Groß *et al. Energy Informatics* 2021, **4**(Suppl 3):13
https://doi.org/10.1186/s42162-021-00172-6

Energy Informatics

RESEARCH Open Access

Comparison of short-term electrical load forecasting methods for different building types

Arne Groß[1,2*], Antonia Lenders[1,3], Friedhelm Schwenker[3], Daniel A. Braun[3] and David Fischer[4]

From The 10th DACH+ Conference on Energy Informatics
Virtual. 13-17 September 2021

*Correspondence:
arne.gross@ise.fraunhofer.de
[1]Fraunhofer Institute for Solar Energy Systems ISE, 79110 Freiburg, Germany
[2]IMTEK, Faculty of Engineering, University of Freiburg, 79110 Freiburg, Germany
Full list of author information is available at the end of the article

Abstract

The transformation of the energy system towards volatile renewable generation, increases the need to manage decentralized flexibilities more efficiently. For this, precise forecasting of uncontrollable electrical load is key. Although there is an abundance of studies presenting innovative individual methods for load forecasting, comprehensive comparisons of popular methods are hard to come across.
In this paper, eight methods for day-ahead forecasts of supermarket, school and residential electrical load on the level of individual buildings are compared. The compared algorithms came from machine learning and statistics and a median ensemble combining the individual forecasts was used.
In our examination, nearly all the studied methods improved forecasting accuracy compared to the naïve seasonal benchmark approach. The forecast error could be reduced by up to 35% compared to the benchmark. From the individual methods, the neural networks achieved the best results for the school and supermarket buildings, whereas the k-nearest-neighbor regression had the lowest forecasting error for households. The median ensemble narrowly yielded a lower forecast error than all individual methods for the residential and school category and was only outperformed by a neural network for the supermarket data. However, this slight increase in performance came at the cost of a significantly increased computation time. Overall, identifying a single best method remains a challenge specific to the forecasting task.

Keywords: Forecasting, Machine learning, Electrical load

Introduction

To mitigate the effects of climate change and protect the environment, Germany set a goal to increase its share of renewable energy in the power generation to 80% by 2050 (Bundesministerium für Wirtschaft und Energie 2017). However, since renewable energy generation from sources such as wind or sun is highly volatile, accurate forecasts of non-controllable electrical load are necessary to flexibly manage and achieve demand-supply balance.

Load forecasting is divided into three types depending on the forecasting horizon: short-term load forecasting (STLF), which is used as a term to denote forecasts horizons of up to one week ahead, medium-term load forecasting (MTLF) ranging from one week to one year ahead and long-term load forecasting (LTLF), which predicts load profiles of one year and more (Hahn et al. 2009). MTLF is necessary for fuel supply planning and maintenance and LTLF is crucial for power systems planning (Kyriakides and Polycarpou 2007). STLF is relevant for day-to-day operations of power systems such as energy trading in deregulated markets and unit dispatching or energy management on the individual building or household level (Gajowniczek and Zabkowski 2014).

Another way to categorize load forecasting besides the forecasting horizon is the level of aggregation of load profiles. The discrepancy in forecasting performance can be substantial as forecasting more aggregated load profiles yields lower forecast errors (Sevlian and Rajagopal 2018). Most of the previous research on load forecasting focused on aggregated load data at for example city or country level (Mirowski et al. 2014; Hayes et al. 2015). However, since smart metering data becomes increasingly available, load forecasting at the level of end-users gains increasing attention (Kong et al. 2017; Shi et al. 2017).

Furthermore, work on load forecasting differs in the granularity of the data used. Most previous research focused on hourly data and only 12% of papers reviewed by (Amasyali and El-Gohary 2018) used sub-hourly data. Even though higher data granularity allows for decision-making at higher frequency, two challenges arise from higher sampled data. First, a courser data granularity yields smaller forecasting errors due to the smoothing of load fluctuations. Vice versa a finer data granularity leads to a higher forecasting error. Secondly, a larger data granularity increases the amount of data points presenting a challenge for computationally intensive machine learning methods.

Forecasting methods used in prior studies

Various methods are used for forecasting tasks in the literature. The most popular methods include machine learning (ML) methods such as artificial neural networks (ANNs) and Support Vector Regression (SVR), statistical methods like ARIMA and regression models (Amasyali and El-Gohary 2018; Kuster et al. 2017). A review in (Yildiz et al. 2017) reports that linear regression models are easier to implement, use and understand compared to the 'black-box' ML methods, while the forecasting accuracy of the ML models was higher in the performed study.

Even though ML methods enjoy great popularity in the timeseries forecasting research field, their suitability is still debated (Makridakis et al. 2018a; Hippert et al. 2001); (Yildiz et al. 2017). This debate is attributed by (Hippert et al. 2001) to a rather unsystematic testing of the models, poor predictive results arising from overfitting of the networks and lack of sufficient comparison of benchmarks.

As for the type of load profile, most studies investigate demand prediction of non-residential buildings, namely 81%, compared to 19%, which explored prediction of residential building demand (Amasyali and El-Gohary 2018).

For prediction of non-residential building consumption not one method was found to be superior to all other, as (Penya et al. 2011) for example found a simple auto-regressive (AR) model to be most successful, whereas (Massana et al. 2016) received lowest forecasting errors with SVR.

