Prediction of Availability and Charging Rate at Charging Stations for Electric Vehicles

Can Bikcora[1], Nazir Refa[2], Lennart Verheijen[3], and Siep Weiland[1]

[1]Department of Electrical Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

[2]ElaadNL, 6812 AR Arnhem, The Netherlands

[3]GreenFlux Assets B.V., 1092 AD Amsterdam, The Netherlands

c.bikcora@tue.nl; nazir.refa@elaad.nl; lennart.verheijen@greenflux.nl; s.weiland@tue.nl

***Abstract*—To enable better smart charging solutions, this paper investigates the day-ahead probabilistic forecasting of the availability and the charging rate at charging stations for plug-in electric vehicles. Generalized linear models with logistic link functions are at the core of both forecast scenarios. Moreover, the availability forecast at a charging point is simply a binomial problem, whereas the charging rate forecast is handled via an ordered logistic model after categorizing the feasible range of values. These two scenarios are evaluated on real data collected from two representatives of the most occupied charging points in the Netherlands, with the focus of the analysis kept at the selection of essential regressors. Based on the ranked probability scores associated with the day-ahead forecasts generated for the last nine months of 2015, it is concluded that the usefulness of predictive models depends highly on the charging station. When contributing substantially to performance, such models possess a simple structure with a few basic lagged and indicator variables.**

***Index Terms*—Demand forecasting, electric vehicles, load modeling, predictive models, time series analysis.**

I. INTRODUCTION

With the penetration of plug-in electric vehicles (PEVs) becoming more pronounced from one year to the next [1], the power and energy sector nowadays is undergoing a transitional phase to address the imposed challenges. As the foremost inevitable outcome associated with higher penet-ration levels, the anticipated boost in the electricity demand is likely to disrupt the current infrastructure and operation due to amplified peaks in daily demands and voltage variations [2]. To alleviate this problem, smart charging (SC) [3] distributes the deferrable portion of the PEV load to time moments at which the base (non-flexible) demand is low, so that a relatively flat total load profile is ensured.

The performance of a SC method ideally relies on the accuracy of the prior knowledge on the load realization and demand to occur in the relevant future timeframe (e.g., in the coming hours or day). For this reason, load forecasts constitute the most crucial inputs to SC algorithms that essentially operate as model predictive controllers in real-time [4–6]. Despite the abundance of real-data investigations concerning short-term forecasting of the base load [7], to the authors' knowledge, there has not been much analysis on forecasting the PEV demands, apart from the work [8] where a synthetic time-series data was used. This is arguably due to the low occupation rates and relatively small number of transactions taking place at charging points.

To contribute to the outlined deficiency in the literature, this paper aims to forecast the availability and the charging rate at charging points, based on measured data from several charging stations in the Netherlands. Having the world's third best-selling market for PEVs [9], with number of PEVs being doubled from 2014 to 2015 [10], the Netherlands manifests itself as a prominent test case in terms of its availability of non-synthetic data. Although charging rate forecasts have more relevance to SC, the availability alone is also an interesting forecast to consider, such as for vehicle routing applications. Since predictions on short-term horizons are relevant in the described contexts, we conduct day-ahead forecasts as a valid example to this category. Moreover, rather than producing *deterministic* (point) forecasts where a single value is returned per specified time moment, we generate *probabilistic* forecasts [11] of the desired quantities. In this way, the full probability distribution of a forecast is computed, which is more informative for control purposes such as SC.

Regarding the theory used in forecasting, both the availability and the charging rate are forecasted by means of generalized linear models (GLMs) [12] [13] with the *logit* (logistic) function employed as the link function. As the main difference between these two cases, the former is a binomial logistic regression problem, whereas the latter is treated in terms of a multinomial (polytomous) model. In more detail, the allowed range of charging rates are first discretized into several categories, and afterwards, an *ordered* (ordinal) *logit model* (also referred to as the *proportional odds model*) [14] is used to compute the forecasted probabilities associated with each category. Such discretization of the charging rate values is a reasonable option to consider due to the fact that the encountered charging rates are mostly in the vicinity of a few particular values. To compare among forecasts, the ranked probability score [15] [16] is chosen, which is a widely adopted proper scoring rule for probabilistic forecasts of categorical variables.

