Clustering of Usage Profiles for Electric Vehicle Behaviour Analysis

Constance Crozier, Dimitra Apostolopoulou, and Malcolm McCulloch
Department of Engineering Science
University of Oxford
Oxford, UK OX1 3PJ
Email: constance.crozier@eng.ox.ac.uk

***Abstract*—Accurately predicting the behaviour of electric vehicles is going to be imperative for network operators. In order for vehicles to participate in either smart charging schemes or providing grid services, their availability and charge requirements must be forecasted. Their relative novelty means that data concerning electric vehicles is scarce and biased, however we have been collecting data on conventional vehicles for many years. This paper uses cluster analysis of travel survey data from the UK to identify typical conventional vehicle usage profiles. To this end, we determine the feature vector, introduce an appropriate distance metric, and choose a number of clusters. Five clusters are identified, and their suitability for electrification is discussed. A smaller data set of electric vehicles is then used to compare the current electric fleet behaviour with the conventional one.**

***Keywords*—Clustering algorithms, Demand forecasting, Electric vehicles, Pattern analysis**

I. Introduction

This paper uses cluster analysis to identify typical vehicle usage profiles. Such an analysis is important to reduce the complexity of vehicle use prediction and to compare the behaviours of the current conventional and electric fleets.

Charging of electric vehicles (EVs) has the potential to significantly alter household electricity demand. If left uncontrolled, this would have serious consequences for the network infrastructure [1]. Therefore, research into methods of charging EVs with minimal impact on the grid, referred to as *smart charging schemes*, are gaining popularity. In order to practically implement smart charging the distribution network operator (DNO) needs to accurately predict the behaviour of the vehicles on its network. Both the energy consumed by the vehicle and when it is likely to plug in are critical pieces of information for planning. In the future it may also be possible for EVs to supply power back to the grid [2], and successfully incorporating this would also require accurate prediction of vehicles' availability.

In the UK EVs currently account for just over 2% of new vehicles sales [3], and it has been shown in existing literature that these consumers display some certain demographic traits with respect to the nation's vehicle owners. [4] report that EV users are typically male, highly educated, have high incomes, and often more than one vehicle in the household. This demographic is likely to result in a different set of usage profiles, and for short-term planning these need to be identified. However, as the penetration of electric vehicles increases it is likely that the behaviour of the electric fleet will converge towards the conventional one. There is therefore value in assessing both how conventional vehicles are used, and how the current EV usage differs from it.

Cluster analysis of data allows patterns to be identified, and the dimensionality of the data to be reduced. In the case of vehicle use, identifying patterns allows the fleet utilisation to be better understood and visualised. If a set of representative profiles is found it also simplifies the computation of the prediction problem; instead of predicting individual requirements we need to predict the number of vehicles belonging to each cluster.

Previous smart charging research has adopted more simplistic methods for modelling EV behaviour. A common approach is to assume discrete probability distributions for energy consumption and arrival time of vehicles (e.g. [5]–[7]). However, these two quantities are unlikely to be independent - the further a vehicle has travelled the later it is likely to arrive home. Furthermore, [8] note that summary travel statistics models like this miss details required for impact analyses at the distribution network level.

Some previous work has incorporated clustering into EV smart charging, although not for the prediction of vehicle behaviour. For example, [9] proposes a charge optimisation scheme which clusters vehicles by location, allowing the charging station used to become a decision variable. Clustering of EV data has also been used in other areas of research. In [10] it is used to improve the efficiency of vehicle-to-infrastructure communication, and [11] uses clustering of vehicles to find the optimal locations for fast charging stations. Clustering of usage profiles has been used more extensively in the area of household electricity prediction (e.g. [12]). In this paper we cluster vehicle usage using techniques which have previously been used to categorise household demand.

