Clustering Trips Identified from Automatic Number Plate Recognition Camera Scans

Pedro M. Pinto Silva * Matthew Forshaw* A. Stephen McGough*

Abstract

In major cities, government agencies increasingly employ automatic number-plate recognition (ANPR) technology in law enforcement and traffic control. In the Tyne and Wear region (UK) the network of ANPR cameras is used to monitor travel times across sensitive roads. However, few works have explored the full potential of number-plate scans for analysing individual and collective travel patterns. In this work we present a methodology for deriving trips from vehicle sightings at fixed camera locations. We illustrate the effect of parameters τ and T on trip discrimination and on the detection of implausible trips. To demonstrate the potential of trip data we apply *k-means* clustering to trips identified from over 40 million plate scans recorded over fifteen weekdays. Results show that whilst private and transit travel modes can begin to be inferred from the resulting clusters, further work needs to be put into developing a more consistent and integrated framework for trip identification in ANPR data.

1 Introduction

The volume of traffic on our roads has been growing steadily for over 25 years, both in terms of the number of vehicles on the road – increasing by 40.6% in the UK [5] – and the distances covered – 325.5 billion miles driven in the UK in the year ending September 2017 which is up nearly 30% in the last 25 years [6]. This is placing ever more burden on the road infrastructure along with those who police and manage it. In order to better understand how we can deal with this increase in demand we need to better understand how the road network is being used. By understanding road usage we can better deal with congestion, handle traffic incidents, plan road modifications and deal with illegal acts.

In a utopian model we would have full disclosure of all journeys made by all vehicles on the road infrastructure. However, this has numerous ethical and technical issues. From an ethical standpoint should we be allowed to know where all vehicles are at any given point in time. From a technical point of view, although every vehicle could be fitted with a GPS tracker – costly in its own right – there would still exist the issue of how we would collect and stream all of this data for future processing. Alternatively one can view the problem the other way around and rather than tracking individual vehicles look at collecting information by observing vehicles passing points within the road networks. A prime example of this approach are Automatic Number-plate Recognition (ANPR) cameras. These cameras are a combination of digital camera coupled with Artificial Intelligence to identify number-plates within the image and convert these into strings of characters. ANPR cameras are normally fixed in location[1] able to view all vehicles passing that location.

For ANPR the problem now becomes that of recovering as much information about vehicle's journey as possible from the limited number of observations. ANPR cameras are normally located on major roads and interchanges, however, this only covers a tiny fraction of the road network. We can, though, estimate routes between cameras by understanding the distances between cameras and the most "sensible" routes between them. This allows us, given a set of ANPR sightings of the same vehicle, to produce a "most likely" route for that journey. It should be noted that we cannot determine the actual start and end of the journey as these will happen in areas not covered by ANPR. It should also be noted that for ethical reasons it is not normal to obtain actual number-plates, but rather the hash of these. Though, for most situations this will suffice.

Once we have a set of sightings of a vehicle using ANPR, we now need to convert these into actual journeys. The first requirement is to identify individual journeys. Although this can't be done with certainty we can apply general rules to distinguish one journey from the next. For example if two sightings are made from ANPR cameras which are connectable by a "sensible" route[2]

*School of Computing, Newcastle University, United Kingdom

[1] Although cameras can be in a vehicle and moved from location to location.

[2] Here "sensible" implies that a route between cameras A and B would not need to go through a third camera C.

in a time interval which is "sensible" then these can be determined to be part of the same journey. However, if the timings between two sightings is significantly longer than what would be expected then this would imply that the vehicle stopped between these two cameras and that the later sighting is part of a new journey.

The process of journey identification needs to be performed on dirty data which contains numerous impurities which need to be handled. These include:

- **Number-plate miss-reads:** Although ANPR cameras have accuracies of around 99, miss-reads are possible. This can lead to sightings being missed or vehicles being wrongly sighted in locations.
- **Timing errors:** The time-stamps of sightings could be erroneous. The minor side of this is implausible journey times, though, more seriously, this can lead to reordering the set of cameras on a particular journey.
- **Clones number-plates:** For various reasons a number-plate may be cloned and used on a different vehicle. This can lead to impossible journeys and journeys that the real vehicle did not make.