The same holds for residential buildings. In Kong et al. (2017), the authors found LSTM to yield the lowest forecasting error for day-ahead predictions, whereas (Humeau et al. 2013) reported the linear regression to perform best at the individual household level in their comparison of linear regression, SVR and MLP. However on the aggregated level, SVR yielded the lowest forecast error. For day-ahead forecasts of 27 households, (Lusis et al. 2017) identified the SVR to be the best performing method.

Overall, for residential as well as non-residential building load profiles no consensus on a single best method for STLF could be determined from our literature research. Additionally, most studies used few methods for comparison and oftentimes only one type of data.

Contribution

In this paper, we aim to fill this identified gap by providing a competitive comparison study of popular load forecasting methods on a large database.

We focused our work on STLF, more specifically on day-ahead forecasting of individual electrical load profiles. The database consists of three different categories including education (school load profiles), industry (supermarket load profiles) and residential data. From the literature review, seven widely popular methods from machine learning and statistics were selected for our comparative study. Furthermore, an ensemble technique was applied and a naïve seasonal model was used as a benchmark. The methods include support vector regression (SVR), multiple linear regression (LR), a simple multi-layer perceptron with one hidden layer (MLP), long short-term memory network (LSTM), random forest regression (RF), k-nearest neighbor regression (KNN) and an auto-regressive integrated moving-average model with explanatory variables (ARIMA). Additionally, a forecast is obtained by combining the individual forecasts in a median ensemble.

Our study makes a comprehensive comparison of these popular STLF methods with the intention to determine the most suitable method for each type of the load profile.

To summarize, the contributions of this paper are the following:

- Comparison of seven different methods from ML and statistics as well as one median stacked ensemble for day-ahead forecasting of electrical load profiles
- Three different datasets, including school, supermarket and residential buildings
- High resolution data of 15 minutes sampling
- Adaptation to a specific type of load profile through optimal feature and hyperparameter selection
- Predictions on the building level as well as a comparison to the forecasting error of predicting aggregated data

After this introductory Section, the paper is structured in the following way: "Methods" section introduces the seven algorithms and one benchmark method used in this study and gives a description of the experiment and the dataset, as well as an outline of the preprocessing and hyperparameter tuning steps. "Results" section presents the results and in the "Discussion" and "Conclusion" sections the key findings are put into context and summarized.

Methods

In the following, the seven methods used in this paper are presented. Figure 1 shows the conceptual idea of forecasting in a supervised fashion, where the method can be any ML or statistics method capable of such a supervised approach.

If not specified otherwise the algorithms were implemented in an explanatory fashion, such that explanatory variables (aka features) were used to predict the future electrical load:

$$y_{t+1} = f(\boldsymbol{x_{t+1}}^T) + \epsilon_{t+1}, \tag{1}$$

where the electrical load one step into the future y_{t+1} is predicted with $\boldsymbol{x_{t+1}}^T$, a vector of a multivariate time series at time $t+1$ comprising of n explanatory variables (Hyndman and Athanasopoulos 2018). ϵ is the error. To predict one whole day in the future the multivariate time series needs to be available for this whole future time period as well. Therefore sometimes forecasts themselves need to be utilized. The features used in this work are specified in "Feature extraction and selection" section.

Some methods are intrinsically not explanatory, such as ARIMA models, which are time series forecasting models (see Auto-Regressive integrated moving-Average model with explanatory variables). The time series regression forecasting framework, is defined as:

$$y_{t+1} = f(y_t, y_{t-1}, ..., y_{t-l}, \boldsymbol{x_{t+1}}^T) + \epsilon_{t+1}, \tag{2}$$

where l is the length of the immediate history, which is additionally to the explanatory variables $\boldsymbol{x_{t+1}}^T$ used to predict one step ahead. Since the aim is to do day-ahead forecasts, predicting only y_{t+1} is not enough. To predict all values of one day an iterative multi-step forecasting method was used for ARIMA as well as the LSTM.

The reason for choosing an explanatory forecasting framework were findings from a preliminary experiment, where explanatory forecasting yielded more promising results compared to time series or time series regression forecasting. However, a more in-depth investigation into this topic was out of scope for this paper and will be left to investigate in future work.

Multiple linear regression

Multiple linear regression (LR) assumes a linear relationship between independent explanatory variables $x_1, x_2, ...x_n$ and the dependent variable y:

$$y_{t+1} = \beta_1 x_{1,t+1} + \beta_2 x_{2,t+1} + ... + \beta_n x_{n,t+1} + \epsilon_{t+1}. \tag{3}$$

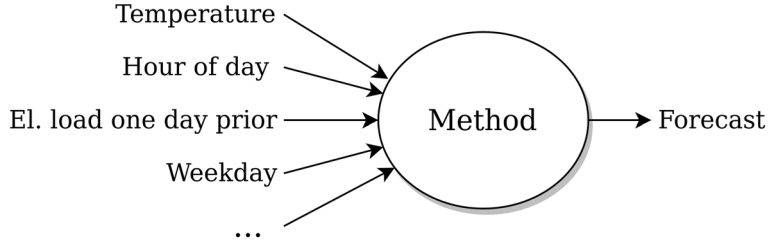

Fig. 1 Conceptual graph of explanatory forecasting. On the left hand the explanatory variables aka features are depicted, which are input to a method. The method finds a function mapping from these input features to a value, which is the forecast

To estimate the β values, the error ϵ_{t+1} was minimized on the training data using the ordinary least squares method.

