Generating Geospatially Realistic Driving Patterns Derived From Clustering Analysis Of Real EV Driving Data

Pedersen, Anders Bro; Aabrandt, Andreas; Østergaard, Jacob; Poulsen, Bjarne

Published in:
Proceedings of 2014 IEEE ISGT Asia Conference

Link to article, DOI:
10.1109/isgt-asia.2014.6873875

Publication date:
2014

Document Version
Early version, also known as pre-print

Generating Geospatially Realistic Driving Patterns Derived From Clustering Analysis Of Real EV Driving Data

Anders Bro Pedersen, Andreas Aabrandt
& Jacob Østergaard
Department for Electrical Engineering
Technical University of Denmark
abp@elektro.dtu.dk

Bjarne Poulsen
Department for Applied Mathematics and Computer Science
Technical University of Denmark
Kgs. Lyngby, Denmark
bjp@imm.dtu.dk

***Abstract*—In order to provide a vehicle fleet that realistically represents the predicted Electric Vehicle (EV) penetration for the future, a model is required that mimics people driving behaviour rather than simply playing back collected data. When the focus is broadened from on a traditional user-centric smart charging approach to be more grid-centric, it suddenly becomes important to know not just when- and how much the vehicles charge, but also where in the grid they plug in. Since one of the main goals of EV-grid studies is to find the saturation point, it is equally important that the simulation scales, which calls for a statistically correct, yet flexible model. This paper describes a method for modelling EV, based on non-categorized data, which takes into account the plug in locations of the vehicles. By using clustering analysis to extrapolate and classify the primary locations where the vehicles park, the model can be transferred geographically using known locations of the same classification.**

I. Introduction

One of the driving forces behind the Danish EV effort is the idea of using the batteries in the vehicles to balance out the fluctuations in the production, which is caused by a growing amount of wind power being produced in the country. The current wind penetration has already exceeded 30%, but with the government aiming for a fossile free production by 2050, the wind penetration is expected to reach 50% already by 2020.

Past projects like the danish funded EDISON project has already investigated Electric Vehicles (EVs) in relation to increased wind penetration, and currently the larger EcoGrid EU demonstration project is aimed at showing exactly this; a prototype powersystem with 50% renewable production [9][10][11].

The primary mechanism in the EcoGrid EU project is a socalled real-time electricity market, with a resolution of 5 minutes. In its most basic sense, an intelligent automated algorithm receives information about the current consumption from smart meters throughout the grid. Based on the current electricity price and the grids perceived need for up- or down regulation, it transmits a price to the customers that is meant to induce the necessary reaction. The general case works on a larger segment of the grid, but the next step would be to have locational prices [14].

Since EVs have yet to start selling in numbers that will cause them to rival existing Internal Combustion Engine (ICE) vehicles, it has been necessary to develop simulations to try to predict their impact on the grid. Since most, if not all, of these simulation have focused on the combined load, the EVs have often been modeled in a cumulative fashion. This technique works well for determining things like the day-to-day load imposed by a vehicle fleet, but does little to help determine the low-level impact such as the congestion you would likely see on a local feeder, should too many in the neighborhood decide to invest in an EV. The main purpose of a vehicle will always be transportation, hence it is also necessary to model the locational side of this behavior, if one wishes to realistically determine the local grid loads that result from an increased EV penetration. Knowing not only the daily consumption of a vehicle, but the detailed movements thereof, is a good platform for investigating both congestion preventive charging algorithms but also more user-centric services, such as route prediction, which could aid in facilitating e.g. automatic reservation of charging stations etc.

