Towards Predicting First Daily Departure Times: a Gaussian Modeling Approach for Load Shift Forecasting

Nicholas H. Kirk  
Institute of Automatic Control Engineering  
Technical University of Munich, Germany  
nicholas.kirk@tum.de

Ilya Dianov  
Department of Informatics  
Technical University of Munich, Germany  
ilya.dianov@tum.de

Abstract—This work provides two statistical Gaussian forecasting methods for predicting First Daily Departure Times (FDDTs) of everyday use electric vehicles. This is important in smart grid applications to understand disconnection times of such mobile storage units, for instance to forecast storage of non dispatchable loads (e.g. wind and solar power). We provide a review of the relevant state-of-the-art driving behavior features towards FDDT prediction, to then propose an approximated Gaussian method which qualitatively forecasts how many vehicles will depart within a given time frame, by assuming that departure times follow a normal distribution. This method considers sampling sessions as Poisson distributions which are superimposed to obtain a single approximated Gaussian model. Given the Gaussian distribution assumption of the departure times, we also model the problem with Gaussian Mixture Models (GMM), in which the priorly set number of clusters represents the desired time granularity. Evaluation has proven that for the dataset tested, low error and high confidence (≈95%) is possible for 15 and 10 minute intervals, and that GMM outperforms traditional modeling but is less generalizable across datasets, as it is a closer fit to the sampling data. Conclusively we discuss future possibilities and practical applications of the discussed model.

Index Terms—Times forecasting, First Daily Departure Times, Vehicle-to-Grid integration, Gaussian modeling, Gaussian Mixture Models, Grid load shifting

I. INTRODUCTION

With an increasing use of Plug-in Electric Vehicles (PEVs), mobile units can be seen as a potential grid-connected energy storage means without compromising their primary mobility functionality: A PEV fleet can store, for instance, power from non dispatchable loads (e.g. solar panel and wind turbine sources) [1]. However, connections of PEVs to the grid, in terms of times and locations, are complex to model given such logistic mobility. This work focuses on how to meaningfully model fleet-level departure times over the commuter time frame 6 am - 9 am, in order to predict the availability of PEVs as grid storage over time. Heuristic assumptions such as over- or under-estimation of arrival/departure times both suffer from shortcomings and will result in inefficient energy use: an accurate forecast is therefore of paramount importance. We exploit First Daily Departure Times (FDDT), which are a key piece of information for connection time estimation in PEV load shifting algorithms [2], but are hard to predict using historical realizations alone or via basic distribution modeling [3]. This research focused on understanding how to accurately predict PEV FDDT for successful load shift scheduling. Such accuracy analysis was first performed via preliminary feature correlation analysis with FDDT (Section III-B), thanks to the availability of a dataset with diverse driving behavior features, which contain, for instance, information sampled from individual drivers regarding average trip length and duration (Section III-A). Given the lack of forecasting capability of such features towards FDDT prediction, the research then makes progress towards approximated Gaussian modeling under specific a priori running assumptions (Section III-C). We provide theoretical background (Section IV-A) to a method for computing lower and upper bounds of PEV FDDT for each time interval (Section IV-B), a time interval scaling method (Section IV-C), and provide a brief validation of such methods (Section V). Conclusively, we provide a summary on the proposed method and possible insights regarding future work (Section VI).

II. RELATED WORK

Previous studies take into account aggregations of driver’s behavior features for activity-based forecasting, aiming at Transportation Demand Management (TDM) [4], congestion planning or logistic network optimality [5]. In particular, behavior aggregation has been useful to understand the actions that provoke inter-relations among individuals, in order to cluster vehicle movement by activity [4].

