Statistics Instead of Stopover—Range Predictions for Electric Vehicles

Christian Kluge, Stefan Schuster and Diana Sellner

Abstract Electric vehicles (EVs) can play a central role in today's efforts to reduce CO_2 emission and slow down the climate change. Two of the most important reasons against purchase or use of an EV are its short range and long charging times. In the project "E-WALD—Elektromobilität Bayerischer Wald", we develop mathematical models to predict the range of EVs by estimating the electrical power consumption (EPC) along possible routes. Based on the EPC forecasts the range is calculated and visualized by a range polygon on a navigation map. The models are based on data that are constantly collected by cars within a commercial car fleet. The dataset is modelled with three methods: a linear model, an additive model and a fully nonparametric model. To fit the linear model, ordinary least squares (OLS) regression as well as linear median regression are applied. The other models are fitted by modern machine learning algorithms: the additive model is fitted by boosting algorithm and the fully nonparametric model is fitted by support vector regression (SVR). The models are compared by mean absolute error (MAE). Our research findings show that data preparation is more influential than the chosen model.

1 Introduction

The use of EVs can play a central role in today's efforts to reduce CO_2 emission and slow down the climate change [10]. Despite research funding and public support, consumers react cautiously to current offers of the EV market. Surveys show that

C. Kluge (✉) · S. Schuster · D. Sellner
Technische Hochschule Deggendorf, Edlmairstraße 6 und 8,
94469 Deggendorf, Germany
e-mail: christian.kluge@th-deg.de

S. Schuster
e-mail: stefan.schuster@th-deg.de

D. Sellner
e-mail: diana.sellner@th-deg.de

C. Kluge
Technologiecampus Grafenau, Hauptstraße 3, 94481 Grafenau, Germany

A. Fink et al. (eds.), *Operations Research Proceedings 2016*,
Operations Research Proceedings, DOI 10.1007/978-3-319-55702-1_8

two of the most important reasons against the purchase or use of an EV are its short range and long charging times [13].

While the problem of long charging times is of technical nature, the problem of short range has also a psychological dimension known as range stress, the fear of running out of energy on an open road. Especially for new users in electric mobility this mental pressure is intensified by a highly unreliable range prediction offered by car itself. The built-in range prognosis of cars is often based on the EPC of the immediate past. Therefore, in mountainous regions, where elevation changes are frequent and high, the range prognosis varies drastically with the elevation profile of the passed route. To better support drivers, the project "E-WALD—Elektromobilität Bayerischer Wald" equips EVs with tablet computers that visualize the remaining range by a polygon drawn on navigation map.

One way to estimate the range of an EV is to predict the EPC along routes that may be travelled. In this study, we describe the development and comparison of different models to choose the best model for estimating the EPC. The considered models are a simple multivariate linear regression fitted by OLS, a linear median regression also known as least absolute deviation (LAD) regression fitted by quantile regression, an additive model fitted by a boosting algorithm and a fully nonparametric model fitted by a SVR. Our approach is driven by the goal to estimate EPC in a way that is as independent from car model specific properties as possible. This will allow to apply the modelling process to a wide variety of vehicles from different car manufacturers.

The structure of this paper is as follows: In Sect. 2 we describe how the data was obtained and prepared. Section 3 presents the process of model development. The model evaluation is given in Sect. 4, and Sect. 5 concludes this work with a short discussion.

2 Data Description and Preparation

Data were collected from Nissan LEAF vehicles that are part of a commercial car fleet operated by the E-WALD GmbH. To store the data, tablet computers which constantly record the car trips have been installed in these EVs.

The data, such as battery power, ambient temperature, speed, heater consumption, as well as GPS coordinates (latitude and longitude), were collected with an interval of 1 s during the trips from September 2014 to January 2015 for 7 Nissan LEAF vehicles. To improve the quality of the data base, erroneous data and outliers have been removed. The features of the data are as follows: length of trips is between 3 and 75 km, duration of trips is between 5 min and 1 h, temperature is between −4 and 25 °C. After filtering, about 385 trips can be used for further analysis.

Our approach is to estimate the EPC independent from specific car models. We therefore concentrate on external factors such as elevation difference and temperature, and investigate their influence on the EPC. To distinguish the influence of ascending versus descending slope on the EPC, we introduce the notion of positive elevation difference (PED) which is defined by the sum of meters a car travelled through ascending slope and negative elevation difference (NED) which is correspondingly defined by descending slope. In this study, a trip is divided into parts of by exactly 3 km travelled distance. In order to estimate EPC in GID (a Nissan LEAF internal unit which amounts to 80 Wh) per 1 km and slope, the entries on EPC, PED and NED have to be divided by the respective distance travelled (distance-based dataset).