In what follows next, we first elaborate on the underlying models and specifications in forecasting, given in Section II. Then, the performance metrics and comparisons are presented in Section III, followed by the closing section where the conclusions are emphasized.

II. Description of the Utilized Models

A. Forecasting of the Availability

If a PEV occupies a charging point (socket), then the corresponding time-series is assigned the value of 1, which is otherwise 0. With this information, let y_t denote the value of this variable at the discrete-time index t where the instants within a day correspond to the times 00:00, 00:15, …, 23:30, 23:45, amounting to 96 moments per day.

Due to the binary nature of y_t, the Bernoulli distribution is well suited to characterize it. To describe its conditional mean μ_t with respect to the available information set Υ_{t-}, an auto-regressive model structure with exogenous inputs (ARX) is used within the framework of GLMs [12] [13]:

$$\mu_t := E\{y_t \mid \Upsilon_{t-}\} = g^{-1}\left(c + \sum_{i\in\Omega}\phi_i y_{t-i} + \sum_{j\in\Phi}\beta_j x_{j,\,t-m_j}\right) \qquad (1)$$

where $g(.)$ is a monotone mapping that is referred to as the link function, Υ_{t-} consists of $\{y_{t-i}\}_{i\in\Omega}$ and $\{x_{j,\,t-m_j}\}_{j\in\Phi}$, the sets Ω and Φ include the positive integers that correspond to the indices of the modeled lagged measurements and exogenous inputs, respectively, the scalars c, $\{\phi_i\}_{i\in\Omega}$, and $\{\beta_j\}_{j\in\Phi}$ are model parameters, and $\{m_j\}_{j\in\Phi}$ are the corresponding lags of the exogenous variable x with the index j. Note that for $g(z) = z$, (1) turns into a classical linear regression problem. In our case, the argument of $g(.)$ is $\mu_t \in [0,1]$ (due to the binary variable y_t) and the range of this function can take any value on the real line. Among numerous monotone functions $g(.)$ having this domain and range, the *logit* function given by $g(z) = \ln(z/(1-z))$ is arguably the most common choice where some justifications for its popularity are conveyed in [17]. By definition, $\mu_t = E\{y_t \mid \Upsilon_{t-}\} = p(y_t = 1 \mid \Upsilon_{t-})$, which is the probability (conditioned on the past data) that there exists a PEV at the charging point. Thus, the evaluation of (1) for the logit function returns

$$\mu_t = p(y_t = 1 \mid \Upsilon_{t-}) = \frac{1}{1 + \exp\left(-c - \sum_{i\in\Omega}\phi_i y_{t-i} - \sum_{j\in\Phi}\beta_j x_{j,\,t-m_j}\right)}. \qquad (2)$$

By rewriting (2) for future time instants, and using the parameter estimates from the associated maximum likelihood estimation of this model, the day-ahead probability forecasts $\mu_{t+1}, \mu_{t+2}, ..., \mu_{t+96}$ can be computed. In particular, we avoid the use of lags less than 96 in both y and x, so that iterated multi-step-ahead forecasts are not needed and only known quantities are used for the computation of each step in future. The reason for this decision is the performance degradation due to the strong influence of previously computed forecasts on the following ones under the presence of such lags. For example, the probability that a PEV departs is very small for any step-ahead forecast when $y_{t-1} = 1$ is included in the model. This observation was first made in [8], which also held for our dataset.