Once journeys have been identified from the sightings we can then progress by using these journeys to identify higher-order issues within the road network. In this paper we demonstrate how we can use this journey information in order to identify the most likely class each vehicle is a member of. By clustering over such characteristics as how many journeys are made each day, average length of journeys, the number of different ANPR cameras seen in a day and the times when journeys are made we can cluster vehicles into buses, taxis, commuters and delivery vehicles.

The rest of this paper is presented as follows. In Section 2 we discuss related work. Section 3 we presents the ANPR data for the Newcastle area. Our process for identifying individual journeys is presented in Section 4 while Section 5 presents our classification approach. We present results in Section 6 before offering conclusions and future directions in Section 7.

2 Related Work

The main use of ANPR data for the UTMC is estimating average journey times for selected or sensitive links in the road network. Furthermore, several authors have extensively researched how to use number-plate data as an extension to link counts for estimating origin-destination matrices and link flows [2, 3, 9].

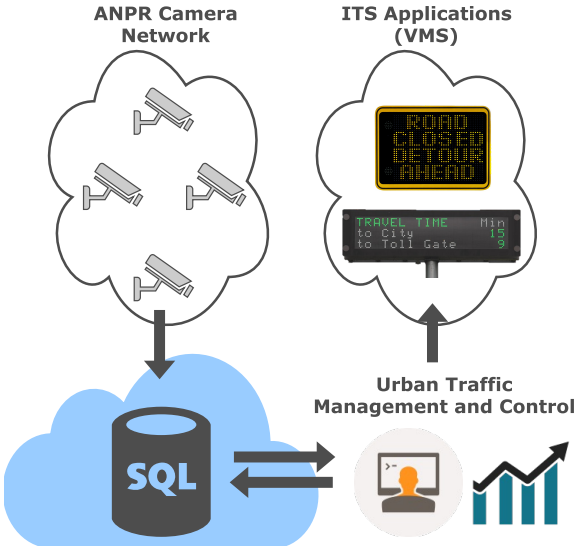


Figure 1: Overview of an ANPR system.

However, very few works have focused on analysing individual or collective travel patterns from number-plate data, particularly across extended periods of time. Moreover, there is no consistent conceptual and analytical framework for transforming number plate data into a historical sequence of trips for each vehicle. Finally, we believe that trip data, properly identified from number-plate data, has the potential to unlock a number of new applications for urban traffic control and law enforcement. Thus, in section 4 we present a conceptual methodology for grouping multiple camera observations of the same vehicle into one or several trips of that vehicle.

Determining the distribution of travel modes is one of the fundamental steps in the four-stage model: an essential traffic modelling methodology for transportation planning [8]. Previous works have used trip information derived from different sources of data to identify travel mode or purpose of trip. More notably, survey data, floating car data and mobile phone data have been used [1, 10]. Although number plate data has been used in [4], to identify different categories of trips, the authors do not differentiate between private or public travel modes and focus instead on categorising trips by time of occurrence. Hence, in section 5 we apply the *k-means* clustering algorithm to derived trip data and based on the results, we discuss the limitations of proposed methods and that of ANPR data.

3 Tyne and Wear ANPR Data

Automatic number-plate recognition (ANPR) cameras are actively employed in urban traffic environments

Vehicle	Camera	Timestamp	Clock Error	Confidence
169239	1031	1454284800.26	0	100
12862943	18	1454284800.97	8	61
16243894	22	1454284801.46	6	86
4817789	52	1454284803.43	13	94
5503486	110	1454284802.19	22	91
15244177	115	1454284802.83	18	87

Table 1: Sample of number plate data. Clock error is given in milliseconds and confidence as a percentile value.

and play an important role in day-to-day intelligent transportation systems. They can be used by government subsidised entities in urban traffic management and control; by commissioned highway agencies in electronic toll collection; or by law enforcement organisations in detecting speeding vehicles and validating number plate registrations. The wide diversity of applications, paired with the large improvements in price-to-performance ratios of ANPR hardware and software systems, has resulted in increased investments of ANPR cameras for urban environments [7, 12].