In [6] the FDDT prediction is based on utility maximization of the vehicle trip and activity participation. The activities are defined as driver intentions such as “being home before work”, while trip is characterized by departure and arrival times. Another method [7] uses a multilevel approach which claims that FDDT is dependent on individual attributes such as gender, age, profession and macro-level attributes such as day of the week, location and household income. Each attribute is modeled using normal distributions and the prediction is based on log likelihood maximization. However both these approaches require private (usually unavailable) information about each driver (e.g. type of activities engaged in after
work). Goedel [3] provides different approach which takes into consideration the day of the week as feature and a vehicle-based analysis of commuters, in order to predict a departure confidence interval. In other work [8], charging profile predictions are based on stochastic analysis of the conditional Probability Density Function (PDF) over FDDT, daily arrival times and daily traveled distances. However, both methods provide a one hour interval precision of FDDT which is not sufficient within the domain of load shift prediction. Given a low correlation among FDDT and driver behavior features, the presented research focused on Gaussian modeling of FDDT data only. Furthermore, the advantage of considering only first daily departure times is that the research can disregard the complexity entailed by modeling the multiple stop factor.

III. DATA UNDERSTANDING

We now describe the reasoning behind the adoption of the training and test set (Section III-A), the feature correlation analysis (Section III-B), and the set of assumptions that are required for this statistical modeling problem (Section III-C).

A. Data Adoption

This project makes use of datasets from NREL’s Secure Transportation Data Project [9], in particular Texas Department of Transportation - Transportation Studies with GPS Travel Diaries.

The main reasons for such adoption are:

- the dataset comprises many real-time features of the trips (e.g. interval times, speeds, accelerations, statistical measures - see Table I)
- the features present high precision and low granularity
- all data has been electronically tracked
- given the geographical location (Texas), we assume climate variability to be low and therefore not influencing departure times

A major downfall of the dataset is that it does not comprise labeling for the day of the week, and furthermore all samplings have been performed only on Tuesdays or Wednesdays.

| Feature                        | Description                                                                 |
|-------------------------------|-----------------------------------------------------------------------------|
| start_tm                      | The start time of the first recorded point for the vehicle                  |
| distance_total                | Total travelled distance in miles                                           |
| percent_fifty_five_sixty     | Percent of total time spent at speeds between fifty five and sixty miles per hour |
| driving_speed_standard_deviation | Standard deviation of driving speed distribution                        |

Table I: Listing and descriptions of examples of features present in the NREL Transportation dataset [9].

B. Feature Analysis

Correlation among potential features and the class to predict (FDDT) is a necessary but not sufficient condition for pattern learning. We analyzed the potential predictive ability of each feature by executing a Correlation-based Feature Subset Selection [10] and a correlation-based Principal Component Analysis (PCA) [11], making use of an Independent and Identically Distributed (IID) assumption. Such feature filters yielded a very low correlation between features and data, making these unserviceable for machine learning (see Table II).

| Correlation-based selected features                                      | Correlation with FDDT (start_tm) |
|-------------------------------------------------------------------------|----------------------------------|
| total_speed_velocity_ratio                                              | +0.15                            |
| percent_distance_fifty_five_sixty                                       | -0.21                            |
| absolute_time_duration_hrs                                              | -0.3                             |
| descending_rate_median_absolute_deviation                               | -0.08                            |
| max_deceleration_event_duration                                         | -0.33                            |
| average_acceleration_event_duration                                     | +0.04                            |
| min_deceleration_event_duration                                         | -0.05                            |

Table II: Features with the highest correlation with FDDT (start_tm), and with the lowest correlation among themselves. For a description of the cited features, see [9].

C. Assumptions

An initial intuition after viewing the variety of available features in the dataset (Section III-A) would suggest the possibility of performing high-dimensional regression with such diverse components. However, this approach is not possible with the current dataset, since features presented very low correlation values and hence low or no learning potential (see Table II). Therefore we proceed in assuming that every sample is Independent and Identically Distributed (IID), i.e. that the FDDT of a vehicle does not influence the FDDT of another. Consequently, we do not consider the problem as a time-series analysis as understood in literature [12]. Given such information, instead of predicting the exact departure time of the sample, it is more convenient to forecast, given historical values, i) how many FDDT will fall within certain time intervals, ii) the confidences of the latter, and iii) a system-level interval granularity itself. By empirical analysis and by assumption we define that the FDDT sampling undertaken for the training dataset is distributed according to Poisson’s definition.

IV. APPROXIMATED GAUSSIAN MODELING

We proceed in describing a method to constrain our problem to the Gaussian modeling domain (Section IV-B, IV-C, given the assumptions in Section III-C).