Many existing EV simulations, besides being cumulative, calculate the consumption based on an average efficiency and an expected daily driving distance. The method propsed in this paper uses recorded EV charging locations, which, through a clustering analysis, is partitioned into typical categories such as *home* and *work*. When identified, the identified locations can either form the basis for the simulation, or be translated elsewhere using existing databases of categorised addresses. Because the model incorporates actual locations, it is possible to create feasable routes using existing routing engines as is known e.g. from GPS navigation. This can provide a more realistic, not to mention flexible, consumption pattern. For example if a vehicle owner lives and works close to a major freeway, the vehicle will more likely be travelling along at a higher average speed, which results in an increased consumption compared to an inner city commuter. Once the vehicles have been modeled in the geospatial domain, the next step would be to attach them to a grid- model to simulate the actual impact. This is, however, slightly outside the scope of this

paper, but will briefly be touched upon in the final conslusion. Section II describes the dataset from the Danish "Test-en-elbil" project, on which the method in this paper is based. This is followed by a generic analysis of the data, looking into the availability of the EVs compared to findings of previous EV studies carried out on data obtained from ICE vehicles. Section III discusses the clustering of charge locations and the algorithms used along with various classification rules to obtain the desired result. Section IV describes the generation of the model(s) based on the clustering analysis. Section V wraps with a conclusion and a discussion of the next steps.

II. Data analysis

The Danish "Test-en-Elbil" ("Test-an-EV") project, which lays claim to the title of Europe's largest EV research project, was founded in 2010 by Clever A/S and is still ongoing [3]. The purpose of the project is to gain a better understanding of ordinary peoples driving styles and -habits when using EVs, and therefor puts great emphasis on logging how- and where the vehicles travel as well as charge. About 180+ Full Electric Vehicles (FEVs), predominantly belonging to a group of models commonly referred to as the "Triplets" (Peugeot iOn, Citroen C-Zero or Mitsubishi i-Miev), are currently used by the participants of the project. Everyone is free to requestion participation, but the ideal users for the project are families that fits the following profile:

- They must already own at least one vehicle.
- They must, to the extent permitted by the range etc., use the EV to cover all their everyday driving needs.

The chosen families are given the an EV for a duration of three months, after which the vehicle is returned. A multitude of measurements from the EVs are collected during the trials, some as fast as once per second, which are stored in the onboard computer before being uploaded to a server. Because some data, such as that from a Global Positioning System (GPS), is logged in a "raw" state, a certain amount of post-processing is required to partition this, so as to determine when the vehicles are driving, charging or just parked. This pre-analysis process is very helpful, since it allows the clustering analysis to easily focus on just the charging locations.

A. Comparitive analysis

In the past, other data collection projects have been carried out in Denmark which have been focused on ICE vehicles, since it was the only thing on this scale available at the time [2]. Interesting questions like consumption in relation to EVs would be speculative at best, but one question that was sought answered was how much time the vehicles stayed parked throughout the day. These numbers are very important from the grid side perspective, since they directly speak of how available the EVs are to potentially participate in balancing. Time parked is of course only an indicator to the true availability, but since an EV's battery is only accessible for balancing when it is plugged in, this is very important. While the data did not allow for the distinction between merely plugged in and actually charging, it was not possible to determine how often the vehicles in fact were grid connected. For this reason, the assumption was made, that when parked (i.e. not driving), the vehicle is grid connected. Many ideas, projects and businesses are addressing ways to incentivise users to plug in the vehicle whenever parked for a longer duration, but this is outside the scope of this paper. Before staring the clustering analysis an initial analysis of the pre-processed data was carried out, primarily to uncover whether there were any major differences between the previously analyzed ICE datasets and this, the first larger all-EV data set collected in Denmark. Previous studies carried out have found that the ICE vehicles were on average parked more than 90% of the time, and nearly all were parked during the critical night hours where the wind is usually most predominant [1][9][10]. The illustration in figure 1, which was borrowed from one such study, shows an overall availability of no less than 94%. This is presumably a result of most of the trips being of a relatively short duration, with little discernible overlap. The figure paints a picture of a typical commuter, with a morning trip to work around 8:00 and a homebound trip again around 16-17:00.