In the region of Tyne and Wear, United Kingdom, there are over 250 active ANPR cameras. Over 1 million license plate detections are recorded by these cameras every day. Figure 2 shows the number of daily scans recorded over a month (February, 2017). Furthermore, every scan is stored in a central database managed by Urban Traffic Management and Control (UTMC) Tyne and Wear, and used to compute travel times across particular links of interest in the road network. These are usually major roads that see high volumes of traffic, or road segments more prone to traffic jams. Average journey times can then be conveyed back to the drivers by the way of Variable Message Signs (VMS) or web based applications. Figure 1 represents this interaction.

Number plate data, in its essence, is a stream of events, each representing a vehicle observed by one camera at a specific point in time. An excerpt of the data can be found in table 1. All number-plates were anonymised by the UTMC through a hashing algorithm before the data was shared. Cameras are uniquely identified by an integer and timestamps are relative to each camera's clock. However, the cameras are connected through a private network which provides clock synchronisation using the Network Time Protocol (NTP). Therefore, the timestamps can be used directly if the synchronisation error is negligible. The following additional information is also captured and provided by each camera: (i) the clock synchronisation error (milliseconds); (ii) the camera's confidence that the identified number plate is the true number plate (percentage); (iii) the direction of travel, away or towards the camera. The confidence in the observation is especially useful as it helps diagnosing license plate recognition errors. On the other hand, the direction of travel is dependent upon the orientation of the camera, which is not provided. Hence, we chose to ignore the latter in this work, and aim to explore this information in future works.

4 Trip Identification

In this section we explain our technique for identifying which sets of sightings can be combined in order to make a trip.

Let the i_{th} sighting of vehicle k be defined as the unordered pair:

$$s_i^k = \{c, t\}, \tag{4.1}$$

where c uniquely identifies a camera, and t is a scalar representing a point in time (e.g. a timestamp).

Let an ordered sequence of sightings of vehicle k define the u_{th} trip of vehicle k:

$$w_u^k = \left(s_{(1)}^k, s_{(2)}^k, \ldots, s_{(n)}^k\right), \tag{4.2}$$

where n is the length of the trip, i.e. the number of sightings. Moreover, let the corresponding journey time sequence, of length $n-1$, be defined as the time difference of consecutive sightings:

$$jt_u^k = \left(t_{(2)}^k - t_{(1)}^k, \ldots, t_{(n)}^k - t_{(n-1)}^k\right). \tag{4.3}$$

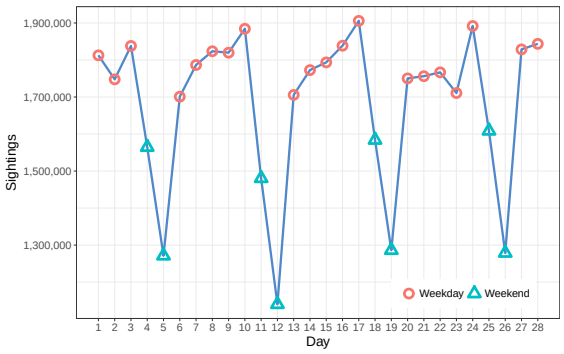
We consider a trip of w_u^k valid if the following conditions hold:

$$n \geq 1, \tag{4.4}$$

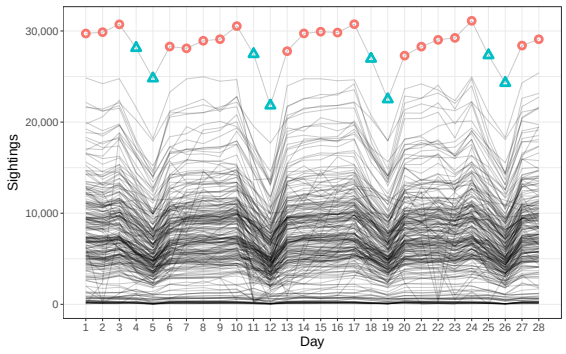
$$\tau_{(i)} < jt_{u(i)}^k < \mathrm{T}_{(i)} \ , \ \forall i \in jt_{u(i)}^k, \tag{4.5}$$

where $\tau_{(i)}$ and $\mathrm{T}_{(i)}$ are the lower and upper bound of the i_{th} element of the journey time sequence.

