Clustering Vehicle Trips Identified from Automatic Number Plate Recognition Camera Scans

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

Abstract

This is the text of my abstract. It is a brief description of my paper, outlining the purposes and goals I am trying to address.

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 [1] – 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 [2]. 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 on the roads.

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] 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 latte 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

*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.

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

In [4], the authors used partial trip information from plate number scans to compute travel demands, in the form of a trip matrix, and estimate link flows from a set of constraints composed of traffic conservation laws, prior knowledge and path flow disaggregation techniques, such as Stochastic User Equilibrium (SUE).

Uses of trip data for understandint travel demands / traffic prediction

Studies about best placement of cameras

Number plate data is not as popular, for urban traffic monitoring and prediction , as other sources of data, namely GPS traces or floating vehicle data. Previous works trips from GPS traces, .

Although significant research has gone into identifying trips from GPS traces, to the best of our knowledge, little research has gone into doing so with number plate

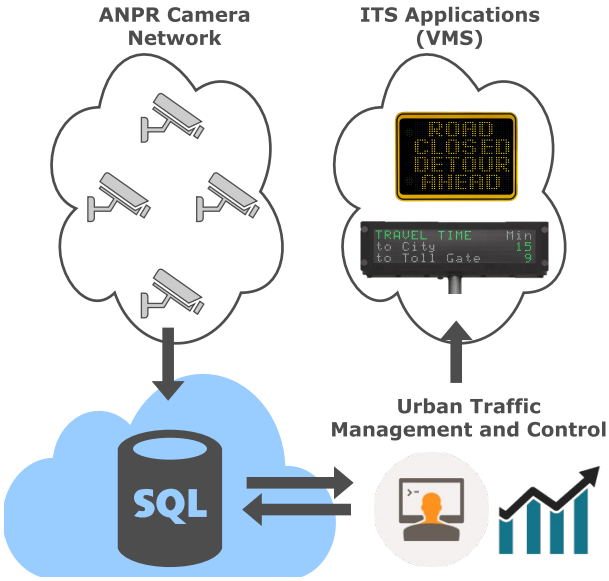


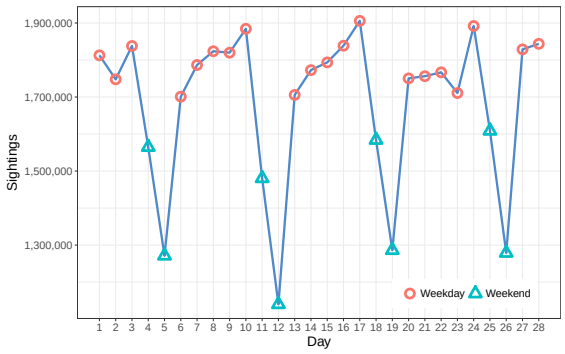
Figure 1

data. Therefore, the contributions of this paper are twofold:

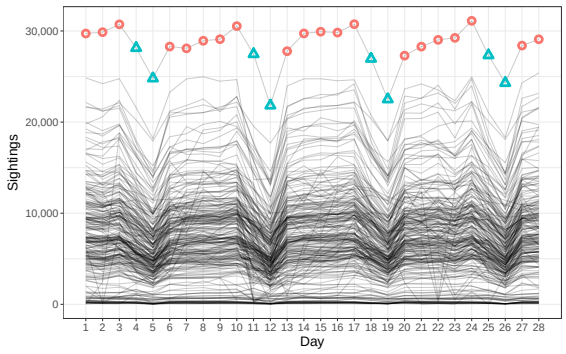
3 Tyne and Wear ANPR Data

Automatic number plate recognition (ANPR) cameras are actively employed in urban traffic environments 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 comissioned 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 [6, 9].

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 the Urban Traffic Management and Control (UTMC) of 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



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

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 plate numbers 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.

The main use of ANPR data for the UTMC is estimating average journey times for selected links in the road network. Furthermore, several authors have extensively researched how to use used number plate data as an extension on link counts for estimating origin-destination

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
6756787	146	1454284801.53	22	99
8487265	2	1454284803.88	10	93

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

matrices and link flows [3, 4, 7]. 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 the following section we present a conceptual methodology for grouping multiple camera observations of the same vehicle into one or several trips of that vehicle.

4 Trip Identification

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

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

where c is an integer that 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 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 k valid if the following conditions are met:

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

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

The first condition 4.4 is straightforward and specifies that every trip should have at least one sighting. The second condition 4.5 defines a minimum and maximum travel times between consecutive observations. Its purpose is twofold: (i) first, to allow trips made by the same vehicle to be differentiated. For instance, given two consecutive sightings of k three hours apart, we want to interpret them as belonging to different trips of k; (ii) second, it allows unplausible trips to be identified. For example, an unplausible 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 plate number travelling in the road network. Evidently, condition 4.5 is only valid for trips of length two or greater. Nevertheless, trips

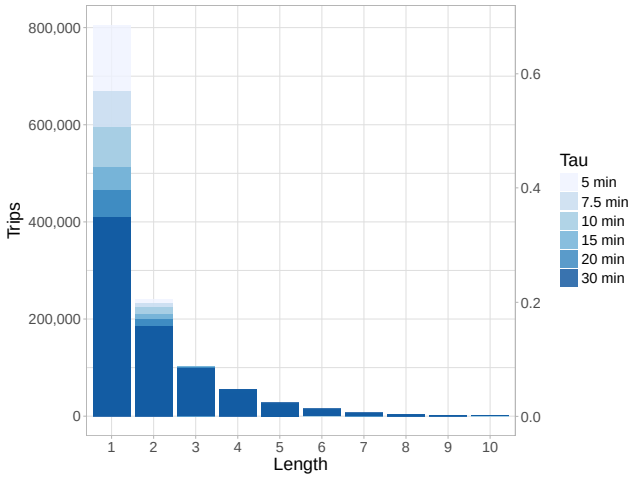


Figure 3: Distribution of trips per length of trip.

