches, with error bounds (thick lines).

For a single calibration route on a single day, the ratio θ is estimated as follows. We assume that the flow bins are mutually independent, and that each flow bin has a Beta distribution over our belief in its ratio,

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}.$$

We begin with a flat (Gelman et al., 2004) prior, ($\alpha = 1$, $\beta = 1$) over its matching ratio θ. We assume that the total (all lanes) number of matches that the number of matches observed is an exact estimate of the population of matches at that flow, N, and that the number of matches observed by the configuration cameras, k is a sample from the number that could have been observed. Under these assumptions, k is drawn from a Binomial distribution with unknown parameter θ. i.e. the situation is identical to estimating the heads-probability θ of a weighted coin, from N

observations, k of which are heads. On observing N and k, the Beta posterior is updated by:

$$\alpha \leftarrow \alpha + k,$$
$$\beta \leftarrow \beta + N - k.$$

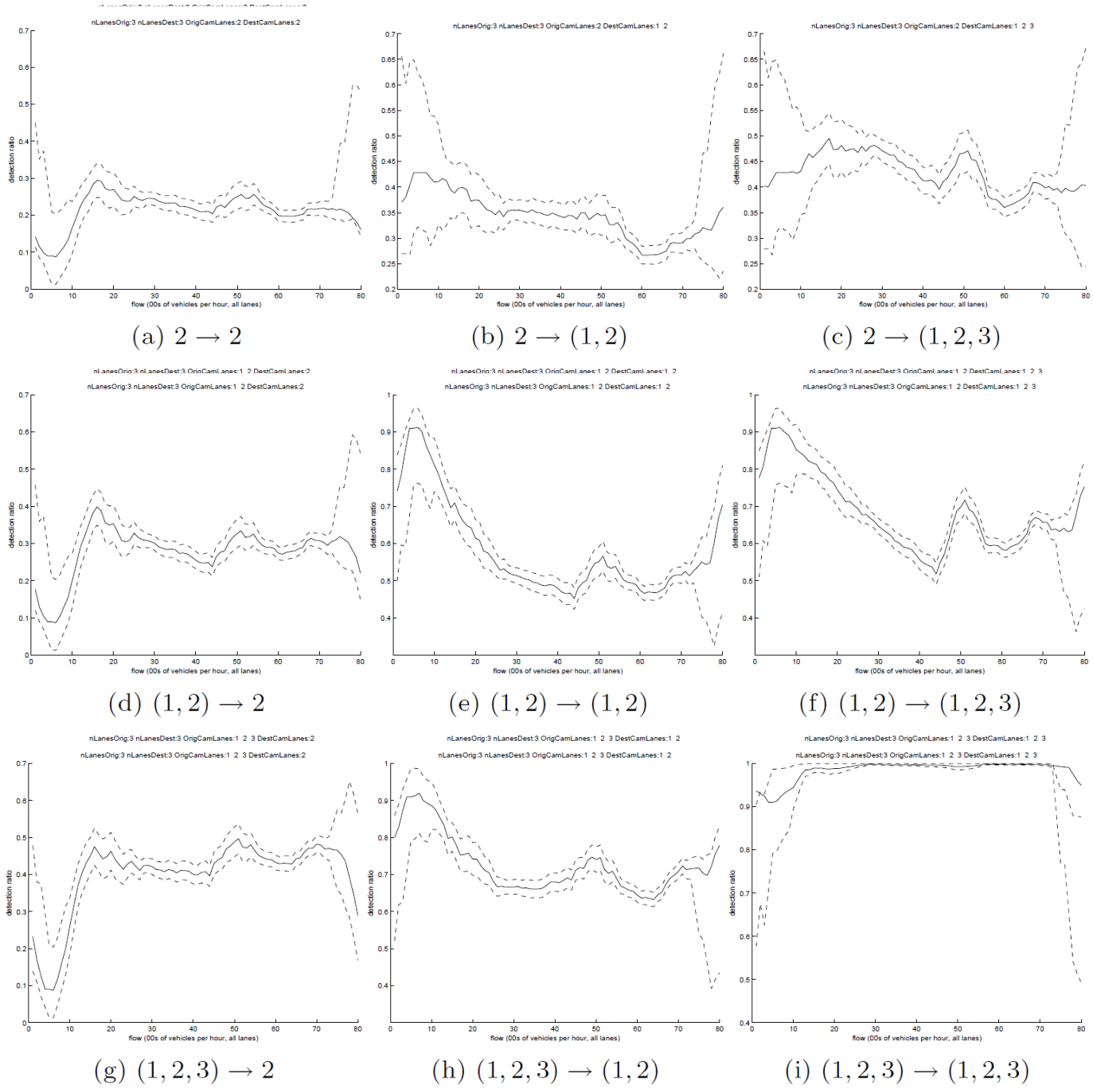
Furthermore, information may be fused from all routes sharing configurations into a single posterior for each configuration and each flow bin, using the same update equation. For example, two routes between three-lane motorways with cameras at lanes 1 and 2 at the origin and lane 2 at the destination. Thus for each origin-destination configuration, we may infer the match ratio as a function of flow. Such inferred functions are presented in Figure 8 for the most common NTCC configurations between 3-lane motorways, and form the principal empirical result of this exercise. As well as providing calibration for the subsequent analysis of NTCC data, they also contain information about how traffic adapts across lanes according to the overall flow. In addition to the ratios in Figure 8, many more ratios have been computed for routes between pairs of roads of different lane numbers and camera configurations, using the same method.