The first condition 4.4 is straightforward and specifies that every identified trip should have at least one sighting. Obviously, vehicles can make trips that do not pass through any ANPR cameras and thus have no associated sightings: $n = 0$. However, this work focus on trips that we can observe and hence we consider that $n > 0$. The second condition 4.5 defines a minimum and maximum travel times between consecutive observations. Its purpose is twofold: (i) first, to allow separate trips



(a) Total number of scans recorded per day in Tyne and Wear. There is a clear seasonal effect caused by decreasing traffic demands at weekends and increasing traffic volume during weekdays.



(b) Number of scans recorded per ANPR camera and day in Tyne and Wear. Inter-camera variability is observed, as some cameras are located in more traffic intensive road sections than others. Decommissioned or temporarily unavailable cameras (due to loss of power, faulty camera, road closed, etc) are depicted at the bottom of the graph.

Figure 2: License plate scans recorded by ANPR cameras during February 2017, in the region of Tyne and Wear, United Kingdom.

made by the same vehicle to be differentiated. For instance, given two consecutive sightings of k three hours apart, we would want to interpret them as belonging to different trips of k; (ii) second, it allows implausible trips to be identified. For example, an implausible trip can result from observing k at a given camera and then a few seconds later at a second camera, several miles apart. Two explanations are common, either one of the cameras made a detection error, or there is another vehicle with a cloned number-plate travelling on the road network. Evidently, condition 4.5 is only valid for trips of length two sightings or greater. Nevertheless, trips can easily be differentiated by first sorting sightings by time of occurrence, then calculating the journey time sequence for the entire sequence and finally comparing each element against T. An example of a trip identified this way can be seen in Figure 4.

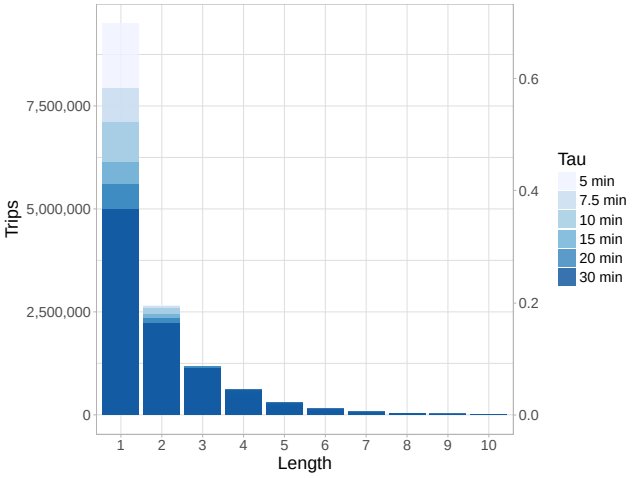


Figure 3: Distribution of trips per length of trip.

The simplest approach to choosing the value of T is to pick a fixed empirical value, such as 5 or 10 minutes. However, if the distance between two cameras is greater than another pair of cameras, then it makes sense that T is relaxed. Similarly, if there is an anomaly in the road network, such as a traffic jam, and the routes connecting the two cameras are affected, then the value of T should also be adapted. Hence, T should be a function of the distance between the two cameras (or, more accurately, of the top n-routes between these) and the distribution of observed journey times. The same rationale can be applied to τ.

On the other hand, a consistent set of models and methodologies for estimating T and τ, under a variety of circumstances and road anomalies, remains an open research problem. Thus, for the purposes of this work, we will consider these parameters to be fixed. Nevertheless, the importance of estimating these parameters should not be understated. Any errors in trip identification that occur due to poor estimation and filtering methods will propagate and will be amplified in posterior analysis done using trip data. Therefore, T and τ estimation is something that needs to be carefully addressed and studied in the future in order to fully understand its impact in the outcome of the research. To briefly exemplify this, figure 3 shows how the number of trips

