Towards Electric Mobility Data Mining

Timo Duchrow*†, Martin Schröer*, Britta Griesbach*, Sebastian Kasperski*, Fabian Maas genannt Bermpohl*, Stefan Kramer† and Frank Kirchner*‡
*German Research Center for Artificial Intelligence (DFKI GmbH), Robotics Innovation Center, Bremen, Germany
†Technische Universität München, Institut für Informatik/I12, Garching, Germany
‡University of Bremen, Robotics Group, Bremen, Germany

***Abstract*—With the advent of electric mobility, new challenges arise for car makers, utility companies, car sharing ventures, and policy makers in planning, deployment, and management of electric vehicle infrastructures. To this end, efficient systems for the large scale acquisition, management, and analysis of mobility and consumption data become an enabling factor. We present a framework for collecting heterogeneous fleet-related data to generate reports about vehicle and fleet usage. Using this framework, we present an analysis of trips and charge intervals from one year of telemetry data gathered from a subset of a pilot electric vehicle fleet in northern Germany. We further investigate which factors are associated with differences of relative levels of energy usage per kilometer.**

I. Introduction

For policy makers and companies with interest in electric mobility, the uncertainty about how large fleets of electric vehicles in use by end users will behave in terms of energy needed, regional distribution of charging points, vehicle availability, and mileage represents a major impediment to commit themselves to EV fleet investments. The prognosis of usage statistics for the large scale deployment of electric vehicles (EVs) remains a hen and egg problem: Reliable data of EV user behavior can only be gathered by sufficiently large fleets which remain risky investments if sufficient data for due diligence work is not available. To address this problem, the model region Bremen/Oldenburg, one of the eight EV model regions in Germany, has installed a fleet trial in which private and commercial users are given EVs and data is recorded from the fleet. The long term goal of the model region includes the prediction of the fleet's energy storage capacity, which has been suggested to be used as a means of stabilizing the power grid for fluctuating renewable energy sources [1], [2], [3].

The data that is recorded during the fleet trial may provide valuable information for car manufacturers and consumers, since usual vehicle testing procedures follow standard driving cycles (such as the NEDC or the FTP-75) and do not take into account driving behavior influenced by the necessities of electric mobility (e.g., state-of-charge, availability of charging infrastructures, etc.). During the fleet trial it is expected that actual usage patterns will generate additional insights into operation parameters of EVs.

The EV model region Bremen/Oldenburg is funded by the German Federal Ministry of Transport, Building and Urban Development (BMVBS). The funding program coordination is carried out by the NOW GmbH National Organization Hydrogen and Fuel Cell Technology (Förderkennzeichen 03ME0400G). TD is supported by a grant from the German National Academic Foundation (Studienstiftung des deutschen Volkes).

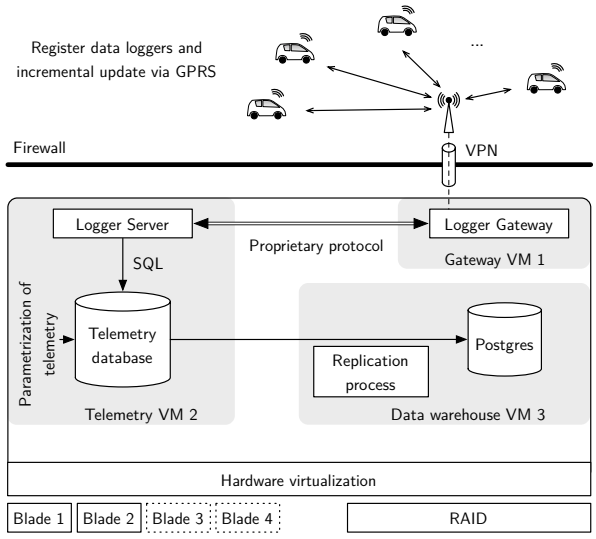

Fig. 1. Architecture of the data acquisition framework. EVs are fitted with data loggers attached to the CAN bus that transmit data via a GPRS connection to a server stack where data is parsed and consolidated in a data warehouse. Parametrization of data loggers can be performed remotely (e.g., to configure which CAN values are to be transmitted via the network).