Auto-Regressive integrated moving-Average model with explanatory variables

A very well known and often used statistical method for forecasting is the auto-regressive integrated moving-average (ARIMA) model (Mirowski et al. 2014; Gross and Galiana 1987; Yildiz et al. 2017).

The ARIMA(p, d, q) model here is defined as:

$$\begin{aligned} y_{t+1}^{(d)} = \beta_1 x_{1,t+1} + ... + \beta_k x_{k,t+1} + \phi_1 y_t^{(d)} + ... \\ + \phi_p y_{t-p}^{(d)} - \theta_1 \epsilon_1 - ... - \theta_q \epsilon_{t-q} + \epsilon_t \end{aligned} \tag{4}$$

where $y^{(d)}$ denotes the d-order difference (Hyndman and Athanasopoulos 2018, Ch.8). After the identification of the model order (the p, q and d value), maximum likelihood estimation is used to find the parameters $\beta_1, ..\beta_k, \phi_1, ...\phi_p, \theta_1, ...\theta_q$.

Support vector regression

The idea of support vector regression, which is an extension of support vector machine, is to find a function, where each prediction y is at most ϵ far away from the target value (Smola and Schölkopf 2004). The support vector regression line is described by

$$y = w^T \phi(x) + b, \tag{5}$$

and the parameters w^T and b are obtained from data using

$$\min_{w,b,\xi_i,\xi_i^*} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*) \tag{6}$$

subject to

$$y_i - w^T \phi(x_i) - b \leq \epsilon + \xi_i^* \qquad i = 1, ..., N \tag{7}$$

$$w^T \phi(x_i) - y_i + b \leq \epsilon + \xi_i \qquad i = 1, ..., N \tag{8}$$

$$\xi_i, \xi_i^* \geq 0, \qquad i = 1, ..., N. \tag{9}$$

where w are the weights, $\phi(\cdot)$ is the transformation of the training data from feature to kernel space and b is the bias. N is the number of samples in the training set. The goal is to get a regression line which is on the one hand flat and on the other hand minimizes the prediction error.

To achieve a flat regression function, one wants to minimize the norm. Deviations from the ϵ-tube are tolerated by the slack variables ξ_i and ξ_i^*. The constant C represents a trade-off between the flatness of the functions and how many predictions can be tolerated outside of the ϵ-tube.

Random forest regression

Random forests (RF) are an ensemble method, comprising a voting committee of n binary decision trees. For each tree a randomly sampled subset of the original training data is used to build it. This is due to single decision trees being prone to overfitting on the training data. The last step is then to average the predictions of each tree to obtain the final prediction of the RF (Bishop 2006, Ch.14).

K-Nearest neighbor regression

In k-nearest neighbor (KNN) regression the Euclidean distance between the query feature vector and every training feature vector is computed. The labels of the k closest vectors are averaged and yield the prediction y_{t+1} (Ahmed et al. 2010).

Multi-layer perceptron

A multi-layer perceptron (MLP) is a feed-forward neural network. One can discriminate three different types of layers: The input layer, where each neuron gets one input dimensions value, the hidden layers, and the output layer.

Apart from the input nodes, all neurons calculate a weighted sum of their inputs including a bias term, apply a differentiable, non-linear activation function and pass the output on to the next layer. In the training process the error between the output of the network and the real labels is minimized by propagating the error gradient back trough the network and updating the weights and biases in the direction of the negative gradient, such that the overall error is decreased (Bishop 2006, Ch. 5).

Long short-Term memory network

In Hochreiter and Schmidhuber (1997) a recurrent neural network architecture called long short-term memory (LSTM) including an input gate, an output gate and a forget gate which dynamically regulate the flow of information, is designed.

The gates can be seen as three filters, which decide what past information is relevant and make it possible to learn long-term dependencies. Naturally the LSTM is intended for sequences and as such the forecasting framework differs from the explanatory forecasting formulation and uses the time series regression framework defined in Eq. 2. For an in-depth description of the LSTM we refer the reader to (Hochreiter and Schmidhuber 1997).

Naïve seasonal model

The naïve seasonal model in this work assumes the electrical load next week is the same as the electrical load of this week. We refer the reader to (Hyndman and Athanasopoulos 2018, Ch. 3.1) for a general naïve seasonal model formulation. Many profiles are very repetitive and follow a strong weekly seasonality. This observation is exploited by the naïve seasonal methods. Therefore, we expect this method to already result in reasonably well predictions, which paired with its straightforward and simple implementation led to our choice to use it as benchmark method.

Median ensemble

In the RF method, a set of decision trees is used to determine the forecast from the resulting ensemble of individual forecasts. Using this idea, the forecasts of all individual methods presented previously, are used to generate an additional forecast. This forecast is obtained by taking the median of all forecasts at every timestep over the horizon. We want to investigate whether this simple combination of predictions in a second level would lead to an increased forecasting performance compared to the individual methods.

Data preprocessing and forecasting setup

Three datasets were used for this study: 19 electrical load profiles of residential buildings, 20 electrical load profiles of schools and 20 electrical load profiles of supermarkets.