Here we are focusing solely on the UK vehicle fleet, although with appropriate data the methods presented could be replicated for other areas. We are also only considering weekday driving, as vehicles behave significantly differently on weekdays and weekends. The former was chosen as vehicle usage is on average higher and more diverse, making the prediction both more important and more challenging.

In this paper we will first explain the data used, then outline

978-1-5386-4505-5/18/$31.00

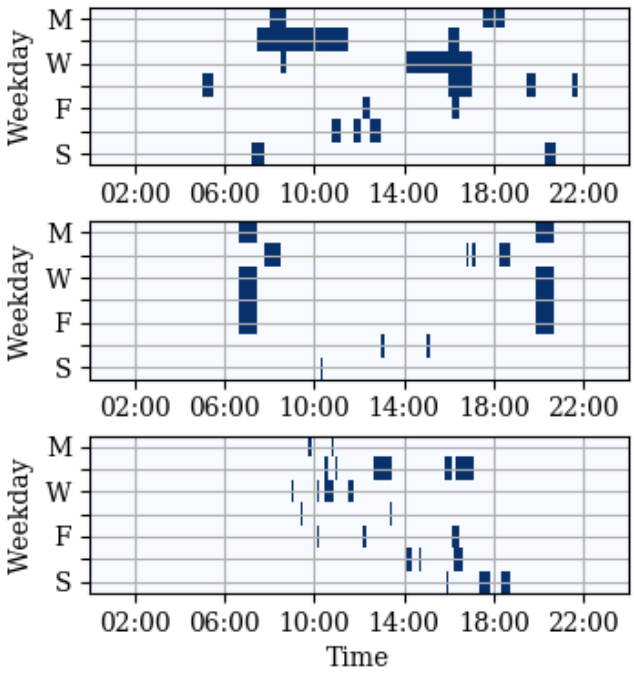

Fig. 1. Example vehicle usage profiles extracted from the NTS.

the clustering algorithm employed. The resulting clusters from the conventional vehicle data are explored, and the differences between the conventional and electric fleet are quantified.

II. Data Sources

For this research a large data set of conventional vehicle usage was used and compared with a smaller set of EV profiles.

A. National Travel Survey

To investigate vehicle use in the UK, the *National Travel Survey* (NTS) was used [13] . This is an annual project which records all of the trips undertaken by a household for a week, where the households are randomly selected from across the country. Journeys are recorded by hand; participants submit a completed travel diary at the end of the week. The full data set contains the time, distance, purpose and mode of transport (among other things) of nearly 2 million journeys. By searching through the data set for trips with the same vehicle ID, over 100,000 individual usage profiles can be extracted. Three example week long vehicle usage profiles are shown in Figure 1, where time of day is on the horizontal axis, week day on the vertical, and blue areas represent times when the vehicle is in use. This allows vehicle specific characteristics to be identified, for instance the second vehicle is shown to have a very regular commute which it carried out on 4/5 weekdays. As the travel diaries are only recorded for a week, this is the longest profile which can be extracted.

Similar surveys are conducted in other countries, such as the *National Household Travel Survey* in the US. The methods produced in this paper could be applied to any such survey, where car journeys are recorded provided the vehicle ID, distance and timings of trips were recorded.

B. Electric vehicle usage

My Electric Avenue (MEA) was a project completed in 2016 which aimed to investigate the impact of EVs on distribution systems [14]. Nissan Leafs were loaned to households for a period of at least 18 months, during which all vehicle use was recorded. The households were in concentrated geographic groups, meaning those within the same group were charging from the same low-voltage network. The project was opt-in so participants behaviour is likely to be representative of EV early adopters.

III. Clustering Analysis

Clustering is the accumulation of data into groups of similar points. In this paper we focus on one of the simplest methods: K-Means clustering. This aims to partition data into K clusters, each of which is defined by a centroid which is the mean position of the points in the cluster. Each point belongs to the cluster whose centroid is closest to it, and the centroid positions are chosen so as to minimize the average distance of a point from its cluster's centroid. The positions can be computed using Lloyd's algorithm, a full formulation of which is presented in [15].