Vehicle	Camera	Timestamp	Trip	Sighting	Journey Time	Trip Id
2362920	1014	2017-02-01 00:00:06	1	1	NA	21
2362920	1044	2017-02-01 00:01:28	1	2	82.38	21
2362920	35	2017-02-01 00:02:32	1	3	63.50	21
2362920	32	2017-02-01 00:04:38	1	4	125.95	21

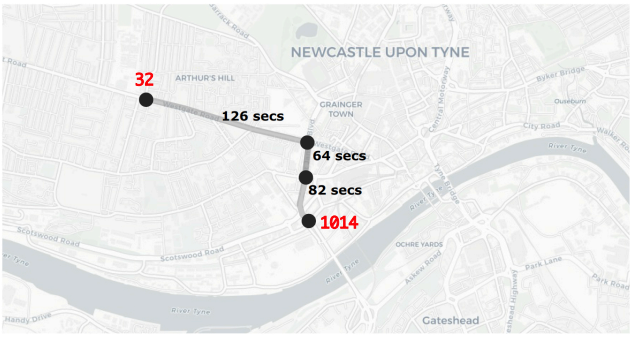


Figure 4: Example of a trip of length 4. On the left side the corresponding data table is shown. The u_{th} trip of each vehicle is given by the variable *Trip*, whereas the i_{th} sighting is given by variable *Sighting*. The variable *JourneyTime* gives the travel time from the previous to current sighting. Lastly, the variable *TripId* represents the unique sequence of cameras that describes the trip. This allows trips to be grouped and summarised not only in terms of their origins and destination, but also routes. On the right side, the same trip is plotted on a map. However, the lines do not represent the true route taken by the vehicle, but instead the fastest driving route between sightings. Even though no routing information is available for each consecutive pair of sightings, the observed journey times can be compared against the distribution of collective journey times to rank the set of most likely routes chosen (which can be determined for instance by Stochastic User Equilibrium [3]).

per length of trip varies by fixing T at different empirical values (τ is ignored).

4.1 Duplicate scannings We need to ensure that every trip of vehicle k is unique from all other trips of vehicle k. I.e. given W^k the set of all valid trips of k:

$$W^k = \left(w^k_{(1)}, w^k_{(2)}, \ldots, w^k_{(N)}\right), \tag{4.6}$$

where N is the number of trips of k, then there should be no two trips containing the same sighting:

$$\begin{aligned} s^k_{(u,i)} \neq s^k_{(v,j)}, \forall u, v = 1, 2, \ldots, N \ , \ u \neq v, \\ \forall i, j = 1, 2, \ldots, n^u \ , \\ \forall i, j = 1, 2, \ldots, n^v \ , \end{aligned} \tag{4.7}$$

where $s^k_{(u,i)}$ is the i^{th} sighting of vehicle k during trip u, and n^u is the number of sightings in trip u.

However, ANPR can identify the same vehicle multiple times in the same run, if for instance the vehicle is stopped at a junction or traffic light. Hence, if two sightings occurred at the same location in a very short period of time, then there is a strong possibility that these are duplicate observations. As a simplification, we can assume that a trip should not contain cycles and that no camera should appear twice in the same trip. Yet, this assumption ignores cases where a vehicle is required to correct its route by passing through the same location as one previously observed in the same trip. Thus, we affirm that two sightings of vehicle k are different if they were observed at two different points in time at different locations, or, if observed at the same location, then the time interval must be greater than a parameter γ, otherwise the two sightings are deemed as duplicates:

$$\begin{aligned} &(\ c^k_i \neq c^k_j \ ) \vee \\ &(\ c^k_i = c^k_j \wedge |t^k_i - t^k_j| < \gamma \ ) \ \Rightarrow s^k_i \neq s^k_j \ , \ i \neq j \end{aligned} \tag{4.8}$$

where c^k_i and t^k_i are the camera and timestamp of the i_{th} sighting of vehicle k.

Although the estimation of γ carries similar considerations and consequences as those of estimating T and τ, most duplicates can be identified in consecutive sightings of the same camera within the same trip. Even though a poor estimation of γ also has an impact on error propagation, this decreases substantially after filtering duplicates according to the heuristic above, due to the low occurrence of cycles in trips. Nevertheless, this issue needs to be fully investigated in future works.