B. Forecasting of the Charging Rate

Charging rates at stations take values within [0, l_u] where l_u =11kW is typically encountered. Since these rates are often concentrated around specific values, depending on distinct types of PEVs, categorization of this bounded interval is a plausible route in forecasting charging rates. In accordance with this idea, we partition [0, l_u] into nc categories defined by the intervals: $[d_0 = 0, d_1]$, $[d_1, d_2]$, …, $[d_{nc-1}, d_{nc} = l_u]$ where $d_0 < d_1 < ... < d_{nc}$. Therefore, as opposed to the binary availability case, a multinomial regression model is hereby treated. We again use the logit link function, but rather than working on the direct extension of (1), we benefit from the inherent ordering in our category choice to reduce the number of problem parameters. This exploitation is achieved by means of the ordered logit model for ordinal response variables [14]:

$$\ln\left(\frac{p(y_t \le d_r \mid \Upsilon_{t-})}{p(y_t > d_r \mid \Upsilon_{t-})}\right) = \ln\left(\frac{\sum_{p=1}^{r}\mu_{p,t}}{\sum_{p=r+1}^{nc}\mu_{p,t}}\right) = c_r + \sum_{i\in\Omega}\phi_i y_{t-i} + \sum_{j\in\Phi}\beta_j x_{j,\,t-m_j} \qquad (3)$$

for $r = 1, 2, ..., nc-1$ and where $\mu_{p,t}$ is the conditional probability that y_t falls into the p-th category. As can be noticed from (3), the coefficients of y and x stay the same in each equation, whereas only the constant term c is permitted to change among the equations described by (3). Thus, this structure assumes the same impact of an explanatory variable on each (cumulative) logit irrespective of the cut-off d_r.

The category probabilities can be solved from the set of equations in (3). It can readily be observed that $\mu_{1,t}$ takes the same form as (2), whereas $\mu_{r,t}$ for $r = 2, ..., nc-1$ can be extracted after subtracting the (r-1)-th equation from the r-th equation and leaving $\mu_{r,t}$ on one side. The forecasts are produced in a similar manner to the previously explained binomial version.

C. Model Selection

Prior to the estimation of parameters with the defined model structures, the regressors to include and the length of the training (modeling) set should be determined, which constitutes the model selection phase. In our day-ahead forecasting, we choose from the set of most recent {1, 2, …, nw} weeks of data to fit the models to. Regarding the auto-regressors, the lags 96 (1-day) and 672 (1-week) are considered as possible options. Concerning the exogenous inputs, the following variables are taken into account:

1) Intraday indicator variables for each hour of a day;

2) Intraweek indicator variables for the days of a week;

3) One option from

a) Day-lagged or week-lagged categorical variable for the other charging point at the station;

b) Day-lagged or week-lagged categorical variables for all charging points of the same provider that are within 0.5 km distance from the modeled one;

c) Same as b), but now for charging points within 1 km distance from the modeled one.

Each charging station consists of two charging points (sockets) and the option a) above considers the data from the other socket, whereas b) and c) extends this setting to also involve nearby stations. Similar to the autoregressor terms, day- and week-lagged variables are taken as candidates. For both the availability and charging rate forecast cases, the corresponding categorical (discrete-valued) time-series are used in modeling.

From possible combinations of the aforementioned regressors and training set lengths, only one must be selected to yield the ultimate model. Our approach for this model selection involves generating day-ahead forecasts on the training set, which resembles cross-validation [18]. Therefore, for each model option, we forecast for every day in the training set, one by one, via a model constructed from the weeks before that day. Afterwards, an error criterion is computed for each option and the model with the lowest error is selected. Since we adopt the ranked probability score (see subsection III.A) as our error measure to evaluate real-time forecasts, it is also used as the criterion in model selection. Rather than following some traditional metric, such as the Akaike information criteria [19] which is more suited to one-step-ahead forecasts, our procedure is more advantageous since it considers 1-through-96–step-ahead errors of interest. To summarize, our scheme basically mimics the real forecast situation on the training set, which is essentially desired in all model selection approaches.