Bulk Data

The reason for estimating match ratios is to provide corrections to aid in estimating the number of origin-destinations routes within the existing NTCC ANPR network. This network consists of 24-bit plate-hashing cameras, positioned at sites in lane configurations. The calibration study suggests how to correct the number of matches on routes, from the sample of matches detected by the existing lane configurations, to form estimates of the actual number of journeys made. The iterative matching algorithm described previously was used to filter the matches made between the origin and destination lane configurations, and the number of journeys at each 5 minute time bin throughout the day was estimated by applying the following adjustments. First, the number of matches in each time bin was divided by the calibration match ratio for origin flow at the start of the journey. Second, it was multiplied by a constant factor of (1/0.75) to account for the ANPR detection rate of individual cameras. (This factor was found to vary somewhat as a function of MIDAS flow, see Figure 9, due to the cameras' processing becoming overloaded by high traffic volumes. However this has been ignored in the present study.) Finally, some NTCC sites monitor roads having $N>3$ lanes, for which we have no calibration data. These cases are treated as three-lane roads, and then their number of matches is crudely multiplied by $N/3$.

Profiles

ANPR data were made available by NTCC for 108 routes, for the eight weeks commencing 22 June, 2009, a dataset some 8Gb in size. The matching and filtering algorithm required about three hours to process this data on an Intel Centrino Duo machine, with the majority of time being spent on inserting the results from Matlab into SQL server via ODBC.



(a) $2 \rightarrow 2$ (b) $2 \rightarrow (1,2)$ (c) $2 \rightarrow (1,2,3)$

(d) $(1,2) \rightarrow 2$ (e) $(1,2) \rightarrow (1,2)$ (f) $(1,2) \rightarrow (1,2,3)$

(g) $(1,2,3) \rightarrow 2$ (h) $(1,2,3) \rightarrow (1,2)$ (i) $(1,2,3) \rightarrow (1,2,3)$

Figure 8: Journey detection rates for common lane configuration pairs on three-lane motorway to three-lane motorway routes.

The final objective was to develop origin-destination route profiles for various day types, so journeys from multiple days of the same type have been averaged out, smoothed with a 15-point convolution, and their number over time of day tabulated for subsequent use with the interactive visualisation tool. Figure 10 shows example route profiles which have been produced by this method, and which show rush hour peaks in the morning and evening.

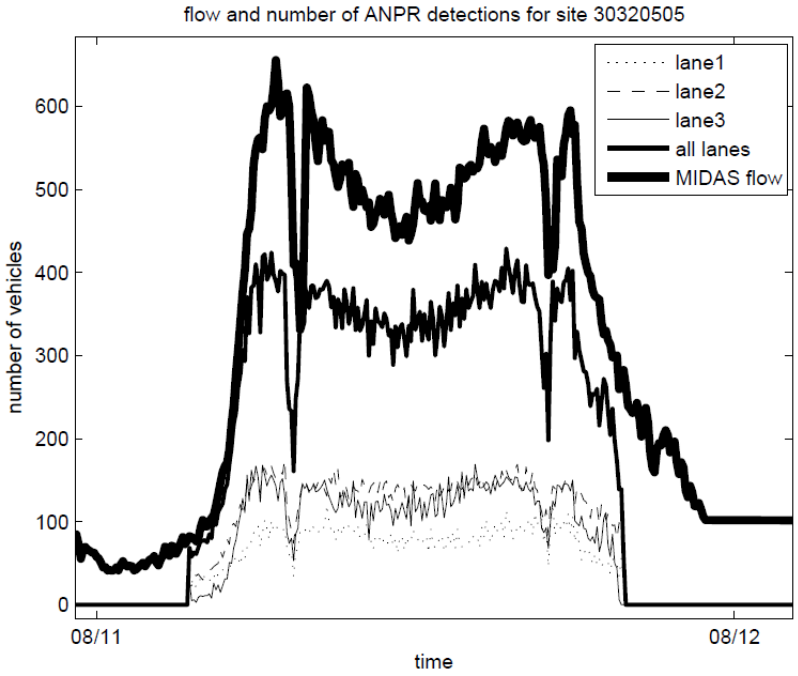


Figure 9: Number of ANPR detections in each lane, and all lanes, compared against MIDAS flow data.