A. Theoretical Framework

1) Poisson distribution: In probability theory, the discrete Poisson distribution expresses the likelihood of a number of events occurring sequentially and independently of each other within a given time frame, knowing that on average a given number λ occurs. For an in-depth description of the mathematical properties of this distribution, we refer to [13]. We exploit the mathematical property that for a hypothetically infinite number of samplings, the superimposition of such Poisson distributions converges to a Gaussian (Normal) distribution
Due to the latter property, it is then possible to model with traditional Gaussian assumptions.

2) Gaussian distribution: A Gaussian (Normal) is a continuous probability distribution that often characterizes real-valued random variables in applied contexts. In this context we model the distribution over time intervals and their confidence. The latter is possible by defining the percentage of values captured by a distance $k\sigma$, where $\sigma^2$ is variance from the mean $\mu$, as seen in [1].

$$\int_{-k\sigma}^{+k\sigma} N(\mu, \sigma^2)$$  \hspace{1cm} (1)

For $k = 2$ we obtain confidence of $\approx 95\%$. For a deeper mathematical description we refer to [13].

B. Computing Time Intervals

By aggregating the timestamp samples of a single sampling session in discrete time intervals, we obtain a Poisson distribution. If we sample a sufficiently large dataset, or a heterogeneous set of sampling sessions, the superimposition of these Poisson distributions converge to an approximated Gaussian distribution [13].

Let $n$ be the index of the current sampling session and $N$ the total number of samplings which have been operated. Let $b$ be an arbitrary constant that defines the number of bins (and therefore the time interval granularity). We then construct a matrix $K$:

$$K^b_n = \begin{pmatrix} K_0^b & \cdots & K_b^b \\ K_1^b & \cdots & K_b^b \\ \vdots & \ddots & \vdots \\ K_n^b & \cdots & K_b^b \end{pmatrix} \hspace{1cm} (2)$$

We compute an estimation of the lower bound and the upper bound of the number of PEV departures in the time interval $t_i \in 0,\ldots,b$ with a probability of 95%:

time interval margins $t_i = \left[ m^i - 2\sqrt{m^i}, m^i + 2\sqrt{m^i} \right] \hspace{1cm} (3)$

where:

$$m^i = \frac{1}{n} \sum_{j=0}^{n} K_{ij} \hspace{1cm} (4)$$

C. Granularity Scaling

We want to hypothetically increase $b$ in order to have a time prediction interval as small as possible. We define the error $\epsilon_{\%}$ as the wanted percentage error of our confidence interval.

We obtain the lowest time granularity possible without lowering the given confidence interval by imposing that:

$$\epsilon_{\%} \leq \frac{m^{{min}}}{\sqrt{m^{{min}}}} \hspace{1cm} (5)$$

where $0 \leq \epsilon_{\%} \leq 1$ and:

$$m^{{min}} = \min(m^0 \ldots m^b) \hspace{1cm} (6)$$

An overview of the entire modeling here discussed in Sections IV-B and IV-C can be viewed in Algorithm 1.

**Algorithm 1: PEV departure number within time interval computation**

**Data:**
- $\text{minDepTime}$, earliest departure time
- $\text{maxDepTime}$, latest departure time
- $TD$, a training set containing $n$ sampling sessions of departure times
- $\text{CIvalue}$, the percentage value of the desired confidence interval

**Result:** $\text{TimeIntMargins}$, PEV departure number for each time interval

**begin**

$\epsilon_{\%} \leftarrow (1 - \text{CIvalue})$

$TD_{cut} \leftarrow \text{trimRange}(TD, \text{minDepTime}, \text{maxDepTime})$

$\epsilon_{\%}^i \leftarrow \left( \min_i \left( M_{\text{matrix}^i} \right) / \sqrt{\min_i \left( M_{\text{matrix}^i}^2 \right)} \right)$