III. Comparison of Forecast Performances

A. Performance Metrics

Strictly proper scoring rules [20] encourage an honest comparison when evaluating probabilistic forecasts. As a short explanation, a probabilistic scoring rule $S(p,y)$ quantifies the loss associated with a forecasted density p when the observation y is realized and is called strictly proper if $Ey_d\{ S(y_d, y) \} < Ey_d\{ S(p, y) \}$ for all $p \neq y_d$ where $E\{.\}$ takes the expectation with respect to y_d, the underlying density of y. Note that in our case, a forecasted density is represented via the discrete set of probabilities $\bar{\mu} := \{\mu_1, \mu_2, ... \mu_{nc}\}$ where the time index is omitted for brevity and all μ add up to 1. As a commonly used strictly proper rule for ordered categorical forecasts, we consider the ranked probability score (RPS) [15] [16]:

$$RPS(\bar{\mu}, j) = 1 - \left(\frac{3}{2} - \frac{1}{2(nc-1)} \sum_{i=1}^{nc-1} \left[\left(\sum_{k=1}^{i} \mu_k \right)^2 + \left(\sum_{k=i+1}^{nc} \mu_k \right)^2 \right] - \frac{1}{nc-1} \sum_{i=1}^{nc} |i-j| \mu_i \right) \quad (4)$$

which takes the above value when category j is realized. The RPS takes its values on the unit interval and, by default, is defined in positive orientation, indicating that better forecasts are closer to 1. Since we prefer to work in negative orientation, we use the version given by (4) in which the default definition [16] is subtracted from 1.

For binary forecasts $\bar{\mu} := \{(1-\mu), \mu\}$ where $nc = 2$, such as the availability case, (4) simplifies into the celebrated Brier Score [21]:

$$RPS(\bar{\mu}, j) = \begin{cases} \mu^2 & \text{if } j = 0 \\ (1-\mu)^2 & \text{if } j = 1 \end{cases} . \quad (5)$$

Note that (4) and (5) have been described for only one time instant. Since performance evaluations are made at many points in time, the error values we provide in the next section are simply the averaged version of the RPS over the applicable time moments.

B. Description of the Data and Forecast Settings

The evaluations are demonstrated by running the day-ahead forecast scenarios on two representative charging stations that are relatively highly occupied with respect to the available set of stations in the Netherlands. In the sequel, we shortly refer to these stations as S1 and S2 and their time-series as {S1s1, S1s2} and {S2s1, S2s2} since there are two sockets (charging points) per station. The occupation rates of S1 and S2 in 2015 were around 40% and 25%, respectively, corresponding to around 1700 transactions for S1 and 1500 transaction for S2 with the two sockets combined. These two specific choices are especially made to illustrate that forecast accuracies can vary considerably among different stations.

The day-ahead forecasts are generated for the days starting from 1st of April 2015 until the end of the year, proceeding day by day, exactly as if this scheme operates in practice. Thus, the days in these 9 months are considered in the computation of the RPS. Whenever the model selection is active, it is run only once in every week since it is reasonable to expect that a considerable portion of the training set must be modified to have an impact on the existing model specifi-cations. Apart from the demonstrations where the model selection mechanism operates by default, we also run the forecasts for various fixed model settings in order to observe whether a fixed structure can compete with a varying one.

To depict the typical encountered charging rates, the April data for S1 and S2 are plotted in Fig. 1 and Fig. 2, respectively, with the complete dataset provided by ElaadNL [22]. As can be observed from these figures, most of the rates are nearby 3.3 kW, corresponding to PEVs using a single-phase Level II charger [23]. Apart from this fact, full-electric vehicles (FEVs) can charge quite fast around 11 kW, whereas plug-in hybrid electric vehicles (PHEVs) charge at much lower rates around 2.7 kW.

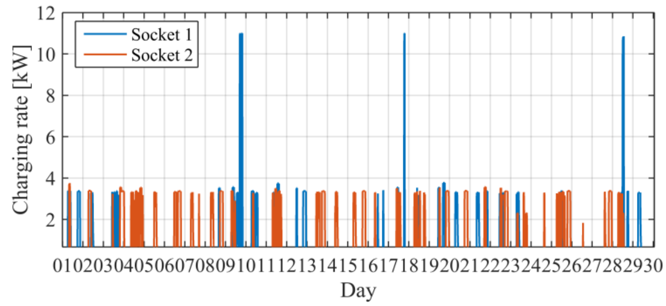

Figure 1. Charging rates at S1 during April 2015.