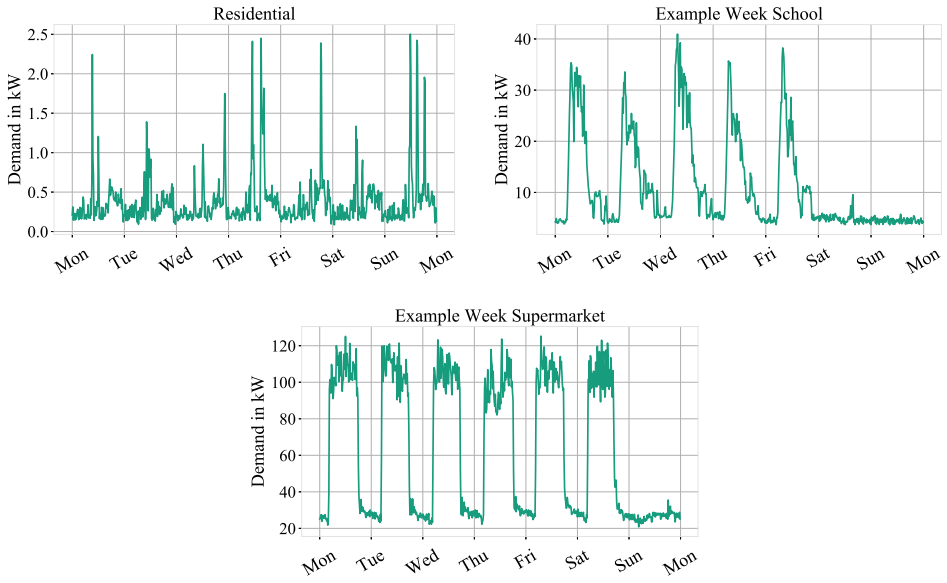

Fig. 2 Example weeks for each category to show the different properties of the electrical load profiles

All time series are from buildings in Germany and were collected in the scope of the Fraunhofer ISE projects synGHD[1] and synPRO[2] (Fischer et al. 2015). The residential and school load profiles each span a time of 18 months, whereas the supermarket load profiles span a time of 10 months. The granularity of the data is 15 minutes. An example week for each profile is shown in Fig. 2. Each building type shows a different degree to which the load profile follows a typical pattern.

To obtain the aggregated load profile of each category the time series in the respective data set were summed up. The resulting aggregated profiles exhibited less fluctuations and a smoother, more regular pattern compared to individual profiles of the specific categories. Especially the aggregated residential load profile showed a more recognisable seasonality in contrast to household demand on the building level.

In all three data categories of individual load profiles a daily seasonality could be observed. Individual school and supermarket time series additionally exhibited a strong weekly seasonality, where the weekend (schools) or Sundays (supermarkets) had substantially lower energy demand compared to workdays. The average autocorrelation values for a lag of 96 (one day prior) of schools and supermarkets are 0.68 and 0.56 respectively and for a lag of 672 the correlation values are 0.78 (schools) and 0.85 (supermarkets) (see Fig. 3).

In the individual household load profiles the weekly seasonality was not as strong. The correlation of lag 96 (one day prior) and lag 672 (one week prior) for the residential category is only at 0.22 for both lag values, which emphasises that individual household only show a weak seasonality. However, the autocorrelation values of the aggregated residential load profile are 0.71 for 'lag 96' and 'lag 672', showing a more distinct daily and weekly

[1] https://www.ise.fraunhofer.de/de/forschungsprojekte/synghd.html
[2] https://www.elink.tools/elink-tools/synpro

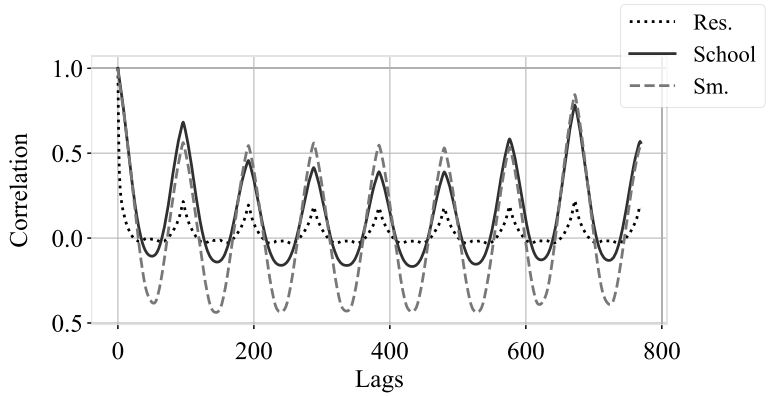

Fig. 3 Average autocorrelation plots for each category. Lag 96 is one day prior and lag 672 is exactly one week prior

seasonality compared to individual residential profiles. Overall, the individual supermarket load profiles display the most regular weekly pattern, followed by school and then the very irregular residential electrical load profiles.

The individual load profiles in the supermarket category differed only slightly from each other apart from different base loads. The school load profiles showed different base loads and also different schools showed a varying course of the school day as to when the school had their lunch breaks and how long a regular school day lasted. Similarly to (Gerossier et al. 2018) we also observed individual residential load profiles to exhibit substantially varying electricity demand.

Data preprocessing

All time series were converted to coordinated universal time. The preprocessing steps are described in the following:

Missing values

Only load profiles with less than six consecutive missing values were included into the datasets. If less than six consecutive values were missing, the values were replaced by linear interpolation.