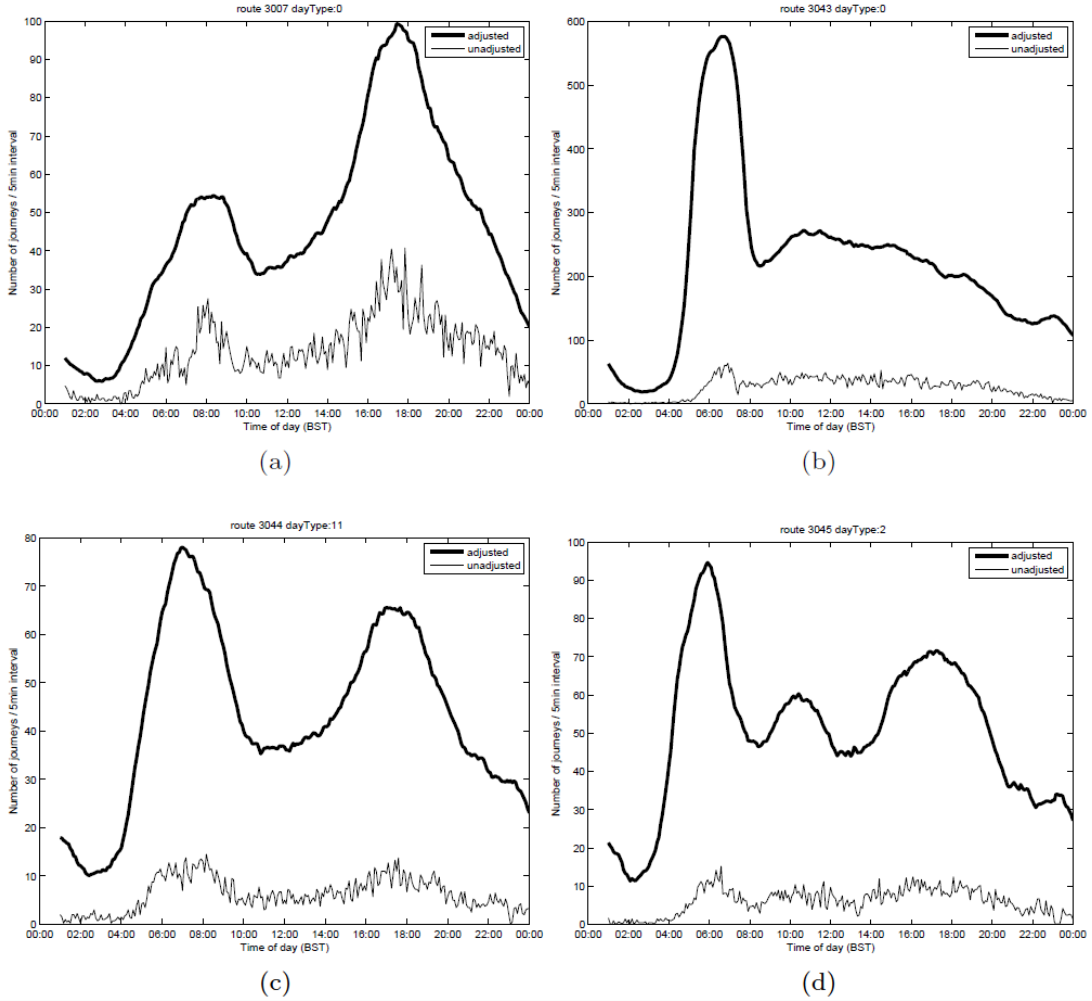


Figure 10: Typical origin-destination profiles.

The unadjusted data shown in these graphs represents the actual matches detected by the NTCC cameras monitoring a subset of lanes. The adjusted data is the result of applying the calibration characteristics for the appropriate start and end lane configurations which have been indexed according to the flow at the appropriate time of day. Such flows have been found from flow profile data supplied by the NTCC for the network links on which the cameras at the origins of each route are situated. The adjusted data therefore represents an estimate of the actual number of O-D movements which has been derived from a combination of NTCC camera observations, flow profiles and calibration characteristics.