In order to implement this type of clustering one needs to decide on: a feature vector, a distance metric and the number of clusters present in the data.

A. Feature Vector

A feature vector is a set of numbers which describe each point in the data set to be clustered. The cluster positions found by the algorithm are highly sensitive to the choice of feature vector, so choosing the variables carefully is critical.

In this study we are interested in identifying typical vehicle usage patterns. One approach would be to choose parameters which we think describe the vehicle use, e.g. distance travelled or number of trips. However, as we are carrying out this research with a view to planning charging it is important to know when, as well as how much, a vehicle is being used. Therefore, it was decided to use the average usage profile for the feature vector, so that each element of the feature vector represents the usage at a certain time of day. As the NTS data does not record energy consumption, here the usage is defined by the distance the vehicle has moved in the time period, or the average speed. There is not enough information to recover the actual velocity profile - we are effectively assuming that each journey is completed at a constant speed. This approximation is justified as it is the distance rather than speed which most strongly dictates energy consumption, so it is the distance we are more interested in. The total distance that a vehicle has travelled can be recovered from the usage profile by integration.

For each vehicle we have 5 weekdays of data, which were combined to create an average profile. The average was used because it describes not only the vehicle usage but how repetitive it is, as the highest values will indicate a fast journey carried out on the same time each day. By averaging the usage profiles we get the expected distance travelled on any given day at that time, so if a vehicle did a long journey one day but not on the others the expectation would have a low value.

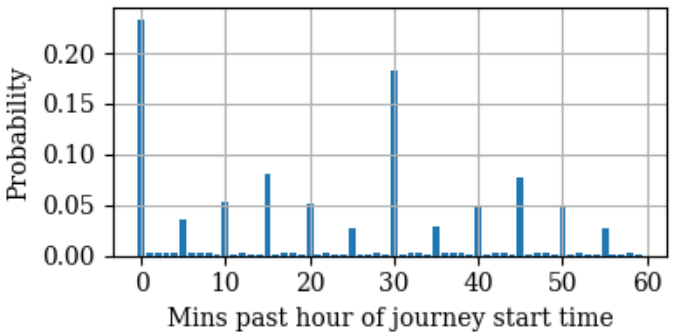

Fig. 2. The probability distribution of minutes past the hour of a recorded journey start time.

A decision has to be made on the time resolution of the profiles used. In the NTS journey times are recorded in minutes past midnight, so theoretically we could cluster on profiles at 1 minute resolution. However, that would result in 1440-point long feature vectors which will be computationally expensive, and will limit the number of points we can cluster on. Using this time resolution would also be pointless because the data is unlikely to be accurate to this level, given they are recorded by hand. In fact, closer examination of the survey data found that *rounder* times were much more likely to be recorded. Figure 2 shows the probability distribution of minutes past the hour of a recorded journey start time. There appears to be a 40% chance of a journey either starting exactly on the hour or half hour which seems statistically improbable, more likely people are rounding their trips to the nearest half hour. It was therefore decided to use a resolution of half an hour, resulting in 48 points per feature vector.

It is common practice in clustering to normalise profiles, so that the relative shape of the feature vector variables are compared rather than their magnitude. However, when considering the use of EVs the distance it has travelled is one of the most important features for planning charging. Therefore, we have not normalised the profiles but uniformly scaled them to have a maximum value of 1 - this limits the size of the distances calculated during the composition. Nine example feature vectors extracted from the NTS data are shown in Figure 3.

B. Distance Metric

Standard K-Means clustering uses the Euclidean distance as the distance metric between two data points. This describes the distance between two vectors $\mathbf{p}$ and $\mathbf{q}$ as:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_i (q_i - p_i)^2}. \quad (1)$$

However, this is not ideal for clustering profiles as Euclidean distances treat every point in the vector independently. To see why this is problematic for time signals consider Figure 4, which shows three synthetic profiles, *(a-c)*, all of which contain exactly two non-zero elements. *(a)* and *(b)* represent the same signal shifted by one time instant while *(c)* is completely different. Clearly we would recognise *(a)* and *(b)* as more similar than *(c)*. Unfortunately, using Euclidean distances, all of these signals are equidistant from another.