Other work has been carried out to evaluate the impact of EVs on several aspects of the ecosystem, the economy, and daily life itself using fleet trials. The Electric Power Research Institute (EPRI) started a fleet trial of 60 GM Volt plug-in hybrid EVs (PHEVs) in 2011, mainly to assess performance and customer satisfaction [4]. There are other approaches that focus on PHEVs in small scale trials [5], [6]. The REV Project is a long-term fleet trial, focusing on standard cars that are converted to EVs and the use of renewable energy sources [7]. The Öko-Institut has analyzed a fleet trial at SAP AG where a number of EVs are in use as part of the company's vehicle fleet. The analysis focuses mainly on CO_2 reduction that can potentially be achieved given different usage profiles within a company's vehicle fleet [8].

In the remainder of this paper, we will first give a brief overview of the infrastructure for data acquisition and an outline of what types of data are gathered. We will then present the major preprocessing steps that are performed to simplify

the subsequent data analysis. Finally, we will describe the actual analysis and some results that were obtained so far.

II. Method

A. Data Acquisition

Data acquisition protocols are usually designed to suit the EV model type and research question at hand. However, the EV product landscape is changing rapidly, as are the type of questions asked while new research alliances are being formed. We therefore advocate a different approach to EV fleet data management. We present a solution that is mostly agnostic of the vehicle type and can be used to record data from heterogeneous EV fleets. Also, we aim to record as much data as possible by tapping into the vehicles' CAN bus and consolidating the data in a central repository.

A stream model of data processing is then employed to analyze the data with respect to specific research questions. Here, we present an analysis of trip distances, battery charge consumed per trip, and charging intervals.

The model region's vehicle fleet comprises a heterogeneous set of 44 vehicles including transporters, standard cars that have been refitted for electric mobility, small series EVs, and electric motorcycles. For the analysis of trips and charging intervals presented here, we concentrated on a subset of 16 Think City vehicles. These cars were selected since charging interval data was available for a full year at the time of writing. After initial analysis, we excluded one vehicle from the data set since no data from the speedometer was available (though GPS-based speed readings were) leaving us with a total of 15 cars for the analysis of trips and charging intervals. For the creation of road utilization maps we used data of the entire fleet. The vehicles were given to private and business users in the model region who gave written consent to the use of their mobility data. The identity of the user for a lease period was not revealed to the data analyst. No record was made of drivers for individual trips.

The data set includes GPS position, vehicle speed, battery parameters (current, voltage, temperature, state-of-charge), static vehicle information, and status and control flags. Weather data is also recorded from three weather stations located inside the model region via the Yahoo! Weather web service [9] and stored alongside the telemetry information. Most of the data analyses that are presented in the following are based on the GPS positions and the battery state-of-charge values of the vehicles. However, the data that is available in total provides the potential for an extended analysis of other operational parameters as well.

To achieve this, all vehicles in the fleet are equipped with a board computer designed for fleet management purposes. The system is equipped with a custom-built firmware that enables reading telemetry from the vehicles' CAN bus and transmitting the recorded raw data via a GPRS connection to a central server stack where the data is decoded and consolidated. Wireless communication is secured via a VPN tunnel to the data center. The server stack consists of three virtualized machines running on a blade center with attached storage.

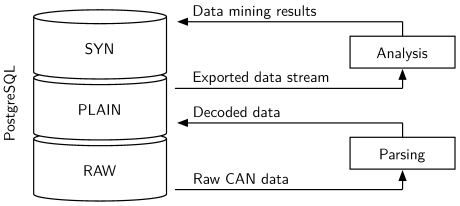

Fig. 2. The three-layered database architecture that separates data at different levels of abstraction.

Data is replicated from the data loggers' gateway software into a Postgres database where it is consolidated before raw telemetry data is decoded. Fig. 1 gives a schematic overview of the data acquisition infrastructure.