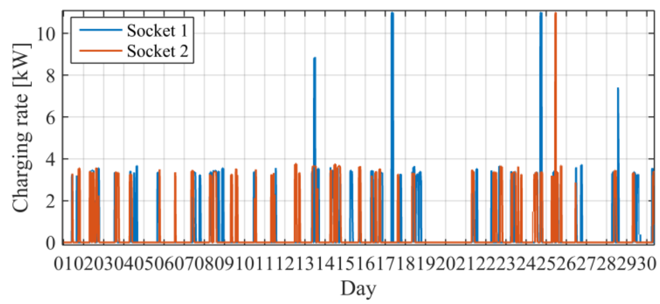

Figure 2. Charging rates at S2 during April 2015.

C. *Availability Forecasts*

For availability forecasts, the data is converted into binary form and the *fitglm* function from the Statistics and Machine Learning Toolbox of MATLAB [24] is used to formulate and estimate the associated logistic GLM.

As the first observation, which is also valid for the charging rate case, the RPS values obtained with model selec-tion can also be achieved by using a fixed model structure and training set length. Thus, although the ideal model specification varies from one charging point to the other, it stays more or less the same in terms of the day that is forecasted within a year. In accordance with this fact, model selection can be carried out much less frequently, only to find out the best settings for a given time-series.

In the second place, no evident proof has been found that the data from the other socket, as well as the data from nearby stations of the same provider, improves the forecasts. Therefore, these variables are not needed as exogenous inputs. Figs. 3–6 illustrate the RPS values when all forecasts are generated with particular fixed settings. In detail, “c” refers to the model where the ARX component in (1) has only the constant term, with no other inputs included. This means that the occupation rate of the training set is returned as the forecasted probability for all moments in the coming day, which is indeed a basic intuitive approach of forecasting. Moreover, “c+96+672+H” refers to the model where day- and week-lagged data and the indicator (dummy) variables for the hours of a day are included in the model. Other options are defined similarly and the horizontal axes in the figures show the taken fixed length of the training set in terms of the number of weeks. In consideration of Figs. 3–6, 4 weeks of length is a suitable size for all cases and the best model structure is dependent on the specific data. For instance, only hourly indicator variables are sufficient for S2.

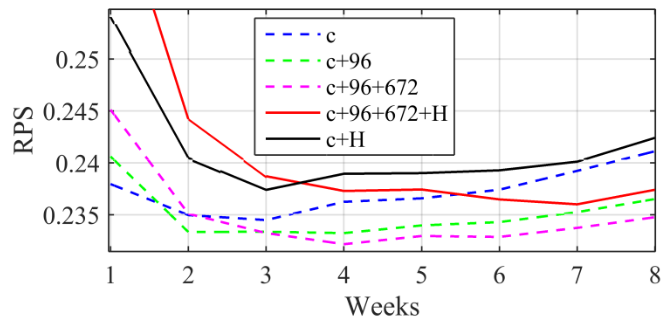

Figure 3. RPS values for various fixed model settings and training set sizes, provided for S1s1.

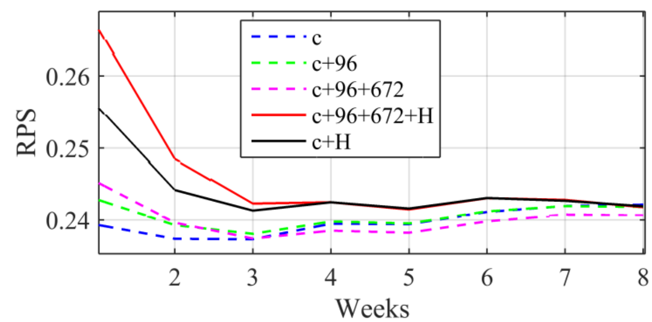

Figure 4. RPS values for various fixed model settings and training set sizes, provided for S1s2.