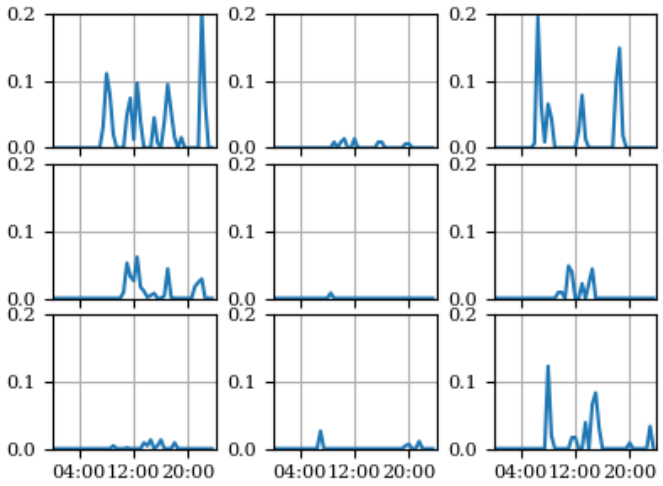

Fig. 3. Example feature vectors as extracted from the travel survey.

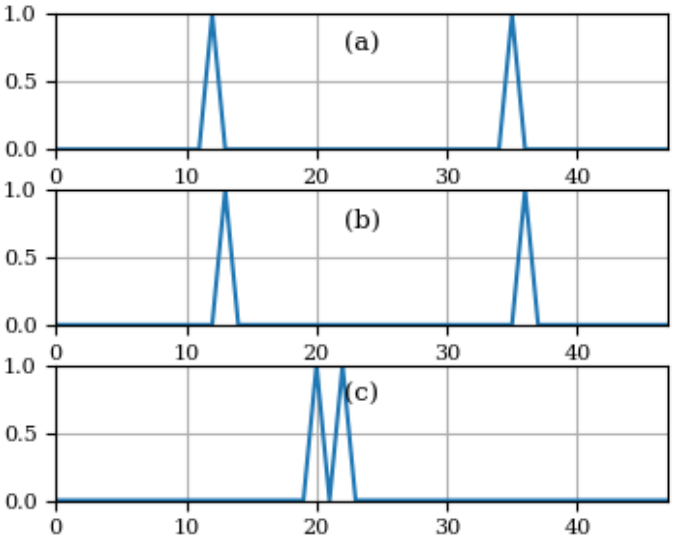

Fig. 4. Three synthetic signals illustrating why Euclidean distance is inappropriate in this case.

One common approach for the similarity of time-series is to use their auto-correlation, which finds the maximum dot product of the signals over all possible relative time-shifts. However, shift invariance is not desirable for planning charging a vehicle which carries out one journey at 2AM is not the same as one that carries one out at 6PM. Instead, it was decided to apply a Gaussian filter to the profiles. This smooths the profiles with time, thereby encoding information about the surrounding times in the features. Figure 5 shows the same three synthetic profiles with filters of 0.5, 1 and 2 point standard deviations applied. When the Euclidean distance is now applied it will identify that *(a)* is much closer to *(b)* than *(c)*. Carefully choosing the filter width is important; too narrow a filter won't be effective, while too wide sacrifices valuable information. Here we chose the 1 point standard deviation filter.