The database architecture consists of three layers named RAW, PLAIN and SYN (see Fig. 2). On the lowest layer (RAW), data that is collected from the vehicles is stored directly without any conversions made. A process, which runs simultaneously to the data acquisition itself, parses the data records stored on the RAW level and converts these vehicle-specific records into a homogeneous format. The results are stored on the second level (PLAIN). For example, while on the RAW level different vehicle models use different CAN IDs to transmit speed readings from the speedometer, these are mapped to the same ID signifying speedometer speed readings on the PLAIN level. Reporting tools and data mining applications can access this set of homogeneous data records, perform their analysis and store the results on the third level (SYN).

By using this three-layer bottom-up strategy, the raw data collection is preserved in its original state, thus allowing for later changes to the raw data decoder or the addition of further steps of preprocessing. Furthermore, the generation of homogeneous data structures on the PLAIN layer allows analysis of data collected from different vehicle models in a unified fashion independent of the specific native format in which the corresponding data was encoded originally. If the formatting on the lower levels changes, subsequent processing applications can simply be re-run to update higher level representations to the most recent state, e.g. in case of new data sources being available on the PLAIN level, or in case of new analysis tools for the SYN level.

B. Preprocessing

The entire preprocessing procedure serves two aims. As explained in the previous section, the vehicle-specific data needs to be converted into a homogeneous format. Second, the available data needs to be sampled, transformed, and cleaned. The second aspect will be addressed in this section.

1) Data Stream Generation: Data is exported from the system as a sequence of triples containing timestamp, data type identifier (e.g., for the telemetry stream: speed, latitude, longitude, battery state-of-charge, etc.) and value (t_i, σ_i, x_i) to the analysis tools. Timestamps need not be evenly spaced, it is only guaranteed that $t_i \leq t_{i+1}$ for two subsequent triples.

Algorithm 1: Route construction from GPS

```
Data: Stream of GPS positions
Result: List of edges in street graph
begin
    Find the start of a trip;
    p ←first GPS point of the trip;
    START ← all edges near p;
    root ← empty hypothesis;
    Add all edges e ∈ START as successors of root;
    TRIP ← empty list of edges;
    while no end of trip detected do
        Get the next GPS point p;
        NE ← nearby edges of point p;
        SH ← active hypotheses including an edge
        e ∈ NE;
        for all hypotheses h ∈ SH do
            Add support to h recursively back to root;
            Find adjacent edges in street graph;
            Add these edges as successors to h;
        while root has only one likely successor do
            step ← most likely successor of root;
            Remove all other successors of root;
            Add root to TRIP;
            root ← step;
    return TRIP;
end
```

The nature of the stream-based approach allows for simple merging of two separate streams and for the attachment of more than one analysis tool to the same stream. For example, one could add external data sources like weather data at a later point in time and attach an additional module that gathers weather statistics per trip. Note that while we are presently exporting data streams from a consolidated database, stream-based approaches that do not necessitate materialization of all incoming data are an interesting prospect (see [10] for a detailed overview on the potentials and the issues related to stream-based approaches to data processing.)

Further, this approach adds the possibility of allowing future online analysis of the EV fleet in operation, which could be interesting for a number of real-time vehicle routing problems (VRPs) on a fleet base (see [11] for an overview of VRPs).

2) Identification of Trips: Vehicle data is also logged while the vehicle is standing. Therefore, it was necessary to perform a trip identification procedure to be able to detect usage intervals. Additionally, later analyses require knowledge of the traveled distance during the trips, which is not available from the CAN bus in the present case. To get a noise-free representation of individual trips we did not directly use the recorded GPS positions to construct the traveled path. Instead, we constructed the trip indirectly by matching GPS positions to a OSM-based road graph using the following algorithm (making use of PostGIS and pgRouting [12], [13], [14]).

The core of this route construction module is Algorithm 1.

Fig. 3. Section of a trip's GPS coordinates (gray dots) mapped onto the road graph using our method (mapped edges are shown by dark gray line). © OpenStreetMap contributors, CC-BY-SA.

This approach allows to solve ambiguities between very close or parallel road sections that could not be resolved with the typically noisy GPS positions. It allows to produce connected routes through the road graph even where very short road sections are not directly confirmed by a GPS position.