$\text{increase } b$

$K_{\text{matrix}} \leftarrow \text{divideInIntervals}(TD_{cut}, b)$

$M_{\text{matrix}} \leftarrow \text{imposeAndAvg}(K_{\text{matrix}}, b, n)$

$\text{while } j < b \text{ do}$

$\text{intMargins}^j \leftarrow \text{compMargins}(M_{\text{matrix}}^j)$

$output \text{intMargins}^j$

$\text{increase } j$

**end**

D. Expectation-Maximization for Gaussian Mixture Models

Given the Gaussian assumption used throughout this text, for which all first time departures follow a Normal distribution, we made use of Gaussian Mixture Models (GMM) to cluster FDDTs, in which each cluster represents a bin as described in Section IV-B. We model time intervals as a Gaussian distribution, and the time inside each interval is additionally characterized by a Gaussian distribution. For this we use a mixture model with $K$ components where each component is a multivariate Gaussian density:

$$g_i(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \hspace{1cm} (7)$$

where $\mu_i$ is a mean, $\Sigma_i$ is a covariance matrix, $x \in D$, where $D$ is a given dataset, $i = 1,\ldots,K$. In order to learn unsupervisedly the parameters of the latent models characterizing the multivariate Gaussian distribution, we make use of the iterative Expectation Maximization (EM) algorithm [13], which finds the maximum likelihood of parameters also with low resolution distribution data, such as in the case of our approximated Gaussian model derived from superimposed Poisson distributions. For a more detailed description of EM for GMM, we refer to [15].
V. Validation Results and Model Usability

![Graphical representation of the confidence intervals obtained with the approximated Gaussian modeling described in Section IV-B](image)

Validation was not possible with the current running assumptions, given the reduced number of samples of our training data after time frame (6am–9am) pruning. In order to increase the number of training samples, we superimposed the pruned sets of different cities which share similar urban and climatic characteristics (namely Austin, San Antonio, Houston, El Paso) to create a training set, and then used the superimposed Gaussian model to validate on a test set formed by samples from Rio Grande Valley. The training set for all three incoming experiments contained 758 instances of data for the time period from 6:00 am to 9:00 am, and the test set for all experiments contained 260 instances of data for the same time period. The set was normalized for explanatory convenience (i.e. every bin defines the percentage of vehicles that depart in the bin time interval). Statistical dispersion was computed with Gauss error function (erf) for our validating model which evaluates the probability that measurement $x$ is within a range from $-\frac{x}{\sigma \sqrt{2}}$ to $\frac{x}{\sigma \sqrt{2}}$. To compute erf we used a following formula:

$$erf(x) = \frac{x}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

\hspace{1cm} (8)

a) First experiment (approximated Gaussian modeling, 12 timeframes): The model derived from the theory presented in Section IV-B validated the model on all bins. Margin computations are shown in Fig. 1 while validation results can be seen in Fig. 2. The resulting erf for such approximated Gaussian modeling is shown in Table III.

The implementation of Algorithm 1 focused on granularity understanding and margin computation, in which we can compute a trade-off between estimation confidence (via modeling the $k$ parameter in Eq. 1), and time interval granularity (i.e. the number of bins).

| Time intervals | Average margin values | erf values |
|----------------|-----------------------|------------|
| 6:00-6:15am    | 10.084                | 0.0955     |
| 6:15-6:30am    | 0.0729                | 0.0821     |
| 6:30-6:45am    | 0.0756                | 0.0851     |
| 6:45-7:00am    | 0.0809                | 0.0911     |
| 7:00-7:15am    | 0.1088                | 0.1223     |
| 7:15-7:30am    | 0.1207                | 0.1355     |
| 7:30-7:45am    | 0.1300                | 0.1459     |
| 7:45-8:00am    | 0.1074                | 0.1207     |
| 8:00-8:15am    | 0.0849                | 0.0956     |
| 8:15-8:30am    | 0.0504                | 0.0568     |
| 8:30-8:45am    | 0.0570                | 0.0642     |
| 8:45-9:00am    | 0.0531                | 0.0599     |
| Average erf value | 0.0938          |            |
| Normalized score on a number of bins | 1.1256          |            |

Table III: Gauss error function results for the approximated Gaussian model on average margin values (12 bins, 15 minutes each).