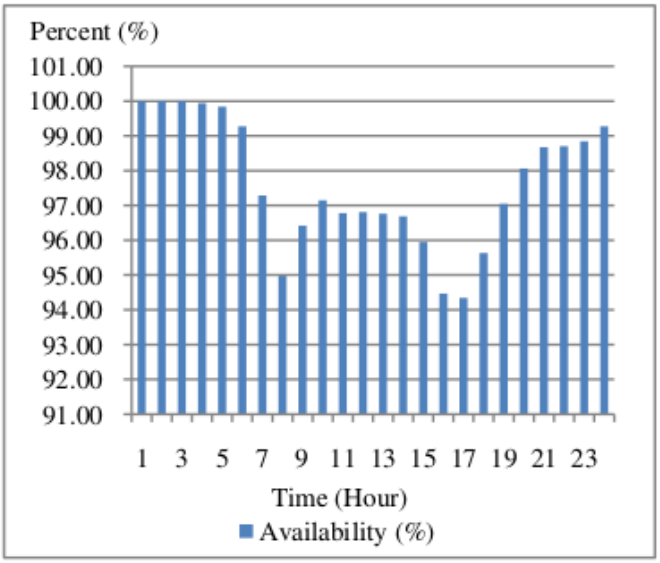

Fig. 1. Illustration of availability from ICE vehicle study (borrowed from [1])

To compare, the same analysis was carried out on the EV data from the "Test-en-elbil" project and the result, as seen in figure 2, shows an overall availability of at least 89-90% with an average of more than 96%. Not only is this in line with the previous findings for ICE vehicles, and because the availability it is not noticably higher, it seem to suggests that the test-drivers did in fact managed to utilized the EV to cover their regular driving needs. When looking at the weekday availability, two dips are clearly visible, resulting from the larger number of trips in the morning and afternoon, which are typically for people going to- and coming *home* from *work*. Since people generally go straight from *home* to *work*, but

have a tendency to run other errands afterwards, the afternoon dip is noticeable larger.

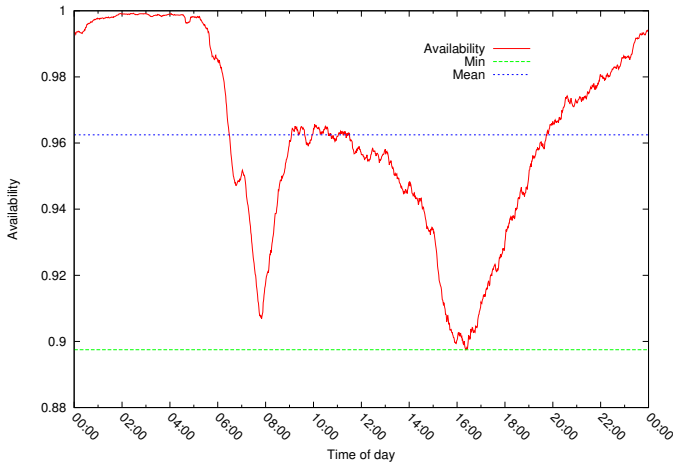

Fig. 2. Illustration of the overall availability on a weekday.

When observing the weekdays individually (not shown here), there is also a clear difference between Monday through Thursday and then Friday. Not surprisingly the weekend looks different, though still with an equally high availability. Since people are not going to *work*, but instead run errands throughout the day, there is only a single larger dip to be observed starting from around 10:00 and slowly leveling out throughout the late afternoon. As seen in Fig 3, which shows the weekday- and weekend availability overlaid for the sake of comparison, another noticeable difference is that weekday ends with a slightly higher availability. Considering that many people tend to visit friends, family and in general run late social errands, this is perhaps not surprising.

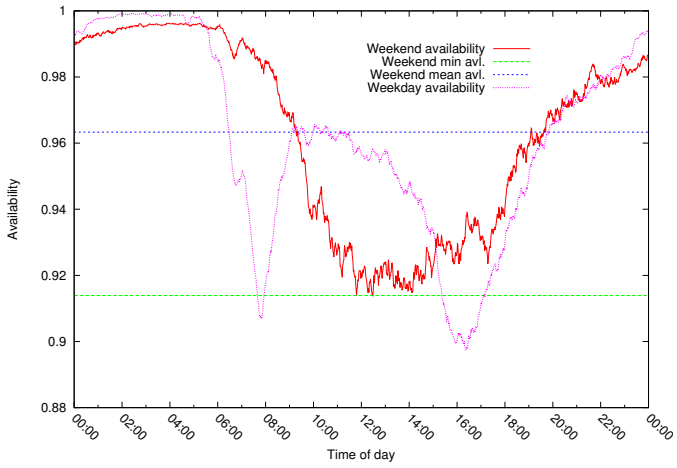

Fig. 3. Overlay comparing the weekday- and weekend availability.