C. Selecting number of clusters

One of the biggest challenges of K-Means clustering is determing the value of K, or the number of clusters. This can

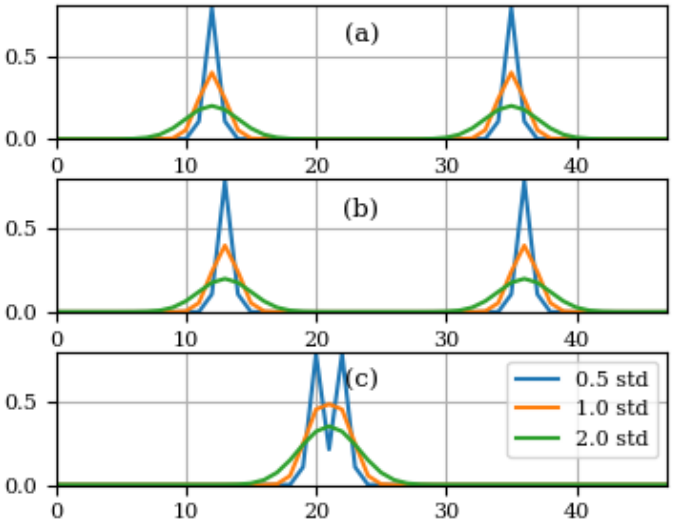

Fig. 5. Three synthetic signals illustrating why Euclidean distance is inappropriate in this case.

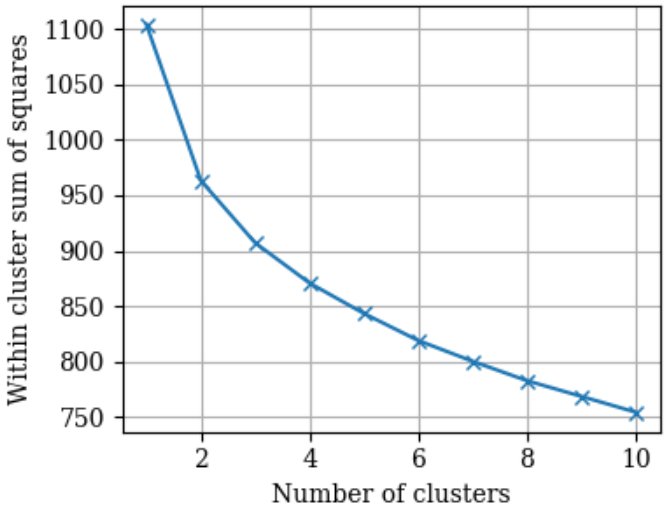

Fig. 6. The variation in within cluster sum of squares with number of clusters for the NTS data.

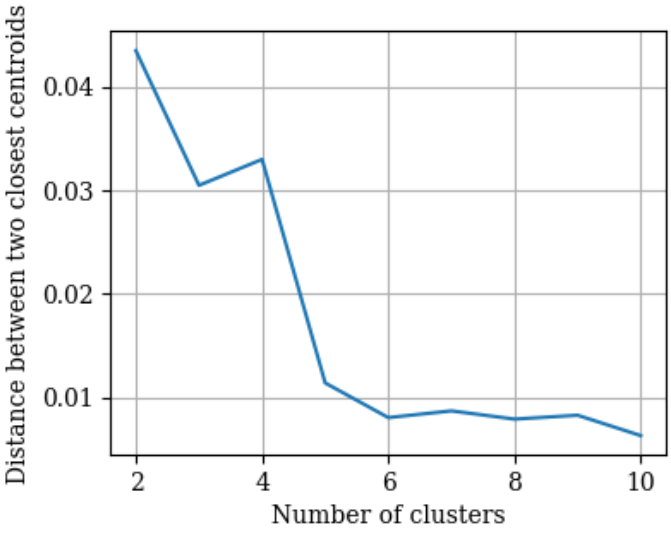

Fig. 7. The variation in distance between the two closest cluster centroids with number of clusters.

be a difficult job, particularly when the goal of the clustering exercise is dimensionality reduction. As K is increased the clusters necessarily describe the data more accurately, however the more clusters there are the less the dimensionality of the problem is reduced.