4.2 Errors in plate scanning Number plate recognition cameras have accuracy rates of 99.9% or higher. If we consider that on average 1 to 10 out of 10000 scans are misclassified number-plates then approximately 200 to 2000 scans everyday are incorrect. We categorise the errors made by ANPR cameras into two types: (i) a passing vehicle is not detected by the camera; (ii) a different vehicle is detected instead. In the first case, there is no observed data. The corresponding trip sequence for the affected vehicle will be missing a sighting but we will have no indication of this. It should be noted that this could also happen in the case where a vehicle does not take the 'normal' route and pass this intermediary camera. For the second case, there is a recorded sighting, but this will be assigned to the incorrect vehicle. Even though, the camera provides a value of *confidence* which helps in diagnosing incorrect scans, errors still occur for high values of confidence. Even so, in this work we do not present a complex solution for this issue. Methods to detect and address these two types of errors need to be designed and implemented in the future.

5 Clustering vehicles

One direct application of trip data is unsupervised learning. In this section we identify groups of vehicles with similar trip patterns based on frequency and diversity of travel. Clusters can represent private vehicles, such as work-home commuters, transit vehicles like buses and taxis, or other types of vehicles such as delivery trucks.

Due to distinctly different traffic behaviour during weekends, we chose to consider only trips occurring during weekdays. We thus used all number-plate data collected between the 6th and 24th of February and exclude data from the 2 weekends in-between. Furthermore, as mentioned in section 4, we used fixed empirical values for T. Table 2 displays the number of trips, average trip length and the proportion of trips of length one, for varying values of T. The effect of incrementing T on resulting trips is clear: increased trip lengths and decreased trips containing just a single sighting. On the other hand, we did not set a value for τ, but instead handled implausible trips by filtering all sightings with confidence below 85%. Duplicates were filtered after identifying consecutive sightings by the same camera occurring within the same trip. Clock synchronisation errors were provided in milliseconds with none exceeding 5 seconds. These were therefore ignored.

Transforming trips into features which can be used in clustering algorithms was a 3-step process: (i) First, every trip was summarised as a single row of data. The following information was extracted: length of trip, origin, destination, route, start and end times. (ii) Second, daily trip information was obtained for each vehicle: number of trips, median of trip lengths, number of sightings, distinct number of origins destinations, and routes, hour of first sighting, hour of last sighting and total rest time between trips. (iii) Finally, daily information per vehicle was collapsed into a single row by averaging this information across the 15 days.

Tau	Trips	Average Length	Proportion of Trips Length 1
5 min	13,603,759	1.46	0.70
7.5 min	12,394,709	1.60	0.64
10 min	11,690,791	1.69	0.61
15 min	10,823,333	1.83	0.57
20 min	10,305,860	1.92	0.54
30 min	9,653,499	2.04	0.52

Table 2: Overview of trip data for varying values of T.

Table 3 depicts a sample of the resulting features vector. A total of 1034107 distinct vehicles were detected. However, because there is a high percentage of trips containing a single sighting, some of these features were highly correlated. We therefore, chose to remove three of the features represented in 3: *Average Sightings*, *Average Distinct Routes* and *Average Hour Difference*, to avoid the obfuscation of the natural clustering [11]. Furthermore we considered that a trip of length one has no destination (which explains values of average distinct destination below zero) and we filtered all instances of vehicles where the total number of trips is lower than 3, resulting in 642006 unique vehicles.

Clustering of vehicles was performed using the Hartigan and Wong *k-means* algorithm, for each value of T. The number of clusters k was varied between 2 and 8 and executed with a maximum of 200 iterations and 100 different starting states of the algorithm. The Calinski-Harabasz criterion is used to pick the best value of k, the one that minimises the within-cluster and between-cluster errors and provides the more natural clusters [11]. The results of vehicle clustering are presented and discussed in the section below.