First, the continuous stream of GPS positions is searched for the possible start of a trip indicated by a major change in the position for a number of consecutive points. Once the start of a trip has been detected, all edges in the routing graph within a predefined search radius from the starting position are selected. Each of these is considered a candidate for the real start of the trip and as such forms a hypothesis for the route that the vehicle might have taken. The algorithm then continues by recursively "predicting" the possible course for each currently active hypothesis for a defined minimum distance ahead. At each junction in the routing graph, a hypothesis is split into several ones, each one marking a new possible route. When the prediction is completed, the algorithm continues by selecting all edges near the next GPS position. These edges are then given a certain support value which is back-propagated to all hypotheses that lead to this edge. If a hypothesis has gained significantly more support than all of its competitors, it is accepted and all competing hypotheses are dropped. This is continued until the end of the trip is found, when there is no further change in the GPS position for a certain amount of time.

Fig. 3 shows an example for a trip that was constructed using this algorithm. The dots mark the recorded GPS positions and the line represents the edges of the routing graph that positions were matched to.

The module for trip detection described above generates a stream of triples that mark entry and exit of routing graph edges, trip start and end, and trip distance calculated from the OpenStreetMap road graph.

C. Data Analysis

The analysis modules that were attached to the generated streams are described below. For the analysis of trips and charging intervals the vehicle telemetry stream was merged with the stream of reconstructed trip data.

1) Road Utilization Statistics: On the basis of the trip data we generated a road utilization map that allows to determine

Fig. 4. Part of the road utilization map of the model region for the initial year of operation. The thickness of road segments indicates the frequency of road utilization. Map data © OpenStreetMap contributors, CC-BY-SA

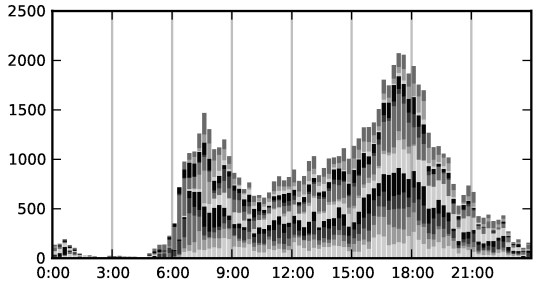

Fig. 5. Aggregated vehicle usage over the course of a day. The y-axis shows the number of trips that fell into the specific time frame. Contributions by individual vehicles are color-coded.

the frequency by which each road segment was actually used by the entire fleet (see Fig. 4). This is valuable information, e.g., for traffic planning or for determining optimal locations for charging points.

2) Trip Statistics: Another module was attached to the stream to generate statistics on the basis of the detected trips, such as battery state-of-charge, trip start and end time, and flags indicating which road segment the vehicle was traveling on. The trip length was calculated using the accumulated edge lengths covered in the road graph.

The state-of-charge is an indicator for the actual charge of the battery. It is a dimensionless number $[0, 1]$ as displayed on the cars' dashboard rather than a physical measure of battery capacity. While the latter is valuable information for energy planning, the battery state-of-charge value is important to EV users as they have to rely on this measure to assess whether there is sufficient charge left for a planned trip.

Let E be the index set of all edges of the road graph. For the purpose of our analysis, we defined a (trip) *leg* l as being a sequence of n (trip) segments $l = (s_1, \ldots, s_n)$ with $s_i = (e, t_{enter}, t_{exit})$ where $e \in E$ is the specific edge identifier, and t_{enter} and t_{exit} the timestamps when the vehicle entered and left the specific road segment.

While the length of a trip segment s_i is given by the length $d(e_i)$ of the specific edge of the road graph, we defined the *distance* of a leg as the sum of the lengths of its segments $d(l) = \Sigma_i^{n_l} d(e_i)$.

The *state-of-charge difference* of a (trip) leg $\delta_{SOC}(l)$ was defined to be the difference between the first and the last state-of-charge value recorded while the vehicle was visiting the leg's road segments. We applied a moving window median to the values of the state-of-charge to reduce the influence of noise effects. On the basis of these definitions we defined the *state-of-charge usage* of a (trip) leg l as the state-of-charge difference per covered distance of the leg

$$U_l = \frac{\delta_l(SOC)}{d(l)}.$$

A *trip* is a leg where the vehicle is standing before the first and after the last trip segment with a certain maximum standing time allowed during the trip.