Outlier detection

The outlier detection included an additive seasonal decomposition with a weekly seasonality. On the residual part of the decomposition an interquantile range (IQR) filter with a lower limit of 2% and an upper limit of 98% detected outliers. The IQR scaling parameter was 1.5 for the school and supermarket data and 3.5 for the residential category as this category has higher and more sudden fluctuations resulting in a more challenging outlier detection (Dawson 2011). The values which were identified as outliers were replaced by linear interpolation between the two neighboring values.

Train/Validation/Test Split

To train the methods a sliding window approach was chosen. The four month prior to the day, which should be predicted, were used as the training set. As validation set six days were selected for 6-fold cross-validation to find the best hyperparameters and features.

These six days were chosen to include two days being either Saturday or Monday since on these days the change from work week to weekend or the other way around occurs. At least two normal workdays and two weekend days had to be included as well as for the school category at least two days had to be school vacation days. The final test set comprised 68 days for the residential and school data and 52 days for the supermarket data as the time series in this category were shorter. The test days included two work weeks (Monday to Friday) from each season. Of these eight work weeks at least two had to be in school vacation times, three weekends from each season and four national holidays. For supermarket data the number of test days is less since spring could not be evaluated as it was in the training data.

Scaling

All datasets were min-max scaled. For this, the training set was scaled. The features of the test set were scaled for forecasting using the same scaler. The final prediction was then inversely scaled back to the original data representation.

Feature extraction and selection

Since we used explanatory forecasting, the features are of major importance to successful prediction. The features which were available belong to three categories: calendar information, history information and weather information. Calendar information includes features describing the seasonal components of the data and features that represent national and school holidays. The seasonalities in the data can be represented by a set of one-hot encoding vectors aka dummy features or by Fourier terms. Seasonality dummy encoding was used to generate two sets of vectors, one set with hour of the week day encoding, featuring 167 vectors and one set with weekday and hour of day encoding, featuring 29 vectors.

The Fourier terms were created according to (Hyndman and Athanasopoulos 2018, Ch. 5.4). Fourier terms allow the reduce the number of features for seasonality compared to seasonal dummy features, depending on the number k of sine and cosine pairs used. The Fourier terms were used to generate features including multiple seasonalities, for example daily, weekly and annual seasonality. Altogether six Fourier term feature sets were extracted. A Fourier feature set with $k = 10$ for weekly as well as daily seasonality was selected most often in the features selection process.

History information included the electrical load value one week prior and the electrical load value one day prior. Weather information was obtained from the European Centre for Medium-Range Weather Forecasts[3] according to the location of the load profiles. The weather information include temperature values and global solar irradiation. Both values could only be obtained hourly, but were upsampled to a quarter-hourly resolution to fit the time series data.

Since not only single features were extracted, but feature sets such as the seasonal dummy features or Fourier terms, filter methods and embedded methods for feature selection were found to be unfeasible (Chandrashekar and Sahin 2014). However, the wrapper method can be applied with feature sets and therefore a wrapper method was utilized, namely sequential forward selection. From each dataset category two time series

[3] https://www.ecmwf.int/

were selected for which a sequential forward selection was used to choose the best features for each method and category individually. The feature selection was limited to use only one feature set describing seasonality (so either one-hot encoding type or Fourier type) and stopped as soon as the forecasting performance of the method did not improve more than 0.1% compared to the last added features. For schools the dummy features for school vacation was set as mandatory beforehand to enforce this information in the final feature set. The features found for the two time series of each category were compared and combined for each method yielding the set of final features used for the hyperparameter tuning (HPT) and forecast of the test set.

The features for the median ensemble comprise the predictions of the individual methods.

Hyperparameter tuning

Hyperparameter tuning was done with one time series from each category, but for every algorithm individually. We used Bayesian optimization to find the optimal hyperparameters for each algorithm (Snoek et al. 2012). The objective function was to minimize the mean root-mean-square error (RMSE) of the six validation days, which were chosen for cross-validation. The LSTM was restricted to use a sequence length of up to one day only in order to limit computational efforts. For the MLP only a single hidden layer was used to evaluate the forecasting performance of the most simple feed-forward neural network.

Evaluation criteria

To compare the results of the final forecast the normalized root-mean-square error (NRMSE) was calculated for all 68/52 days of every time series and averaged. This was done for every method. Therefore, every method in each category has an averaged NRMSE of 68/52 times 20/19 values. The normalized root mean square error has following equation:

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}}{\bar{y}}, \tag{10}$$

where y_t is a measured sample at time t, $\hat{y}_t$ is the prediction at time t, T is the number of samples and $\bar{y}$ is the mean of all observations y. All forecasting measures suffer from specific drawbacks (Shcherbakov et al. 2013) as quantifying the quality of a forecast is not straightforward (Hyndman and Koehler 2006). In this case, the NRMSE is prone to the influence of large outliers.

Furthermore, the time in seconds for every algorithm was measured. The times are divided into fit and predict operations, where fit describes the process of building the model and fitting the training data and predict referring to the step of predicting one day ahead.

Results

After the preprocessing steps, the individual feature selection for every method in every category and the individual tuning of hyperparameters, the final forecast was conducted. For the final forecast the selected test days were predicted for each time series in all categories and with all methods. For these test days the forecasting performance was evaluated with the error measure from Evaluation criteria. Additionally, the final forecast was

performed for the aggregated load profile in each category. The same hyperparameters and features, which were found for the individual building level were used for the forecast at the aggregate level.