6 Results and Discussion

Table 4 provides a summary of multiple runs of *k-means*. For each value of T, the optimal number of clusters is selected by picking the value of k that maximises the Calinski-Harabasz criterion. Furthermore, the measures of inter-cluster (betweeness) and intra-cluster (withiness)

Total Trips	Average Trips	Average Length	Average Sightings	Average Distinct Origins	Average Distinct Destinations	Average Distinct Routes	Average First Hour	Average Last Hour	Average Hour Difference	Average Rest Time
41	3.42	1.25	5.75	2.75	1.00	3.00	15.33	19.25	3.80	3.70
3	1.50	1.00	1.50	1.50	0.00	1.50	14.00	14.50	0.73	0.73
7	2.33	1.33	3.33	2.33	0.67	2.33	11.00	13.00	2.44	2.41
12	2.40	1.10	3.60	2.40	0.60	2.40	14.40	16.40	2.00	1.95

Table 3: Sample of extracted features from trips taken from 15 weekdays of number plate data.

are depicted relative to the corresponding total sum of squares (total betweeness and total withiness). It is noteworthy that the best value of k increases inversely to T. Although a higher value of k seems to suggest that trips with smaller values of T are able capture the variance in the data better, we have to consider that varying T affects the average trip length in the same direction whilst affecting the total number of trips in the opposite direction (table 2).

Tables 5 and 6 depict the cluster centres for T equal to 7.5 minutes and 20 minutes respectively. These partially meet our expectations. For instance, we were expecting to find a relatively small cluster representing taxis with a high average number of trips per day, occurring over a variety of origins and destinations and over a large time frame. Clusters 2 and 7 for T = 7.5 and cluster 4 for T = 20 do indeed fit this profile. What differentiates cluster 2 from 7 in the first case is essentially a lower mean number of trips per day and smaller average time at which the first and last trips of the day occur. Additionally, buses to some extent fit this category as well, however we expected that buses could be separated from taxis by showing less diversity in the number of origins and destinations as these essentially do multiple runs of the same trip throughout the day. Still, this can be explained by the fact that buses take routes through main and secondary roads. As most cameras are placed in main roads, the one long bus trip can be perceived as multiple small trips as the bus alternates between arterial and main roads. This is in fact, one of the downsides of ANPR data and is something that the methodology presented in section 4 needs to develop in the future.

On the other hand, we expected to observe a group representing home to work commuters with the first trip of the day starting approximately at eight in the morning and a second trip terminating between five and six in the evening. Although we observe one or two groups with those characteristics, these contain a higher average of trips per day than expected, which may represent for example work to school trips. However, we observe at least one big group of trips occurring mostly during lunch hour. Several interpretations are possible but further work is needed to provide a more consistent interpretation and validation of these results. A big contributing factor however is the fact that a high proportion of trips contains only a single sighting (table 2). It may be the case that commuters choose other routes than those going through ANPR cameras. Still, we need to devise further models and methods that are able to capture this uncertainty.

Tau	Best k	Betweenss	Average Withinss	Calinski-Harabasz
5 min	8	0.923	0.125	1,116,962
7.5 min	7	0.904	0.143	1,003,744
10 min	7	0.896	0.143	911,044
15 min	6	0.865	0.167	794,557
20 min	4	0.786	0.250	754,241
30 min	3	0.710	0.333	743,001

Table 4: *k-means* performance for several values of T.

7 Conclusion and Future Work

Most urban cities in the world employ a network of ANPR cameras that are used for law enforcement as well as traffic monitoring and control. Number plate data collected in the Tyne and Wear area is stored and leveraged by the UTMC for computing average journey times across a selection of sensitive roads. However, number plate data could be used more extensively to identify and study individual and collective travel patterns. In this paper, we have presented a set of definitions and constraints that establish a conceptual foundation for identifying vehicle trips from number plate detections. We also identify two parameters, τ and T as critical in the discrimination of plausible and implausible trips. Hence, future work should first and foremost focus on developing formal methods to estimate these parameters from observed distributions of travel times and by applying knowledge about the structure of the road network. Moreover, methods for addressing issues concerning camera performance, namely wrong and duplicate scans, should be further developed and researched.