Analogous to trip legs and trip segments, *route legs* and *route segments* designate the corresponding edges of the road graph without timestamps indicating when they were entered and exited.

Trip detection was performed to assess the distribution of trip durations and covered distances.

3) Identification of Charging Intervals: A module was attached to the stream that gathers statistics on the charging intervals (such as start and stop times). A charging interval was defined to start whenever the measured battery current as reported by the vehicles' energy management system was negative. Note that this definition includes recuperation happening during driving.

III. Results

A. Trip Statistics

We detected a total number of 5505 trips in the initial year of operation that covered a minimum distance of $d_{min} = 1\,\mathrm{km}$ and had no significant gaps in telemetry coverage (max 60s). Interestingly, although the vehicle model has a range of well over 100km, most trips were very short as can be seen in Fig. 6 (top). Despite this, the majority of trips was started with highly charged batteries, although battery capacity would have allowed to recharge the vehicles less often as can be seen in Fig. 6 (bottom). There seems to be a linear relationship between trip distance (as calculated from the road graph) and U_{trip} ($R^2 = 0.946$). We found that the spread of state-of-charge differences per traveled kilometer was significantly higher when distances were calculated from GPS positions alone (data not shown; $R^2 = 0.441$) as we calculated on a subset of data (n = 3855). This indicates that using GPS positions alone to measure total trip length leads to less accurate results. We therefore plan to integrate these findings into a road planner for the model region that can show the user the expected range on the road graph given his vehicles current charge.

We also analyzed vehicle usage over the course of a day (Fig. 5) and could for the first time characterize the usage profile of our fleet. The profile indicates two usage peaks consistent with rush hours in the region.