3 Model Development

In literature, there are a lot of different methods for fitting linear models. The most prominent method is OLS regression. Besides, least absolute deviation (LAD) regression is also often used. While OLS is based on estimating the mean of a distribution, LAD is based on estimating the median. The additive model is fitted by a boosting algorithm. The first boosting algorithm in machine learning was designed for binary classification [3, 4]. According to Friedman [5], boosting can be interpreted as a gradient descent algorithm in a function space. Bühlmann and Yu [2] introduced component-wise functional gradient descent boosting for additive models. An overview is given by [1]. The variant of boosting algorithm that was used is based on estimating the median. The fully nonparametric model is fitted by SVR. SVR is a generalization of support vector machine (SVM), which was originally designed for binary classification [11, 12, 14]. These methods belong to the wide class of methods which are based on penalized risk minimization and, therefore, are most suitable for fitting nonparametric models as they balance the trade-off between complexity and goodness of fit, c.f. [7, Chap. 5].

Model Assumptions. At first, the dataset of the recorded tracks is used for a descriptive analysis to reveal interdependencies and relevant variables that are useful predictors for the EPC. Possible variables are shown in Table 1. Therefore we selected PED and NED as important variables and assumed a linear influence on the EPC. So the following basic functional structure was chosen:

$$\frac{\text{EPC}}{\text{km}} = \beta_0 + \beta_1 \cdot \text{PED} + \beta_2 \cdot \text{NED} + \beta_3 \cdot \text{Temp}^2 + \beta_4 \cdot \text{Temp} \tag{1}$$

where $\beta_0, \ldots, \beta_4$ denote the parameters to be estimated.

Table 1 Correlation analysis on continuous data of Nissan LEAF, most relevant data are bold

Variable	PED/km	NED/km	Temperature	Mean velocity
r (EPC/km)	**0.4084**	**−0.4413**	−0.0446	0.0470

The Models. The dependent variable is EPC and independent variables are PED, NED, and temperature. Three models with different degrees of generality have been investigated. The simplest model is the linear model

$$y = \beta_0 + \beta_1 \cdot x_{pos} + \beta_2 \cdot x_{neg} + \beta_3 \cdot x_{temp}^2 + \beta_4 \cdot x_{temp} + \varepsilon \tag{2}$$

where y denotes the EPC, x_{pos} the PED, x_{neg} the NED, x_{temp} the temperature, ε the error term and β_i the parameter vector. A convenient generalization of a linear model is the additive model [6].

$$y = \beta_0 + f_{pos}(x_{pos}) + f_{neg}(x_{neg}) + f_{temp}(x_{temp}) + \varepsilon \ . \tag{3}$$

The difference to the linear model is that the additive model also captures nonlinear effects (f_{pos}, f_{neg} and f_{temp} are continuous functions). The study was done using the statistical software R where we applied the function gamboost with smooth P-spline base-learners PED, NED, and temperature [1, 8, 9]. Finally, we also considered the fully nonparametric model

$$y = f(x_{pos}, x_{neg}, x_{temp}) + \varepsilon \ . \tag{4}$$

As the additive model, the fully nonparametric model captures nonlinear effects. In contrast to the additive model, it also captures all kinds of interactions between independent variables so that the fully nonparametric model, in fact, is more general than the additive model. This was done using the R package e1071.

4 Results

As a measure for quality, the MAE has been chosen. Where n denotes the number of data points, y_i denotes the EPC (in GID) of data point number i and $\hat{y}_i$ contains corresponding estimate from the model, the MAE is given by

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \ . \tag{5}$$

In case of more advanced nonlinear methods like Boosting and SVR, simply calculating MAE on the whole dataset is not appropriate; In order to avoid the problem of overfitting and to obtain honest values, the MAE was calculated using 10-fold cross-validation [7, Chap. 7]. Table 2 shows the results of the different models. All estimators which are calculated nearly have the same quality. The MAE of the LAD regression has the lowest value. Results were also compared with the global mean. It is simply the mean of the whole dataset. In doing so, the estimate $\hat{y}_i$ is always equal to the mean so that $\hat{y}_1 = \hat{y}_2 = \cdots = \hat{y}_n = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$. The global mean acts

Table 2 Results of MAE for each model

Model	MAE	Improvement to global mean (%)	Improvement to OLS (%)
Global mean	1.098	0	−47.12
OLS	0.746	32.03	0
LAD	0.742	32.45	0.62
Boosting	0.744	32.25	0.33
SVR	0.743	32.37	0.51

as a benchmark because this is the result which could be obtained without collecting any data in the car. The 3rd and 4th column show the percentaged improvement to global mean and OLS respectively. Because all applied models have nearly the same performance, it is entirely sufficient to take the much simpler linear methods (OLS and LAD regression) for predicting the EPC.