There are several heuristic methods which can be used to select a number of clusters, and their suitability is context dependant. In this case we attempted the *elbow method* (e.g. [16]). This involves plotting the variation of *total within-clusters sum of squares* (the sum of the squared distance of every point from its cluster's centroid) with number of clusters. The relationship is necessarily monotonically decreasing as adding more clusters will always reduce the average distance of a point from a cluster, however typically there is a sharp initial decrease which levels off. The point at which the decrease transitions from sharp to gradual (the elbow) is chosen as the number of clusters. The logic is that if introducing an additional cluster significantly reduces cluster variance then it is valuable, but if only a moderate reduction in variance is achieved then it is not. The variation of within cluster sum of squares with number of clusters for the NTS data is shown in Figure 6. In this case there is not an obvious elbow, however a transition from steep to gradual seems to occur around 4 or 5 clusters.

As we are interested in determining distinct vehicle usage patterns it is also useful to visualise how different each additional cluster is compared to the pre-existing ones. This can be accomplished by looking at the variation of distance between the two closest centroids with number of clusters, which is shown in Figure 7. It should be noted that, unlike the within cluster sum of squares, this tells us nothing about how well the clusters describe the data, but should highlight when there are two similar clusters. In this case we can see that after 4/5 clusters the distance between the closest two clusters becomes relatively small. This implies that, after this point there are two centroids which are very similar. Therefore it was decided to use K=5 clusters for the remains of the analysis.

IV. Results

The clustering algorithm described in Section III was carried out on 30,000 vehicle usage profiles from the NTS. The resulting clusters are displayed in Figure 8. Each centroid is displayed on a separate subplot using a solid line, the shaded region covers a 90% confidence interval for the points belonging to the cluster. In brackets below the cluster number is the percentage of the data set which occupies that cluster. The legend shows the distance covered by the centroid profile.

Three of the the five clusters seem to represent commuting vehicles (clusters 1, 3, and 4) which together comprise only 18.6% of the vehicles in the data set. Of these number 4 is the most common, and also the shortest. Number 5 has a similar but both leaves and arrives home earlier, it the pre-processing blur is increased to several hours then these clusters merge. Number 1 is the least populated cluster and appears to

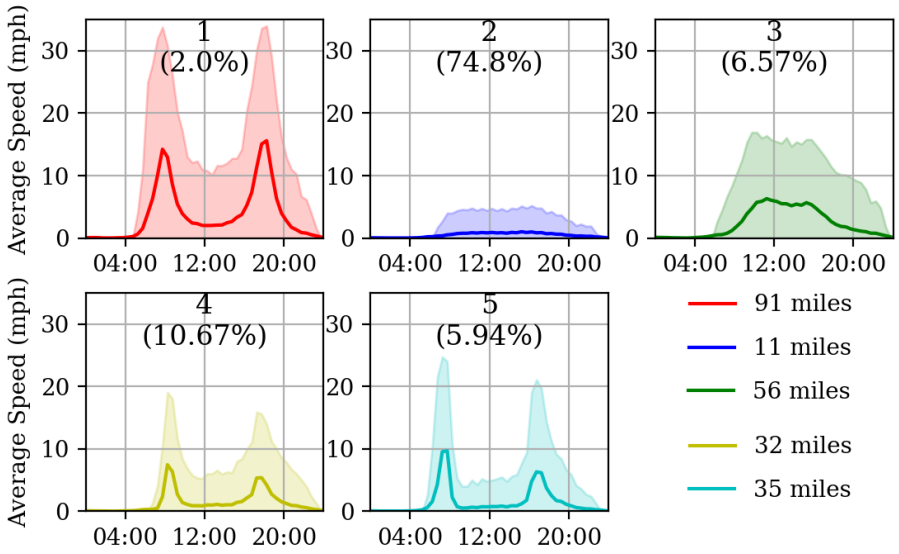

Fig. 8. The centroids, 90% confidence intervals, average distances, and size of the 5 clusters resulting from the NTS vehicle use profiles.

represent vehicles which both commute and drive throughout the day, this is the new cluster that emerges when you move from K = 4 to 5. Numbers 2 and 3 show very flat usage profiles, suggesting the vehicles are used uniformly throughout the day - the divisor is the difference in distance they cover.