b) Second experiment (Gaussian Mixture Modeling, 12 timeframes): For the EM method we used training data containing only FDDTs (start tm) for different vehicles as feature and validated it on the previously mentioned test data (Rio Grande Valley set). Each instance of the dataset is associated with a single vehicle and the resulting model of the EM algorithm, shown in Fig. 3, illustrates the dependency between departure times and number of vehicles departing at the particular time interval. We modeled 12 clusters, i.e. 12 timeframes of 15 minutes each. The model shows that the highest amount of vehicles was departing at 7:00 am - 7:15 am which makes this interval the most probable for future predictions under the current assumptions. The resulting erf values are shown in Table IV. The average erf values are 0.094 and 0.07 for approximated-Gaussian modeling and EM for GMM respectively (i.e. first and second experiment).

c) Third experiment (Traditional and Gaussian Mixture Modeling, 18 timeframes): We repeated both the EM method for Gaussian Mixture Modeling and traditional approximated Gaussian modeling, this time with 18 timeframes, i.e. 10 minute intervals, using only FDDTs (start tm). Results are visible in Fig. 4 and 5, while Gaussian error values are displayed in Table V and VI. Overall, given the graphical and error results, we can confirm the Gaussian assumption on such real data model. We can observe that approximated Gaussian
modeling preserves the form factor across departures per time intervals, whereas GMM provide a closer approximation to the training data.

VI. CONCLUSIONS AND FUTURE WORK

Modeling First Daily Departure Times (FDDT) of electric vehicles is of paramount importance for smart grid load shift planning, as these can be used as temporary energy storage units. By making a Gaussian distribution assumption of such departure times, we have provided a i) traditional Gaussian modeling approach with confidence and time interval size modeling, and ii) a Gaussian Mixture Model approach to compute clusters associated to time intervals. Evaluation has proven that for the dataset tested, low error and high confidence ($\approx 95\%$) is possible for 15 and 10 minute intervals. By inspection of the normalized score on a number of bins for both methods (Fig. 6), we notice that GMM method is more subject to error when increasing time interval granularity, but requires less data to formulate the model. Future work will be oriented towards testing the presented Gaussian model on large datasets, implementing error propagation when relaxing the IID assumption (i.e. assuming that all cars depart), and considering confidence and error trade-off for practical applications. Furthermore a collaboration with transport survey research centers would be useful to gather more vehicle-related and activity-related data, in order to cluster by the latter and by points of interest, to then verify feature correlation with FDDTs.

ACKNOWLEDGEMENTS

The authors thank Christoph Göbel (Technical University of Munich, Germany) for his valuable feedback on an initial version of this paper.

REFERENCES

[1] W. Kempton and J. Tomić, “Vehicle-to-grid power implementation: From stabilizing the grid to supporting large-scale renewable energy,” Journal of Power Sources, vol. 144, no. 1, pp. 280–294, 2005.
[2] C. Goebel and D. S. Callaway, “Using ICT-controlled plug-in electric vehicles to supply grid regulation in California at different renewable integration levels,” Smart Grid, IEEE Transactions on, vol. 4, no. 2, pp. 729–740, 2013.
[3] C. Goebel and M. Voß, “Forecasting driving behavior to enable efficient grid integration of plug-in electric vehicles,” IEEE GreenCom, pp. 74–79, 2012.
[4] R. Kitamura, “Applications of models of activity behavior for activity based demand forecasting,” Activity-Based Travel Forecasting Conference, New Orleans, Louisiana, 1996.
[5] R. H. Emmerink, K. W. Axhausen, P. Nijkamp, and P. Rietveld, “The potential of information provision in a simulated road transport network with non-recurrent congestion,” Transportation Research Part C: Emerging Technologies, vol. 3, no. 5, pp. 293–309, 1995.
[6] D. Ettema and H. Timmermans, “Modeling departure time choice in the context of activity scheduling behavior,” Transportation Research Record: Journal of the Transportation Research Board, vol. 1831, no. 1, pp. 39–46, 2003.
[7] M. Chikaraishi, A. Fujiwara, J. Zhang, and K. W. Axhausen, “Exploring variation properties of departure time choice behavior by using multilevel analysis approach,” Transportation Research Record: Journal of the Transportation Research Board, vol. 2134, no. 1, pp. 10–20, 2009.
[8] A. Ashtari, E. Bibeau, S. Shahidinejad, and T. Molinski, “Pev charging profile prediction and analysis based on vehicle usage data,” Smart Grid, IEEE Transactions on, vol. 3, no. 1, pp. 341–350, March 2012.
[9] J. Gonder, E. Burton, and E. Murakami, “Establishing a secure data center with remote access,” Federal Committee on Statistical Methodology (FCSM) Research Conference, 2012.
[10] M. A. Hall, “Correlation-based feature subset selection for machine learning,” PhD diss., University of Waikato, 1998.
Figure 6: Gauss error function (erf), normalized over the number of bins, across both GMM and approximated Gaussian models for 10 minute time intervals (left), and 15 minute intervals (right).