B. Conclusions

Generally, the trends observed from the EV data seems to support that of previous ICE studies. Because the availability is nearly as high as for the ICE vehicles, the EVs appear to be able to cover most of the regular driving needs for the users. On the other hand, with the avialability being so high, it also bodes well for their reliability in relation to grid balancing. Since the vehicles in the project change hands every three months, it can be difficult to filter out any initial "enthusiastic" driving, which would differ from the normal routine. There is not much that can be done about that at this stage, but a record of their normal driving behavior, prior to receiving the keys for the EV, could be beneficial to work as a baseline. An interesting observation from the initial analysis of the dataset, was a somewhat large number of very short trips, lasting as little as e.g. 10 seconds. Initially it was thought to be a data-logging error, but when noticing similar trends in the logs for other EVs on the local campus, it was discovered that the likely reason is people briefly turning the ignition key to see the State Of Charge (SOC). Not counting online access, like what you will find in e.g. a Nissan Leaf, perhaps other OEMs should consider adding a function for retrieving this information without having to turn the key.

III. Geospatial Clustering

In order to be able to efficiently map the derived model to known addresses, the locations from the dataset have to be classified in order to identify if it is *home*, *work* or *other*. A dataset for a real-world vehicle will inevitably contain a lot of noise. In the case of GPS, assuming tracking itself is accurate, the noise in the data consists of scattered locations, which the vehicle visited few- or sometimes just once. figure 4 shows a sample of data for a vehicle, which has visited the same location several times.

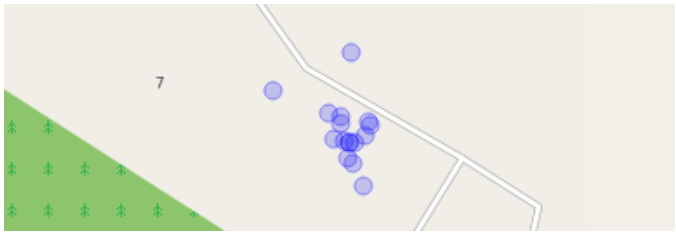

Fig. 4. Sample showing the error in precision typically observed in GPS measurements.

Because of this error induced dispersal it is necessary to perform an additional cluster analyses in order to determine which measured locations are really part of the same true location.

A. Clustering algorithms

The initial approach for deriving the number of clusters, i.e. the number of classified locations in the set, was to utilize the K-means algorithm [4]. The problem with the K-means algorithm is that it tries to force data into a pre-specified number of clusters, which is not ideal when this number is unknown. The resulting process is a brute force attempt at solving the problem, whereby testing for different numbers of clusters while observing an error-function, eventually results

in the actual number. Suffice to say this process can take a little time, especially when looking at larger dataset. K-means has another downfall, which is that it is not very well suited for classifying clusters of data that are not uniformly separated. This should, however, not constitute a problem when attempting to classify groups of GPS coordinates, as they will usually be grouped in a uniform manner. As a better alternative to K-means, the DBSCAN algorithm, was considered, which has a somewhat different approach to identifying clusters [4]. The algorithm, which is a so-called density based algorithm, takes two parameters: the base clustering radius (usually denoted ϵ), which defines when locations belong to the same cluster, and the minimum number of points required to form a cluster (hereafter referred to as *MP*).