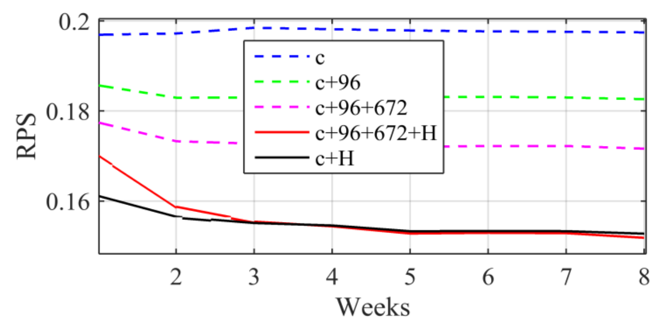

Figure 5. RPS values for various fixed model settings and training set sizes, provided for S2s1.

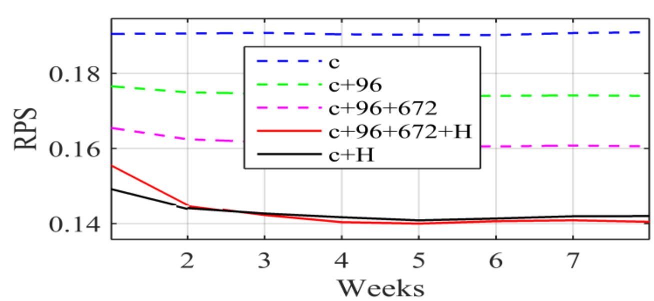

Figure 6. RPS values for various fixed model settings and training set sizes, provided for S2s2.

Furthermore, note that a dummy forecast that assigns the probability of 0.5 to all time moments has the RPS of 0.25, as one can readily observe from (5). Under this information, it is deduced from the plots for S1 that this station operates quite randomly, in the sense that only minor performance gain can be achieved via a model. Unlike S1, S2 has a convincing improvement by the use of a model. Therefore, S1 and S2 are especially chosen and demonstrated in this work as examples where a model may become either ineffective or dominant, respectively.

D. Charging Rate Forecasts

For charging rate forecasts, we define $nc = 10$ categories within the range [0, 11] kW where the edges of the intervals are set to {0, 0.5, 1.5, 2.5, 3.2, 3.8, 5, 6.5, 8, 9.5, 11}, implying that the first category is [0, 0.5], followed by [0.5, 1.5], …, and ending with [9.5, 11]. In a similar manner, any other discretization can be designed as well, in order to capture a desired detail level. In this case, the *mnrfit* function from MATLAB [24] is employed to help with the computations.

Table I lists the achieved minimum RPS values, with the availability case had already been treated. Regarding the RPS values for this case, it is immediate that they all are considerably smaller than the ones for the availability forecasts. This occurs due to the number of categories being increased from 2 to 10 and should not be interpreted as an improvement in predictability for the charging rates. Moreover, Table II presents the RPS values (rounded to three digits after the decimal point) for the training set length fixed to 4 weeks, which still holds to be suitable. Here, "c" can be interpreted as the (intuitive) forecaster that returns the frequency of each category in the training set as the forecasted probability for the time instants of the next day. The same conclusions of the previous section, such as the superior predictability of S2, are again valid.

TABLE I. LOWEST RANKED PROBABILITY SCORES.

	S1s1	S1s2	S2s1	S2s2
Availability	0.232	0.237	0.151	0.139
Charg. rate	0.0574	0.0564	0.0462	0.0453

TABLE II. RANKED PROBABILITY SCORES FOR VARIOUS SETTINGS.

	S1s1	S1s2	S2s1	S2s2
c	0.060	0.060	0.051	0.051
c+96	0.060	0.059	0.050	0.050
c+96+672	0.059	0.059	0.049	0.049
c+96+672+H	0.058	0.057	0.046	0.045
C+H	0.057	0.056	0.047	0.046

IV. CONCLUSIONS

This work investigated several ways to forecast with the available charging rate data from the current PEV charging stations in the Netherlands. In particular, the day-ahead forecasting of the availability and the charging rate are treated in a probabilistic framework where logistic GLMs have been utilized. Irrespective of the forecast scenario, model-based forecasts are proved to be useful at some charging points, whereas the opposite statement is found to be true at some others. In addition, a fixed model structure is determined as being satisfactory if it is correctly determined by model selection.