Cluster	Size	Total Trips	Average Trips	Average Length	Average Distinct Origins	Average Distinct Destinations	Average First Hour	Average Last Hour	Average Rest Time
1	108868	30.67	2.91	1.56	2.65	1.09	9.67	16.11	6.35
2	2948	170.08	13.69	1.47	9.20	4.67	6.89	18.69	11.65
3	309748	5.95	2.01	1.44	1.92	0.68	12.52	14.50	1.93
4	163982	16.99	2.30	1.46	2.13	0.77	11.12	14.99	3.81
5	45414	50.30	4.26	1.52	3.67	1.57	9.04	16.86	7.69
6	10380	86.99	7.16	1.43	5.55	2.49	8.41	17.60	9.03
7	666	331.50	26.43	1.57	9.31	5.17	5.73	19.59	13.14

Table 5: Clusters sizes and centres for T = 7.5 minutes.

Cluster	Size	Total Trips	Average Trips	Average Length	Average Distinct Origins	Average Distinct Destinations	Average First Hour	Average Last Hour	Average Rest Time
1	371318	7.19	1.76	1.68	1.68	0.71	12.30	14.55	2.14
2	49852	44.85	3.73	1.90	3.16	1.73	8.98	17.02	7.76
3	189224	22.82	2.29	1.84	2.11	1.06	10.00	15.76	5.60
4	4593	114.34	9.12	2.30	5.92	3.98	6.93	18.21	10.43

Table 6: Clusters sizes and centres for T = 20 minutes.

Once trip data has been computed, a range of interesting applications are available. This work tries to identify groups of vehicles by clustering information about frequency and diversity of travel. By associating a vehicle with a cluster that represents taxis, or home-to-work commuters, one can begin estimating trip mode usage across the city. However, the results presented here could benefit from extra work and further validation. Finally, future work can focus on using trip data to solve interesting research problems such as: (i) real-time route recommendation using probabilistic graphical models; (ii) detection of abnormal trip patterns for helping law enforcement in the identification of suspect vehicles or behaviour; (iii) modelling how drivers make routing choices in the presence of anomalies in the road network.

References

[1] Alexander, Lauren, et al. "Origin-destination trips by purpose and time of day inferred from mobile phone data." Transportation research part c: emerging technologies 58 (2015): 240-250.

[2] Castillo, Enrique, et al. *Optimal use of plate-scanning resources for route flow estimation in traffic networks.* IEEE Transactions on Intelligent Transportation Systems 11.2 (2010): 380-391.

[3] Castillo, Enrique, Jos Mara Menndez, and Pilar Jimnez. *Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations.* Transportation Research Part B: Methodological 42.5 (2008): 455-481.

[4] Chen, Huiyu, Chao Yang, and Xiangdong Xu. *Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data.* Journal of Advanced Transportation 2017 (2017).

[5] Department for Transport *Provisional Road Traffic Estimates Great Britain: October 2016-September 2017.* url (visited January 2017): https://www.gov.uk-/government/statistics/provisional-road-traffic-estimates-great-britain-october-2016-to-september-2017.

[6] Department for Transport *Vehicle Licensing Statistics: Quarter 3 (Jul-Sep) 2017.* url (visited 31 January 2017): https://www.gov.uk/government/statistics-/vehicle-licensing-statistics-july-to-september-2017.

[7] Hamilton, Andrew, et al. *The evolution of urban traffic control: changing policy and technology.* Transportation planning and technology 36.1 (2013): 24-43.

[8] McNally, Michael G. "The four-step model." Handbook of Transport Modelling: 2nd Edition. Emerald Group Publishing Limited, 2007. 35-53.

[9] Parry, Katharina, and Martin L. Hazelton. *Estimation of origin-destination matrices from link counts and sporadic routing data.* Transportation Research Part B: Methodological 46.1 (2012): 175-188.

[10] Schssler, Nadine, and Kay W. Axhausen. *Identifying trips and activities and their characteristics from GPS raw data without further information.* Arbeitsbericht Verkehrs-und Raumplanung 502 (2008).

[11] Steinley, Douglas, and Michael J. Brusco. *Choosing the number of clusters in -means clustering.* Psychological Methods 16.3 (2011): 285.

[12] Zhang, Junping, et al. *Data-driven intelligent transportation systems: A survey.* IEEE Transactions on Intelligent Transportation Systems 12.4 (2011): 1624-1639.