B. Choosing the parameters

Ideally, a location like e.g. *home* would always be represented by a unique location, but unfortunately the nature of GPS tracking is that it is somewhat imprecise and sufferes from a variying degrees of error. A typical example of this can be seen in figure 4. One way to go could be to derive a standard spread from recorded coordinates for a stationary real-world object and use that to derive ϵ. However, to make things a little easier and assuming that clusters have a certain degree of separation, ϵ was fixed to 0.1 (100m) since the main clusters like *home* and *work* are unlikely to be located that close to each other anyway. The basic assumption is that people will choose not to drive between locations that are in such proximity.

TABLE I
RESULT OF DBSCAN ANALYSIS ON CHARGING LOCATIONS FOR DIFFERENT CLUSTER MINIMA (*MP*).

DBSCAN clustering example			
	$\epsilon = 0.1$ *MP*=5	$\epsilon = 0.1$ *MP*=15	$\epsilon = 0.1$ *MP*=25
Classified noise	246	395	433
Cluster #0	209	199	195
Cluster #1	164	164	159
Cluster #2	44	40	33
Cluster #3	31	22	
Cluster #4	15		
...			
Cluster #18	5		

With ϵ chosen, a series of trials were carried out to determine a suitable *MP* that would result in the largest clusters. Depending on the chosen ϵ value, a larger *MP* could be restricting the formation of smaller initial clusters, which will then not have the opportunity to merge and become even larger clusters. Table I shows the number and size of the clusters for various values of *MP*, which seems to support this hypothesis. As expected, the smaller *MP* value results in a larger number of clusters. If the value of *MP* is too great, the number of clusters shrink, since more locations are not allowed to form clusters and are instead labeled as noise. Because the goal is to identify *home*, *work* and *others*, it is ideal to have at least two dominating clusters. Looking at table I, the smaller *MP* seems a better fit. After the clusters have been formed, a heuristic is employed to select- and classify the desired clusters. Figure 5 illustrates the intended goal, namely to classify the largest clusters. The left of the figure shows the unfiltered locations, the middle the result from clustering and the rightmost the largest clusters.

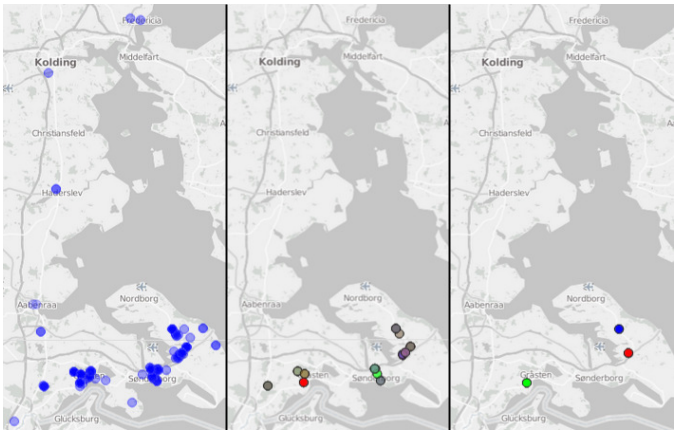

Fig. 5. Illustration showing first no clustering (left), then clustering with too small minima (middle) and lastly with suitable minima resulting in three distinct clusters (right).

Once the clustering analysis has been completed, the identified clusters have to be classified according to type. The heuristics used are as follows:

- The largest cluster and the one with the most stops spanning midnight is said to be *home*. If the two criteria results in different clusters, the dataset is abandoned in favor of better ones.
- The second largest cluster with an average stop-over starting time before noon is declared to be *work*. Of course some people work according to a non-A type schedule, but since it is tricky to verify the assumption here is "work before noon".
- The remaining cluster are characterized as *other*.

More complicated heuristics, to determine e.g. daycare were considered, but were left for future works to easy the initial implementation.

C. Conclusion

As seen in TABLE I, the scans with the smaller *MP* value seem to yield not only the greatest number of clusters, but also the largest individual clusters. The former is perhaps not that surprising, but one could be forgiven for expecting to see larger clusters for greater values of *MP* and not vice versa. Since a set of heuristics are employed for the final classification of the clusters, the smaller clusters will likely be filtered out, while the larger clusters remain. Because of this, the number of clusters is not as important as the size of the largest clusters, which is why a small *MP* is preferable.