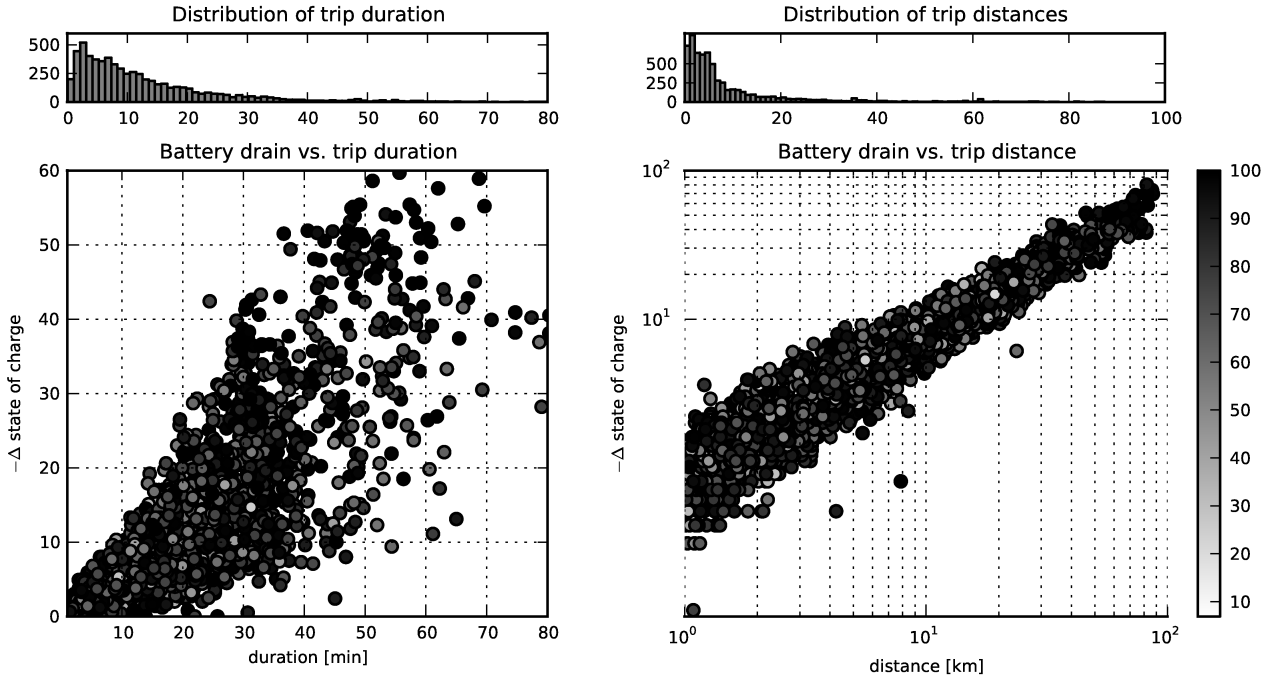

Fig. 6. Trip statistics. Each point in the scatter plot represents a trip. The battery state-of-charge at the start of the trip is color-coded. The upper charts show histograms of duration (left) and covered distance (right) of the recorded trips.

X	$r_{X,U_{trip}}$
Mean speed	0.171
Trip duration	0.080
Trip distance	0.112
Battery temperature	0.101

TABLE I
PEARSON'S CORRELATION COEFFICIENTS OF TRIP ATTRIBUTES AND STATE-OF-CHARGE USAGE U_{trip} (TWO-TAILED P-VALUE < 0.001 FOR ALL X, N = 5505).

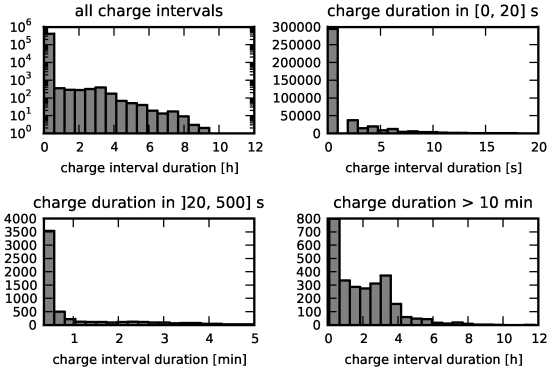

Fig. 7. Histogram of charge interval durations. Note the log-scale in the upper left panel.

As can be seen in Fig. 6, the relative state-of-charge differences are not constant per traveled kilometer. This is unfavorable since the remaining state-of-charge value is used as an indicator by the vehicle user to determine if there is sufficient charge available for a certain trip of known distance. We were thus interested if the observed variability is due to measurement noise or can be attributed to other factors that can be taken into account to make more accurate estimates of U_{trip}.

To demonstrate the flexibility of the system, we compiled a list of possible external factors that are expected to have an influence on the state-of-charge usage. We then calculated the correlation of each factor and the state-of-charge usage per trip. We found mean speed to be most strongly correlated to the state-of-charge usage per trip as can be seen in Table I.

Note that while this is a simple analysis, it can be further segmented by additional information that is contained in the stream as well (e.g., road segments, lending intervals, wind and weather conditions). Future work will describe correlations and possible interactions between attributes in order to find those that can be used to predict state-of-charge usage for a planned trip most accurately.

B. Charging Intervals

We detected over 420.000 battery charging intervals. Fig. 7 gives an overview of the distribution of interval durations. Note there seem to be three peaks in the distribution of charge interval durations. While the first peak between 0s and 2s can be explained by recuperation charging during vehicle deceleration, the charging intervals between 20s and 1min remain to be explained. We plan to use time series knowledge mining for the characterization of these intervals (see outlook).

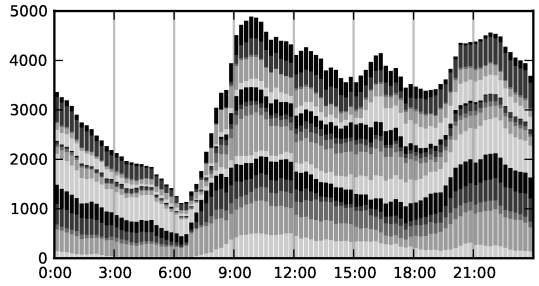

Fig. 8. Aggregated charge interval coverage over the course of a day. The y-axis shows the number of charge intervals that fell into the specific time frame. Contributions by individual vehicles are color-coded.