As the primary observation, when forecasting at the extreme disaggregated level of charging points (sockets), quite simple model structures, possibly consisting of hourly indicator variables, day-lagged and week-lagged measured variables, were determined to be sufficient. Since this occurs because of the high impact of randomness, it is an interesting extension, as future work, to analyze these approaches also at aggregated levels.

REFERENCES

[1] T. Trigg, P. Telleen, et al., "Global EV Outlook: Understanding the electric vehicle landscape to 2020," International Energy Agency, Tech. Rep., Apr. 2013, accessed on 25-04-2016 from http://www.iea.org/publications/globalevoutlook_2013.pdf.

[2] S. McGrath and L. Flader, "Smart Charging: steering the charge, driving the change," *A EURELECTRIC paper*, Mar. 2015, accessed on 25-04-2016 from http://www.eurelectric.org/media/169888/20032015_paper_on_smart_charging_of_electric_vehicles_finalpsf-2015-2301-0001-01-e.pdf.

[3] S. Rajakaruna, F. Shahnia, and A. Ghosh, "*Plug in electric vehicles in smart grids: charging strategies*," Springer, 2015.

[4] L. Gan, U. Topcu, and S. H. Low, "Optimal decentralized protocol for electric vehicle charging," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 940–951, 2013.

[5] N. Chen, L. Gan, S. H. Low, and A. Wierman, "Distributional analysis for model predictive deferrable load control," *53rd IEEE Conference on Decision and Control*, pp. 6433–6438, Dec. 2014.

[6] Z. Li, Q. Guo, H. Sun, S. Xin, and J. Wang, "A new real-time smart-charging method considering expected electric vehicle fleet connections," *IEEE Transactions on Power Systems*, vol. 29, no. 6, pp. 3114–3115, 2014.

[7] T. Hong and S. Fan, "Probabilistic electric load forecasting: a tutorial review," *International Journal of Forecasting*, in press, 2016.

[8] R. J. Bessa and M. A. Matos, "Global against divided optimization for the participation of an EV aggregator in the day-ahead electricity market – Part I: Theory," *Electric Power Systems Research*, vol. 95, pp. 309–318, 2013.

[9] J. Cobb, "Top Six Plug-in Vehicle Adopting Countries – 2015," accessed on 25-04-2016 from http://www.hybridcars.com/top-six-plug-in-vehicle-adopting-countries-2015/.

[10] Rijksdienst voor Ondernemend Nederland (RVO) (January 2016). "Special: Analyse over 2015" [Special: Analysis over 2015], accessed on 25-04-2016 from http://www.rvo.nl/sites/default/files/2016/01/Special%20Analyse%20over%202015.pdf

[11] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 2014.

[12] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Boca Raton: Chapman & Hall/CRC, 1989.

[13] A. J. Dobson, *An Introduction to Generalized Linear Models*, 2nd ed. New York: Chapman and Hall/CRC, 2002, ch. 8.

[14] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[15] A. H. Murphy, "The ranked probability score and the probability score: a comparison," *Monthly Weather Review*, vol. 98, no. 12, pp. 917–924, 1970.

[16] E. S. Epstein, "A scoring system for probability forecasts of ranked categories," *Journal of Applied Meteorology*, vol. 8, no. 6, pp. 985-987, Dec. 1969.

[17] P. Allison, "What's so special about logit?," 2015, accessed on 10-04-2016 from http://statisticalhorizons.com/whats-so-special-about-logit.

[18] R. Picard and D. Cook, “Cross-validation of regression models,” *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 575–583, 1984.

[19] S. Konishi and G. Kitagawa, Information Criteria and Statistical Modeling. New York: Springer, 2008.

[20] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[21] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[22] ElaadNL, “About us,” accessed on 28-04-2016 from https://www.elaad.nl/ organisatie/over-ons/about-us/.

[23] A. Ipakchi and F. Albuyeh, “Grid of the future,” *IEEE Power Energy Mag.*, vol. 7, no. 2, pp. 52–62, 2009.

[24] *Statistics and Machine Learning Toolbox*™ *User's Guide*, The MathWorks, Inc., 2015.