IV. GENERATING THE MODEL(S)

A cumulative simulation often focuses on e.g. the total daily consumption and therefore settles for addressing when-

and how much the vehicles are driven and not where. The proposed method takes a slightly different approach; since one of the main goals is the ability to provide virtual loads as inputs to a grid simulation, the vehicles need to transition between their typical plug-in locations. The basic model is based on a three state Markovian inspired stochastic process, with the two main states being *home* and *work*. Initial attempts have thus far failed to yield a good heuristic for individually classifying the remaining clusters, so for now they have collectively been labeled *other*. Activities that could belong to this group are shopping, sports, visiting friends etc. Given the nature of the data, it would be intuitive to model the state transisions based on the probabilities for the real-world vehicles to move between the identified clusters. This does, however, have one or two drawbacks. Because the aim is a simualtion that behaves not just statistically correct, but also mimics real driving behavior, adhering strictly to using the transition probabilities could result in undesired behavior.

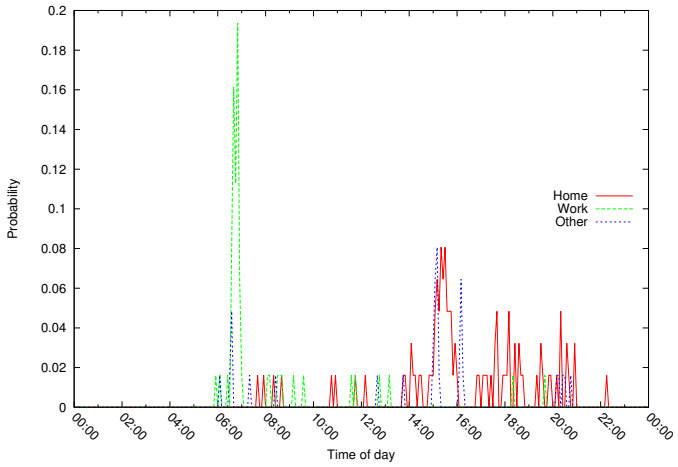

Fig. 6. Plot showing the probabilities of driving to the three primary locations.

Looking at figure 6, which shows the probabbilities of driving to the three different states respectively, one will observe a peak probability for driving to *work* in the morning. Later in the afternoon the probability of driving *home* increases, but starts to diminish from around the same time as the peak for driving to *other*. The behavior typically observed in real-life drivers is one that starts at *home* and ends at *home* again. With the probability of driving to *home* fading, so is the probability of the vehicle returning at the end of the day. For the real-life vehicles there are only a certain number of trips recorded on a daily basis, and most of them are recorded during the day. In fact, using the driving probabilities as basis for the model, it would become statistically unlikely that the vehicle would return *home* after approximately 22:00. Another less than realistic behavior that can arise from using the trip probabilities, mainly due to the greater fluctuations, is that the vehicle is more likely to jump eradically between locations. For example most people will arrive at work in the morning, will stay there for 7-8 hours and rarely leave to go shopping for an hour at random times during the day.

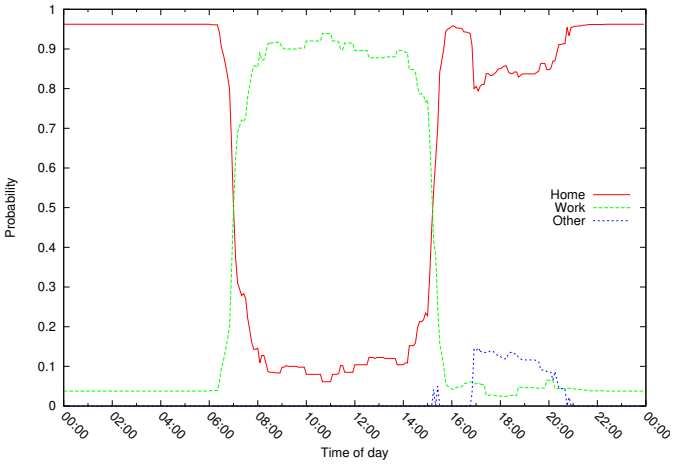

Fig. 7. Plot of the probabilities for being in the state of *home*, *work* and *other* for any given time during an average weekday.