Perhaps the most striking result is that the vast majority (nearly 75%) of the vehicles are in cluster 2. This is the lowest usage group, covering an average of just 10 miles a day, and suggests that the national fleet is under-utilised. If the eventual electric fleet has this composition then only a small percentage of vehicles will need charging on a daily basis.

As well as the average usage profiles it is interesting to look at the distribution of distance travelled within the clusters. Figure 9 shows the probability distribution functions of the daily distance travelled within each of the identified clusters. 85 miles is the typical range of an EV, so the probability that a vehicle in the cluster will travel less than that in a day is also displayed. This shows that the vast majority of vehicles from clusters 2, 4, and 5 would routinely manage on one charge a day. In cluster 3 12% of vehicles would either require a larger battery or additional charging, and in cluster 1 49% of vehicles would. This tells us that of the vehicles regularly driving more than 85 miles a day, 56% are them will be commuters and 44% are more uniformly used.

A. *Comparison with existing EV fleet*

The early adopters of EVs are likely to have usage profiles which are distinct from the current conventional vehicles. Cluster analysis of the MEA fleet, with only 200 vehicles, is not likely to yield any notable results. However, by assigning the vehicles to the clusters identified from the NTS the composition of the two fleets can be compared. Figure 10 shows the cluster distribution for both data sets.

The dominating low use cluster is significantly smaller in the MEA data (although still the most populated). It makes sense that participants in an EV trial would drive more than average. Also notable is that there were no vehicles in the MEA

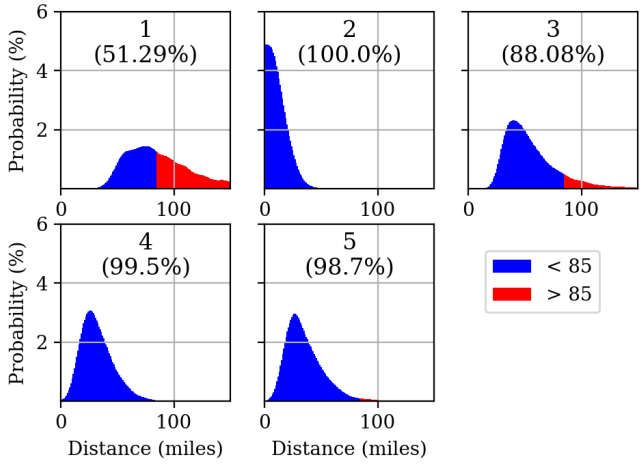

Fig. 9. The distribution of average daily distance within clusters, in brackets is shown the chance of travelling less than 85 miles a day on average.

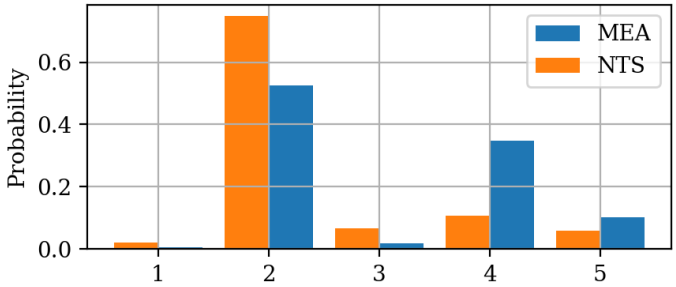

Fig. 10. The probability distribution of cluster composition for both the NTS and the MEA vehicles.

dataset belonging to cluster 3, the highest use class. This is also unsurprising as the EVs in question have an insufficient range to complete the centroid usage on one charge. MEA used 24kWh Nissan Leafs which have an 84 mile range, while cluster 3 travels an average of 92 miles.