Fig. 8 shows charge interval coverage (minimum duration 10min) over the course of the day.

IV. Conclusion

We presented a system for large scale recording, integration, and analysis of data from heterogeneous EV fleets. Using a stream-based approach, analysis tools and data sources can be easily added to the setup to extract and aggregate relevant information from the recorded telemetry data. We presented a number of approaches to such analyses, specifically road utilization statistics, the segmentation and construction of individual trips within the data stream, and the computation of statistics about the vehicles operational parameters (for both charging and trip intervals). We showed that calculation of trip distances from the road graph rather than directly from GPS positions yields more robust distance estimates when mileage readings are not available via telemetry.

V. Future Directions

Future modules will allow analysis of discharging coefficients for individual route segments to build accurate and model-specific EV range planners. Further studies will have to elucidate whether the availability of such planners can help strengthen users confidence in taking EVs for longer trips.

We are also planning to use the available weather data to more thoroughly assess the effects of temperature, wind speed, and direction on the possible range.

We are also integrating modules for mining temporal patterns to aid in the discovery of user or user group-specific mobility patterns using Time Series Knowledge Mining (TSKM) [15].

Acknowledgments

We thank the group of Matthias Busse and Gerald Rausch from the Fraunhofer Institute for Manufacturing Technology and Advanced Materials in Bremen (IFAM) for providing a large part of the fleet.

References

[1] A. Brooks and T. Gage, "Integration of Electric Drive Vehicles with the Electric Power Grid – a New Value Stream," in *18th International Electric Vehicle Symposium and Exhibition*, 2001, pp. 20–24.

[2] W. Kempton and J. Tomic, "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.

[3] C. Quinn, D. Zimmerle, and T. Bradley, "The effect of communication architecture on the availability, reliability, and economics of plug-in hybrid electric vehicle-to-grid ancillary services," *Journal of Power Sources*, vol. 195, no. 5, pp. 1500–1509, 2010.

[4] EPRI Electric Transportation, http://et.epri.com/ETinfo.html, accessed: 2012-02-08.

[5] S. Midlam-Mohler, S. Ewing, V. Marano, Y. Guezennec, and G. Rizzoni, "PHEV Fleet Data Collection and Analysis," in *Vehicle Power and Propulsion Conference, 2009. VPPC '09. IEEE*, sept. 2009, pp. 1205–1210.

[6] Q. Gong, S. Midlam-Mohler, V. Marano, G. Rizzoni, and Y. Guezennec, "Statistical analysis of phev fleet data," in *Vehicle Power and Propulsion Conference (VPPC), 2010 IEEE*, sept. 2010, pp. 1–6.

[7] The REV Project, http://www.therevproject.com, accessed: 2012-02-08.

[8] P. Kasten and W. Zimmer, "CO_2-Minderungspotenziale durch den Einsatz von elektrischen Fahrzeugen in Dienstwagenflotten [Potentials for CO_2 mitigation by deployment of electric vehicles in company vehicle fleets]," Öko-Institut e.V., Tech. Rep., 2011, (in German).

[9] Yahoo! Weather, http://weather.yahoo.com/, accessed: 2011-10-14.

[10] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ser. PODS '02. New York, NY, USA: ACM, 2002, pp. 1–16. [Online]. Available: http://doi.acm.org/10.1145/543613.543615

[11] G. Ghiani, F. Guerriero, G. Laporte, and R. Musmanno, "Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies," *European Journal of Operational Research*, no. 1, pp. 1–11, 2003.

[12] OpenStreetMap, http://www.openstreetmap.org/, accessed: 2011-10-14.

[13] PostGIS, http://postgis.refractions.net/, accessed: 2011-10-14.

[14] pgRouting, http://www.pgrouting.org/, accessed: 2011-10-14.

[15] F. Mörchen and A. Ultsch, "Efficient mining of understandable patterns from multivariate interval time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 181–215, 2007.