As a more dependable alternative, the probabilities for being in a certain location at a given time of the day, were considered. The plot in figure 7 shows this probability for the three main states and compared to that of figure 6, it better reflects the intended behavior. The likelyhood of the vehicle being home in the beginning of the day is very high. At roughly 7-8:00 there is a shift from *home*- to *work* being the most dominant. In the late afternoon the probability of going elsewhere is highest and finally there is a much higher probability of the vehicles returning *home* at the end of the perceived workday. Since nearly all drives in figure 7 result in the vehicle returning *home*, it works much better as a basis for the model.

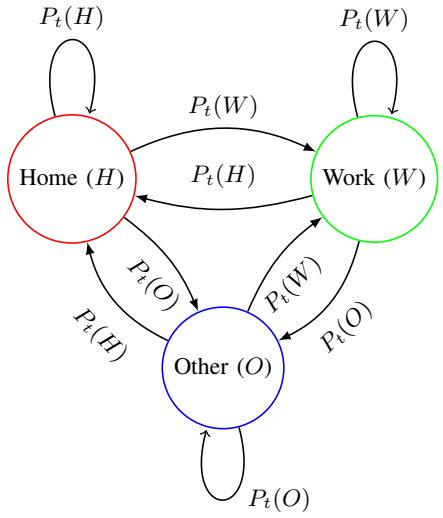

Fig. 8. Illustration of the stochastic model for the vehicle based on state likelyhood.

By using the state probabilities, the vehicle is essentially forced to return to *home* and is much less likely to stay out overnight. Because of a few observed late night trips to work

in the original data, there is still a chance of a late night stay at *work*, but it is negligible. Figure 8 shows the three state model based on the state probabilities, where $P_t(S)$ denotes the probatility that a transition to state S will occur at time t. Once a decision has been made to transition to a certain state, a route should be calculated, along which the vehicle can start to drive. The route provides speed and duration and it is not until the vehicle arrives at the destination, that it should be allowed to once again transision. A way, to reduce computational overhead when simulating large amount of vehicles, could be to determined the duration of the given stop-overs and essentially stop the simulation for that period. This could be done stochastically based on the density distribution for the given location, arriving at the given time. It does, however, not help with the stranding issue as the needed transisional probability could have decreased during the stop-over.

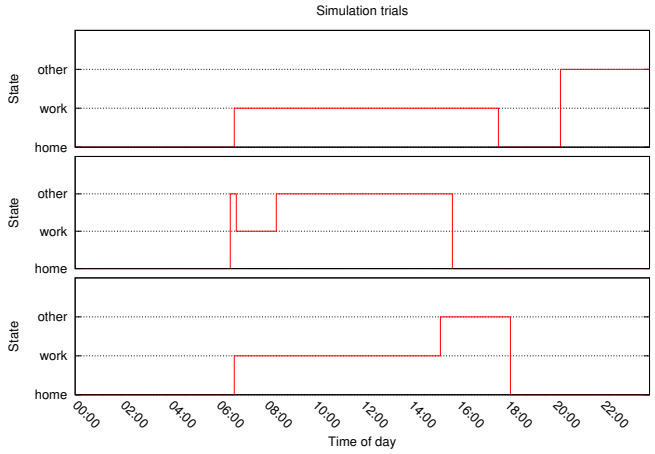

Fig. 9. Output of a random selection of weekday simulation trials.

The plot in figure 9 shows a series of simulation trials for the probabilities plotted in figure 7. With the exception of the end of the first trial, where the vehicle seems to stay out past midnight, the observed patterns appear to reflect a realistic behavior.