Building level results

The forecasting accuracy of the different categories varied greatly. The mean NRMSE of the best methods range from 0.18 in the supermarkets to 1.01 for the residential buildings.

The mean NRMSE for the residential dataset ranged from 1.01 to 1.41 for the different algorithms. This can be seen in Fig. 4 top left. All methods in the residential category have a lower mean NRMSE than the benchmark naïve seasonal method. The best performing method is the KNN (1.01 mean NRMSE) with an improvement of 28.8% compared to the benchmark method in terms of mean NRMSE. A two-tailed paired Student's t-test between the distributions over all buildings of the NRMSE for the KNN-method and the benchmark resulted in a p-value of $2.68 \cdot 10^{-10}$. The observed, greater forecasting performance of KNN is therefore significant.

For the school category the neural networks had the lowest NRMSE, followed by SVR and RF. The simple one layer MLP led to the lowest forecasting error with 0.28 compared to the mean NRMSE of 0.44 of the lowest performing multiple linear regression method. This is the only category where the naïve seasonal method did not consistently show the highest NRMSE compared to the other methods. However, this most likely stems from a disadvantageous selection of features. The improvement of the highest performing MLP compared to the benchmark method is 23.5% in terms of mean NRMSE. The difference

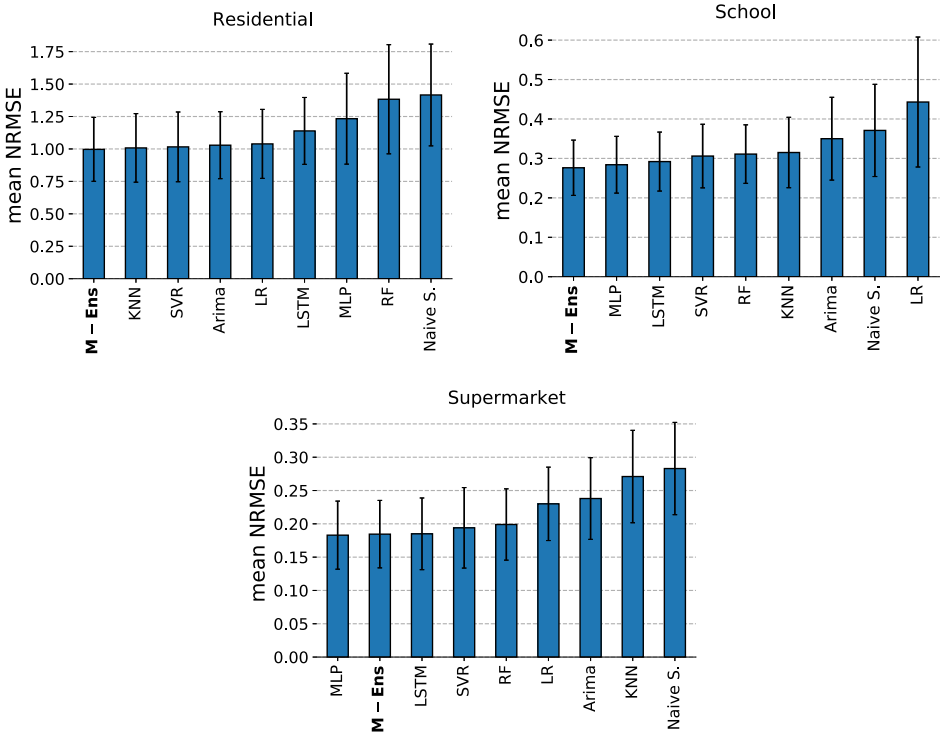

Fig. 4 Mean NRMSE of each method for the test days. On the upper left for the residential category, the upper right the school category and the supermarket category in the bottom row

between the MLP compared to the naïve seasonal method was found to be significant by a paired Student's t-test ($p = 1.08 \cdot 10^{-5}$).

Like the school category, the simple MLP resulted in the lowest forecasting errors with a mean NRMSE of 0.18. The naïve seasonal method again performed lowest (mean NRMSE of 0.28), which leads to an improvement of 35.2% in terms of mean NRMSE. As with the residential and school category the difference between the MLP method and the benchmark was significant according to a paired Student's t-test ($p = 2.58 \cdot 10^{-8}$).

We also found that the individual buildings in one category varied in the magnitude of the forecasting error. For households KNN achieved NRMSE of 0.66 and 1.56 for two different buildings, the same behaviour can be observed for schools (MLP; NRMSE of 0.155 vs. 0.43) and supermarkets (MLP; NRMSE of 0.12 vs. 0.32). This is the reason why the standard deviations in Fig.4 are quite large for all methods. However, the relative order of performance of the individual methods did not change much between the households.

Another reason for the high standard deviation can be explained by breaking down the forecasting accuracy for different day types. The supermarket and school datasets exhibited a higher NRMSE for national holidays compared to every other daytype (weekdays, Saturdays, Sundays and school holidays). In future work this could be counteracted by treating national holidays like Sundays in the feature representation.

Ensemble

The results of creating a median ensemble (M-Ens) from the predictions of all other individual methods, can also be seen in Fig. 4. For the residential and school categories the ensemble resulted in the lowest mean NRMSE (0.996 and 0.276 respectively) and for the supermarket dataset (0.184) only the MLP has a lower mean NRMSE. However, especially for the residential time series it is noteworthy, that the four to five methods with the lowest forecasting error are very close together.