Compared with the national average, the EV data contained a significantly higher proportion of commuting vehicles (25% rather than 13%). This aligns with the findings of [4] that early EV adopters have high incomes.

V. Conclusion

In this paper K-Means clustering was performed on vehicle usage profiles extracted from the National Travel Survey. The resulting clusters were examined and a smaller EV data set was used to compare the behaviours of the electric and national fleets.

Five typical conventional vehicle usage profiles were identified, three of which represented commuting vehicles. Nearly 70% of vehicles were in the lowest use group, meaning that in the UK 65% of the fleet mileage is completed by 30% of vehicles. If the EV fleet behaviour converges to the current conventional one then smart charging will rely on accurately identifying this 30% of vehicles. There would also be significant potential for vehicles to provide services to the grid, as so many of them could be relied on to be available.

In the short term however, EVs appear to have a higher than average level of use. Almost double the number of commuting vehicles were identified in the electric fleet data than in the conventional one. This means that the first vehicles on the grid will have significant charging demands and less availability to provide services to the grid.

References

[1] L. Fernandez *et al.*, "Assessment of the Impact of Plug-in Electric Vehicles on Distribution Networks," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 206–213, 2011.

[2] W. Kempton and J. Tomić, "Vehicle-to-grid power fundamentals: Calculating capacity and net revenue," *Journal of Power Sources*, vol. 144, no. 1, pp. 268–279, 2005.

[3] "Society of motor manufacturers and traders (smmt) ev registration data," https://www.smmt.co.uk/2018/01/december-ev-registrations/, accessed: 2018-03-11.

[4] S. Haustein and A. F. Jensen, "Factors of electric vehicle adoption: A comparison of conventional and electric car users based on an extended theory of planned behavior," *International Journal of Sustainable Transportation*, pp. 1–13, 2018.

[5] Z. Darabi and M. Ferdowsi, "Aggregated impact of plug-in hybrid electric vehicles on electricity demand profile," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 4, pp. 501–508, 2011.

[6] O. Sundstrom and C. Binding, "Flexible charging optimization for electric vehicles considering distribution grid constraints," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 26–37, 2012.

[7] S. Huang and D. Infield, "The impact of domestic Plug-in Hybrid Electric Vehicles on power distribution system loads," *POWERCON*, 2010.

[8] N. Daina, A. Sivakumar, and J. W. Polak, "Modelling electric vehicles use: a survey on the methods," pp. 447–460, 2017.

[9] J. Zhang, C. Yang, and F. Ju, "Optimization of Ordered Charging Strategy for Large Scale Electric Vehicles Based on Quadratic Clustering ," in *4th International Conference on Information Science and Control Engineering*, 2017.

[10] P. Luoto, M. Bennis, P. Pirinen, S. Samarakoon, K. Horneman, and M. Latva-Aho, "Vehicle clustering for improving enhanced LTE-V2X network performance," in *EuCNC 2017 - European Conference on Networks and Communications*, 2017.

[11] A. Shukla, K. Verma, and R. Kumar, "Consumer perspective based placement of electric vehicle charging stations by clustering techniques," in *2016 National Power Systems Conference, NPSC 2016*, 2017.

[12] T. K. Wijaya, M. Vasirani, S. Humeau, and K. Aberer, "Cluster-based Aggregate Forecasting for Residential Electricity Demand using Smart Meter Data," in *IEEE International Conference on Big Data*, 2015.

[13] K. Lepanjuuri, P. Cornick, C. Byron, I. Templeton, and J. Hurn, "National travel survey: 2015 report," Department for Transport, Tech. Rep., 2016.

[14] "My electric avenue," http://myelectricavenue.info/.

[15] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003, ch. 20, pp. 284–292.

[16] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means in wsn," *International Journal of Computer Applications*, 2014.