Other than where NTCC tag data was not available (for example due to ANPR sites being out of service) a complete set of profiles has been produced for each O-D route and each day type. A graphical user interface for selecting and viewing the profiles has been constructed and is illustrated in Figure 11.

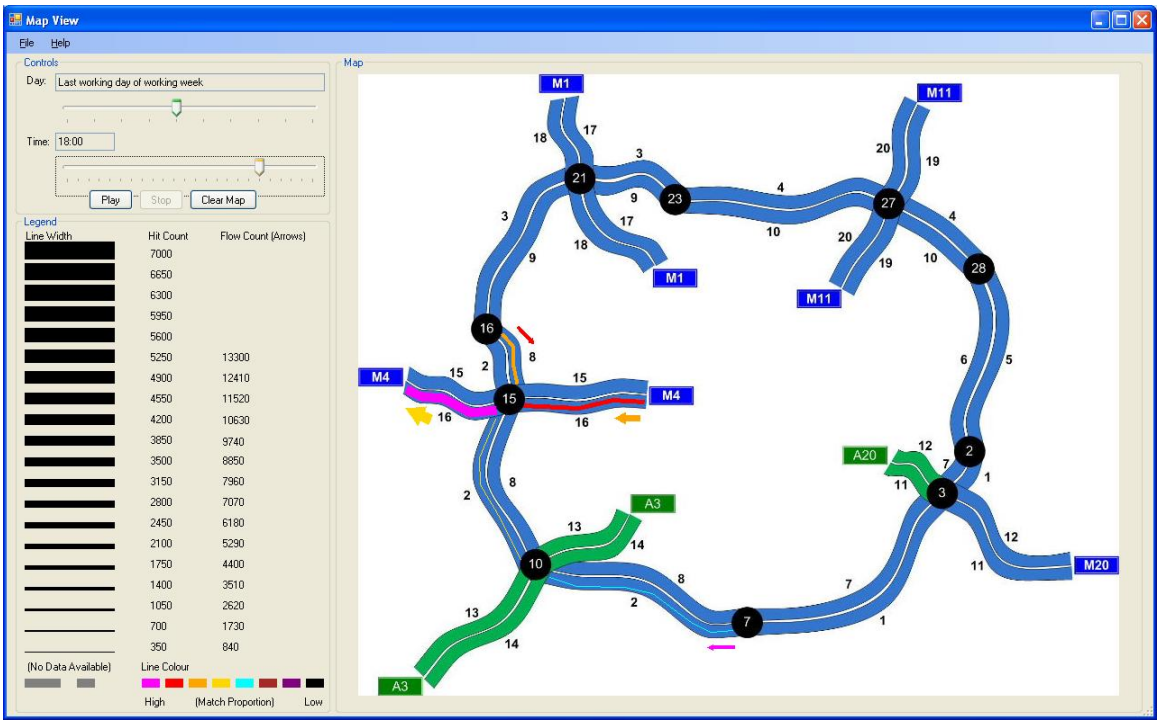


Figure 11: Screenshot from O-D Profile Graphical Visualisation tool.

Discussion

We have presented methods for rapid (3 hours to process 8Gb of data) matching and filtering of noisy ANPR origin-destination data, and for calibrating partial lane configuration matches, as a function of MIDAS traffic flow, to estimate total journey number profiles for routes over time and type of days.

The methods could be further improved by more detailed data modelling. Calibration data could be obtained for roads with four and more lanes; and anomalies such as

lane closures and hard-shoulder running could be accounted for. The assumption of independence of match ratios in neighbouring flow bins was pessimistic, though heuristically smoothed out later in the profile process. This combination of independence and smoothing could perhaps be replaced by (or shown to be equivalent to) a Gaussian Process (Rasmussen and Williams, 2005) which explicitly models local correlations. Finally we noted an effect of MIDAS flow on the detection rates of individual cameras which could be taken into account.

The output profiles could for example be used by the Highways Agency to identify particular popular routes between two arteries of the M25 and hence make recommendations about constructing new direct links (roads or public transport) between their origins and destinations, reducing congestion on the orbital.

References

A.P.Dempster, N.M. Laird, and D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38, 1977.

A. Gelman, J. B. Carlin, H. S. Stern and D. R. Rubin. *Bayesian Data Analysis,* CRC Press, 2004.

C.E. Rasmussen and C.Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

Acknowledgements

The co-operation and assistance of the National Traffic Control Centre in the provision of bulk data is gratefully acknowledged.