Aggregate level results

The results of the day-ahead forecast of the aggregated load profiles yielded smaller forecasting errors compared to the forecasts of the individual building level (see Table 1). The residential category had the largest improvement from the aggregation of time series with 73,6% in terms of mean NRMSE for the KNN, whereas schools had an improvement of 42.95% compared to the individual building level for MLP. Day-ahead forecasts of supermarkets on the aggregated level improved by 53% compared to the individual building level in terms of mean NRMSE for the MLP.

Table 1 The mean NRMSE of all methods for the individual building level and for the aggregate level

Category	SVR	RF	KNN	MLP	LR	LSTM	ARIMA	Naïve S.
Residential	1.02	1.38	**1.01**	1.23	1.04	1.14	1.03	1.41
Residential Agg.	**0.26**	0.28	0.27	0.29	0.29	0.51	0.28	0.33
School	0.31	0.31	0.32	**0.28**	0.44	0.29	0.35	0.37
School Agg.	0.20	0.19	0.2	**0.16**	0.37	0.18	0.26	0.25
Supermarket	0.19	0.20	0.27	**0.18**	0.23	0.19	0.24	0.28
Supermarket Agg.	0.11	0.11	0.18	**0.09**	0.15	0.10	0.15	0.17

The mean NRMSE of the building level consists of all test days and time series in the respective category. The mean NRMSE of the aggregated load profiles is only comprised of the test days

Computational cost results

In addition to forecasting performance one should keep in mind the computational and time costs needed by a method for real life applications. The computational and hence time resources varied greatly (see Table 2). LSTM and ARIMA took most time and resources, which in some applications could be unsuitable to the task or require custom hardware. The multiple linear regression needed the least amount of time, followed by KNN.

Discussion

The ordering of forecasting accuracy by building category follows the extent to which the electric load profiles follow a repeating weekly pattern (cf. Fig. 3). Other studies (e.g. Makridakis et al. (2018a)) suggest that decomposing the data into trend, seasonality and residuals prior to the forecasting procedure may improve forecasting performance depending on the specifically used model and the data. However, we selected to evaluate the forecast methods using minimal preprocessing to facilitate application of the methods.

Then, the seasonality in the data can easily be represented by the temporal features and the seasonal patterns can be learned by the forecasting method. Therefore, it makes sense that with an explanatory forecasting method the error is smallest for the most regularly repeating load profiles. Overall, the supermarket time series show the strongest weekly seasonality, followed by schools and then residences with only weak weekly seasonality, which is reflected in the magnitude of the forecasting errors. Here, especially the neural networks and LSTM performed well which are often successfully applied to detect signals in data afflicted with noise. As discussed, a regular seasonal pattern can be identified as such a signal.

The load profiles of individual households did not follow a seasonal pattern to the same extent. Furthermore, the residential data exhibited rapidly changing and unique fluctuations of load profiles on the building level. For the class of load profiles with this characteristic, LSTM and MLP were not in the best performing methods. Instead, KNN performed best, but was closely followed by SVR, ARIMA and LR. Since the forecasting errors are close together, this provides no clear indication which method is truly the best for households, although from a computational resources perspective LR was fastest. As the fluctuations were smoothed, the residential category showed the highest increase in

Table 2 Times in seconds for the fit and predict of each method for all categories

	Residential		School		Supermarket	
Method	**fit**	**predict**	**fit**	**predict**	**fit**	**predict**
SVR	2.97 ±0.91	0.02	3.21 ±0.73	0.03	11.83 ±1.97	0.04
RF	1.61 ±0.11	0.01	0.6 ±0.05	0.01	3.53 ±0.17	0.01
KNN	0.25 ±0.04	0.12	0.14 ±0.04	0.01	0.22 ±0.05	0.09
MLP	35.74 ±11.03	0.03	157.95 ±43.95	0.03	32.67 ±25.34	0.03
LR	**0.12 ±0.01**	**0.01**	**0.09 ±0.01**	**0.01**	**0.1 ±0.01**	**0.01**
LSTM	529.47 ±274.19	0.13	412.38 ±169.39	0.1	1076.52 ±292.27	0.11
ARIMA	656.81 ±230.55	0.02	1052.88 ±771.1	0.03	18.84 ±11.13	0.01

Mean NRMSE over all test days and all time series. The error is given as standard deviation. The standard deviation of the predict operation is not given, since it is smaller or equal to 0.02 s for all methods and categories

forecast quality through aggregation of several data sets. Note that the accuracy of MLP is close to the best performing method on the more regular aggregated profiles.

Overall, the forecasting errors of the supermarket and school categories compared to the residential category are substantially smaller, emphasising how challenging the task of residential load prediction on the individual building level is.

Generally, these observations are in line with other studies. SVR was found to perform well by (Massana et al. 2016) for non-residential profiles. In Kong et al. (2017), the authors found that LSTM performs well for residential load profiles. In contrast, LSTM performed particularly weak on residential data in our studies. However, the profiles used in (Kong et al. 2017) were restricted to households with an electric heating system showing a strong daily pattern.

This study was restricted to using forecast algorithms on the unprocessed data without deseasonalizing or detrending which reduces the effort for the practitioner. This implies that a load forecast method should be selected based on the degree to which the profile to be predicted follows a seasonal pattern. However, for all three load profile categories SVR was in the best performing three methods with reasonably small computational costs.