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.

The simplest approaching 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 origin-destination (od) 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 estimating 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 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 per length of trip varies by fixing T at different empirical

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

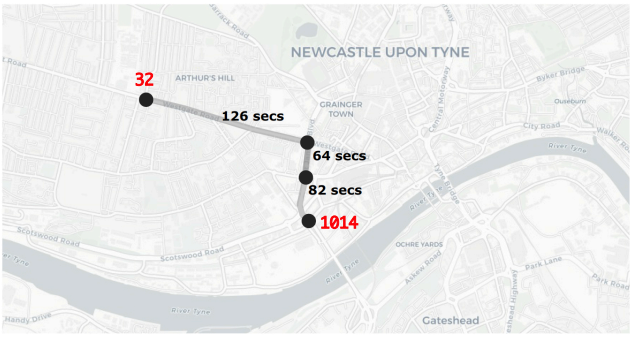


Figure 4: Example of a trip of length 4. On the left side the correspoding data table is shown. The u_{th} trip of each vehicle is given by the variable *Trip*, whereas the i_{th} sighting is given by variable *Sighting*. The variable *JourneyTime* gives the travel time from the previous to current sighting. 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 [4]).

values (τ is ignored).

4.1 Duplicate scannings To ensure that every trip of vehicle k is unique in the sequence of all valid trips of vehicle 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:

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

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:

$$t^k_i \neq t^k_j \Rightarrow s^k_i \neq s^k_j \ , \ i \neq j \tag{4.8}$$

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 impact in error propogation, 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 Even though number plate recognition cameras have accuracy rates of 99.9% or higher, on average 1 observation out of 1000 is a misclassified plate number. We categorise the errors made by ANPR cameras into two types:

1. a passing vehicle is not detected by the camera;
2. 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. On the second case, there is a recorded sighting, but this will be assigned to the incorrect vehicle.

5 Clustering vehicles

One direct application of trip data is unsupervised learning. Previous works have focused on clustering trips in order to identify travel modes or purpose of trip [5]. Estimating a distribution of trip mode travel, is one of the fundamental steps in traffic modelling . In this work we focus on identifying groups of vehicles with similar trip patterns based on frequency and diversity of travel. For instance, taxis usually . However, due to its tiny coverage of the road network, it is not clear if ANPR data can capture the travel characteristics that describe the several groups. Thus, we aim with this work to evaluate how good ANPR data is for providing trip information and ultimately that of the underlyng road network. In the remaining of this section we provide detail on the choice of parameters and filtering methods used in trip identification, as well as in the choice of features. Then, we

Due to distinct traffic behavior during weekends, we chose to consider only trips occurring during weekdays. We thus used all number plate data collected during between the 6th and 24th of February and excluded data from the 2 weekends in between. Furthermore, as mentioned in section 4, we used fixed empirical values for T, mostly due to time constraints. As such, we ran the trip identification sequence for the following values of T: $[5, 7.5, 10, 15, 20, 30]$ minutes. The total number of trips identified during the 3 weeks, for each threshold, was approximately [1.36M, 1.24M, 1.17M, 1.08M, 1.03M, 0.97M] respectively. Moreover, we did not set a value for τ, but instead handled unplausible 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 that could be used in clustering algorithms was a 3-step process:

1. 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.

2. 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.

3. Finaly, daily information per vehicle was collapsed into a single row by averaging this information across the 15 days.

Table /reft:features depicts a sample of the resulting features vector. Each The aim was to choose features that captured different aspects of trip frequency and diversity. However, because a high percentage of trips contain a single sighting. 1034107 vehicles

We identified correlated features and removed the following features .. because ..

Filtered all cases where TotalTrips ¡= 3. (HOW MANY?)

We ran the kmeans algorithm: k = 2:8 iter.max = 200, runs = 100

6 Results and Discussion

7 Conclusion and Future Work

References

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

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

[3] 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.

[4] 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.

[5] 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).

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

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

Tau (min)	Best k	Betweenss	Whithinss	Calinski-Harabasz
5	8	0.92	0.09	1,116,962
7.5	7	0.90	0.10	1,003,744
10	7	0.90	0.18	911,044
15	6	0.86	0.19	794,557
20	4	0.79	0.19	754,241
30	3	0.71	0.30	743,001

Table 3: Kmeans performance comparison for each best k

[7] 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.

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

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

[10] Zhu, Shanjiang, and David Levinson. *Do people use the shortest path? An empirical test of Wardrops first principle*. PloS one 10.8 (2015): e0134322.

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 4: Clusters sizes and mean centers for T = 7.5 minutes