V. Conclusion

This paper suggests- and discusses a method for deriving a more realistic driving pattern, with the purpose of individually simulating large amount of EVs and their respective position throughout the day. This is achieved by cluster analyses of existing EV driving data, to group and classify the most typical types of stop-overs. Following this, a statistical analysis of the data is used to feed into a Markov inspired stochastic model used in the simulation.

A. Future work

Grid models have already been developed for the island of Bornholm, which has been the Smart Grid focus point in Denmark for a while [9][10]. Linking those models with the EV simulation would make for an ideal framework for large scale testing of everything from charging algorithms to grid-centric ancillary services. Another topic, that is rarely tested in practice, is the true scalability of existing- and proposed communication solutions for smart grid applications. Many projects and studies have dealt with various control algorithms, but few tried to tackle the pratical issues that arise with growing penetrations [5][12][13]. In a cumulative simulation the consumption would normally be determined by random trials of driving and parking, likely coupled with an average, EV consumption over distance. Several EV studies have been carried out, mostly arriving at 150Wh/km as a good average but this may not be ideal when simulating individual vehicles. Early trials used online routing services such as the Microsoft Bing Service [8]. While excellent and easy to program against, a dependence on an online service is not feasible when the simulation will result in thousands of route queries being made again and again. A promising alternative is the Open Source Routing Machine, which as the name suggests is a free routing solution [7]. It is based on the Open Street Maps [6] from which its highly optimized algorithm can find a way from A to B in mere milliseconds. While routing on a global scale is definitely possible, a lighter alternative is to simply download a subsection of the map representing the simulation area.

VI. Acknowledgement

The authors would like to thank Clever A/S for use of the data collected during the “Test-en-elbil” project. Since an initial filtering had already been performed by Clever A/S, the needed processing time was reduced, freeing time to focus on more essential areas of the analysis.

References

[1] Q. Wu, A.H. Nielsen, J. Østergaard, S.T. Cha, F. Marra, Y. Chen, C. Træholt, *Driving Pattern Analysis for Electric Vehicle (EV) Grid Integration Study*, IEEE ISGT Europe, 2010.
[2] P. Nørgaard, L. Christensen, *Estimated Impact of the Uncertainties in the Driving Pattern on the Power System Flexiblity Provided by Electrical Vehicles*, 5th Nortic Wind Power Conference, 2009.
[3] “Test-en-elbil” (“Test-an-EV”), http://www.clever.dk/test-en-elbil.
[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st. ed. Springer Oct 2006.
[5] A. B. Pedersen, E. B. Hauksson, P. B. Andersen, B. Poulsen, C. Træholt, D. Gantenbein, *Facilitating a generic communication interface to distributed energy resources: Mapping iec 61850 to restful services, in Smart Grid Communications*, IEEE SmartGridComm, 2010.
[6] Open Street Maps, http://www.openstreetmaps.org.
[7] Open Source Routing Machine, http://www.project-osrm.org
[8] Microsoft Bing Routes, http://msdn.microsoft.com/en-us/library/ff701705.aspx.
[9] The EDISON project, http://www.edison-net.dk.
[10] The EcoGrid EU project, http://www.eu-ecogrid.net.
[11] J. M. Jørgensen, S. H. Sørensen, K. Behnke, P. B. Eriksen, *EcoGrid EU - A Prototype For European Smart Grids.*, IEEE PES General Meeting, 2011.
[12] B. Pedersen, D. Winther, A. Pedersen, B. Poulsen, C. Træholt, *Integrating Intelligent Electric Devices into Distributed Energy Resources in a Cloud-Based Environment*, IEEE ISGT Europe, 2013.
[13] L. D. Orda, J. Bach, A. B. Pedersen, B. Poulsen, L. H. Hansen, *Utilizing a Flexibility Interface for Distributed - Energy Resources Through a Cloud-Based Service*, IEEE SmartGridComm, 2013.
[14] R. Li, Q. Wu, O. S. Schmul, *Distribution Locational Marginal Pricing for Optimal Electric Vehicle Charging Management*, IEEE Transactions on Power Systems, 2013.