5 Discussion

The perhaps most interesting aspect of the results is that the performance of models hardly makes a difference which estimator is chosen. During analysis it was also investigated how another data preparation will change the results. According to one possible way to prepare the data is to divide the trips into parts of 1 GID (of consumed energy) and to extrapolate the travelled distance to 1 km distance (energy-based data). So energy-based dataset and distance-based dataset (Sect. 2) in this study can be compared. As you see in Table 3 the estimated regression coefficients, the influence of independent variables are larger for the distance-based approach than for the energy-based approach. The MAE of the OLS with energy-based dataset was 0.886, very much higher than the MAE of OLS of the distance-based dataset (0.746, see Table 2). So the quality of estimators heavily depends on the way how the dataset is prepared but not which model is chosen. This is remarkable that the vast majority of research in data analysis is concerned with the choice of model and not with the topic of data preparation. In our case, the distance-based dataset is much smaller than the energy-based dataset (n = 1476 vs. n = 4656) but yields much better results. This

Table 3 Estimated regression coefficients (rounded)

Model	$\hat{\beta}_0$	$\hat{\beta}_{pos}$	$\hat{\beta}_{neg}$	$\hat{\beta}_{temp^2}$	$\hat{\beta}_{temp}$
OLS (energy-based data)	2.61	0.028	0.028	0.00014	−0.026
OLS (distance-based data)	2.64	0.067	0.041	0.00084	−0.061
LAD (distance-based data)	2.55	0.068	0.042	0.00064	−0.055

demonstrates, it is more important to have the right dataset, not the biggest dataset. In order to further improve quality of forecasts, it is interesting to investigate the history of forecasts separately for each trip. The current estimates are static. Therefore, it seems to be promising to improve estimations by adding dynamic and adaptive components.

Acknowledgements The E-WALD project and this study have been funded by the Bavarian State Ministry for Economic Affairs and Media, Energy and Technology.

References

1. Bühlmann, P., Hothorn, T.: Boosting algorithms: regularization, prediction and model fitting (with discussion). Stat. Sci. **22**, 477–505 (2007). doi:10.1214/07-STS242
2. Bühlmann, P., Yu, B.: Boosting with the l_2-loss: regression and classification. J. Am. Stat. Assoc. **98**(462), 324–338 (2003). doi:10.1198/016214503000125
3. Freund, Y., Schapire, R.E. (eds.): Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning Theory, Morgan Kaufmann Publishers Inc., San Francisco (1996)
4. Freund, Y., Schapire, R.E.: A decision—theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). doi:10.1006/jcss.1997.1504
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001). doi:10.1214/aos/1013203451
6. Friedman, J.H., Stuetzle, W.: Projection pursuit regression. J. Am. Stat. Assoc. **76**(376), 817–823 (1981). doi:10.1080/01621459.1981.10477729
7. Hastie, T.J., Tibshirani, R.J., Friedman, J.H.: The elements of statistical learning: Data mining, inference, and prediction, 2nd edn., corr. at 7. printing edn. Springer Series in Statistics. Springer, New York (2001)
8. Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B.: Model-based boosting 2.0. J. Mach. Learn. Res. **11**, 2109–2113 (2010)
9. Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B.: mboost: Model-based boosting (2016). http://CRAN.R-project.org/package=mboost. R package version R package version 2.6-0
10. Li, C., Cao, Y., Zhang, M., Wang, J., Liu, J., Shi, H., Geng, Y.: Hidden benefits of electric vehicles for addressing climate change. Sci. Rep. **5**(9213) (2015). doi:10.1038/srep09213
11. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press, Massachusetts (2002)
12. Steinwart, I., Christmann, A., Jordan, M., Kleinberg, J., Schölkopf, B. (eds.): Support vector machines. Information Science and Statistics. Springer, New York (2008). doi:10.1007/978-0-387-77242-4. http://www.springerlink.com/content/uk1165
13. Türnau, M.: Befragung von elektrofahrzeug–mieterinnen. Mobilität, Gesellschaft und Technik (2014). http://digital.bib-bvb.de/webclient/DeliveryManager?pid=7179727Źcustom_att_2=simple_viewer
14. Vapnik, V.N.: Statistical Learning Theory. A Wiley-Interscience Publication. Wiley, New York (1998). http://www.loc.gov/catdir/description/wiley032/97037075.html