| Time intervals | Predicted values | erf values |
|----------------|------------------|------------|
| 6.00-6.10am    | 0.0269           | 0.0304     |
| 6.10-6.20am    | 0.0231           | 0.0260     |
| 6.20-6.30am    | 0.0385           | 0.0434     |
| 6.30-6.40am    | 0.0423           | 0.0477     |
| 6.40-6.50am    | 0.0692           | 0.0780     |
| 6.50-7.00am    | 0.0577           | 0.0650     |
| 7.00-7.10am    | 0.0731           | 0.0823     |
| 7.10-7.20am    | 0.1090           | 0.1125     |
| 7.20-7.30am    | 0.0846           | 0.0953     |
| 7.30-7.40am    | 0.1500           | 0.1680     |
| 7.40-7.50am    | 0.1038           | 0.1168     |
| 7.50-8.00am    | 0.0769           | 0.0886     |
| 8.00-8.10am    | 0.0192           | 0.0217     |
| 8.10-8.20am    | 0.0500           | 0.0564     |
| 8.20-8.30am    | 0.0385           | 0.0434     |
| 8.30-8.40am    | 0.0154           | 0.0174     |
| 8.40-8.50am    | 0.0115           | 0.0130     |
| 8.50-9.00am    | 0.0152           | 0.0174     |
| **Average erf value** | **0.0625** |            |
| **Normalized score on a number of bins** | **1.1250** |            |

| Time intervals | Predicted values | erf values |
|----------------|------------------|------------|
| 6.00-6.10am    | 0.0451           | 0.0509     |
| 6.10-6.20am    | 0.0411           | 0.0464     |
| 6.20-6.30am    | 0.0451           | 0.0509     |
| 6.30-6.40am    | 0.0504           | 0.0568     |
| 6.40-6.50am    | 0.0557           | 0.0628     |
| 6.50-7.00am    | 0.0504           | 0.0568     |
| 7.00-7.10am    | 0.0756           | 0.0851     |
| 7.10-7.20am    | 0.089            | 0.0977     |
| 7.20-7.30am    | 0.0849           | 0.0956     |
| 7.30-7.40am    | 0.0915           | 0.1030     |
| 7.40-7.50am    | 0.0822           | 0.0925     |
| 7.50-8.00am    | 0.0637           | 0.0718     |
| 8.00-8.10am    | 0.0584           | 0.0688     |
| 8.10-8.20am    | 0.0438           | 0.0494     |
| 8.20-8.30am    | 0.0332           | 0.0374     |
| 8.30-8.40am    | 0.0398           | 0.0449     |
| 8.40-8.50am    | 0.0292           | 0.0329     |
| 8.50-9.00am    | 0.0311           | 0.0464     |
| **Average Gaussian error (erf) value** | **0.0626** |            |
| **Normalized score on a number of bins** | **1.1268** |            |

Table V: Gauss error function results for the GMM model validated against Rio Grande Valley values (18 bins, 10 minutes each).

Table VI: Gauss error function results for Approximate Gaussian model validated against Rio Grande Valley values (18 bins, 10 minutes each).

[11] I. Jolliffe, “Principal component analysis,” Springer Science & Business Media, 2002.
[12] G. E. Box, G. M. Jenkins, and G. C. Reinsel, “Time series analysis: forecasting and control,” John Wiley & Sons Ltd, vol. 734, 2011.
[13] F. A. Haight and F. A. Haight, “Handbook of the poisson distribution,” Ann. Math. Statist., vol. 40, no. 1, pp. 326–327, 1967.
[14] R. M. Dudley, “Real analysis and probability,” Cambridge University Press, vol. 74, 2002.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” Journal of the royal statistical society. Series B (methodological), pp. 1–38, 1977.