The ensemble did not, in contrast to expectation (e.g. in Makridakis et al. (2018b)), lead to a considerably better forecasting performance and only resulted in minor improvements of forecasting accuracy. We observed a stabilizing effect due to the ensemble, albeit at a high time and computational effort as all individual predictions are necessary to create it.

From the individual methods, LSTM and ARIMA took most time and resources, which in some applications could be unsuitable to the task or require custom hardware. The mismatch in time required by ARIMA in the supermarket category compared to the other two categories can be ascribed to ARIMA using no MA terms in the supermarket case. Both the ARIMA and the LSTM methods used a time series forecasting approach and since our data has a high resolution, including a long immediate history increases computational load significantly.

In most cases, the slight decrease in forecast error of these methods or the median ensemble will not justify the much higher computational time compared with more basic ML methods such as MLP or SVR. However, setting up such a more basic ML model can lead to up to 30 % improvement of forecast accuracy at reasonable computational costs compared to a naïve seasonal approach.

The hyperparameter tuning and features selection was conducted on one and two time series respectively to minimize computational effort. For this, we selected the time series in the medium range of annual consumption compared to the other buildings in the category. We assumed the chosen building load profile would be representative of the category. This is a limitation of the study as especially the residential category could potentially profit from clustering the residential buildings and finding hyperparameters and features for each cluster individually.

Conclusion

This paper gives a comprehensive comparison of popular methods for day-ahead forecasting on individual school, supermarket and residential load profiles. All methods are compared against a naïve seasonal benchmark method. Especially for the residential load, forecasting on the consumer level is challenging compared to the forecasting problem on

aggregated data. However, forecasts on the individual building level are crucial due to the increasing integration of volatile renewable energy generation.

We found that all methods, apart from the LR in the school category, outperformed the benchmark method. Furthermore, the different load profile categories were predictable according to the regularity of their patterns.

The neural networks, especially the MLP, worked best for school and supermarket data. Even though the KNN yielded the smallest forecasting error for households, the forecasting errors of the first four methods were so close together, that it is difficult to pick one best performing on forecasting error alone. For all datasets the SVR performed well and has reasonable computational cost.

The median ensemble narrowly led to the best forecasting performance for the residential and school load profiles and was only slightly outperformed by the MLP method for the supermarket data. However, the computational effort is significantly larger as all individual forecasts must be generated for the ensemble.

We conclude that investing the extra time and computational cost for setting up a learned model compared to the benchmark method is justified as the learned method can achieve a better prediction by up to 30% less error in terms of the NRMSE.

Some ideas for future work are clustering the buildings prior to HPT and feature selection, other ways of pre-processing and a more in-depth investigation of the different forecasting frameworks. The addition of meta data and occupancy behaviour are worth exploring in the future as well.

About this supplement
This article has been published as part of *Energy Informatics Volume 4 Supplement 3, 2021: Proceedings of the 10th DACH+ Conference on Energy Informatics*. The full contents of the supplement are available online at https://energyinformatics.springeropen.com/articles/supplements/volume-4-supplement-3.

Authors' contributions
AG supervised the design and implementation of experiments and discussed the experimental results. AL planned, implemented and carried out the experiments. AG and AL wrote the paper. DF provided the data, directed the project and supervised the process. FS was regularly consulted and supervised design and progress of experiments. DAB supervised and contributed to result analysis selection. All authors read and approved the final manuscript.

Funding
This research was supported by the German Ministry for Economic Affairs and Energy (BMWi) via SynGHD (03ET7534A). Publication funding was provided by the German Federal Ministry for Economic Affairs and Energy.

Availability of data and materials
Data is not published due to legal restrictions but will be made available on request

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Fraunhofer Institute for Solar Energy Systems ISE, 79110 Freiburg, Germany. [2]IMTEK, Faculty of Engineering, University of Freiburg, 79110 Freiburg, Germany. [3]Ulm University, Institute of Neural Information Processing, 89081 Ulm, Germany. [4]greenventory GmbH, 79108 Freiburg, Germany.

Published: 13 September 2021

References
Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H (2010) An empirical comparison of machine learning models for time series forecasting. Econ Rev 29(5-6):594–621
Amasyali K, El-Gohary NM (2018) A review of data-driven building energy consumption prediction studies. Renew Sust Energ Rev 81:1192–1205
Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, Berlin

Bundesministerium für Wirtschaft und Energie BMWI (2017) Das Erneuerbare-Energien-Gesetz. https://www.erneuerbare-energien.de/EE/Redaktion/DE/Dossier/eeg.html. Accessed: 19 Feb 2020
Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
Dawson R (2011) How significant is a boxplot outlier? J Stat Educ 19(2)
Fischer D, Härtl A, Wille-Haussmann B (2015) Model for electric load profiles with high time resolution for german households. Energy Build 92:170–179
Gajowniczek K, Zabkowski T (2014) Short term electricity forecasting using individual smart meter data. Procedia Comput Sci 35:589–597
Gerossier A, Girard R, Bocquet A, Kariniotakis G (2018) Robust day-ahead forecasting of household electricity demand and operational challenges. Energies 11(12):3503
Gross G, Galiana FD (1987) Short-term load forecasting. Proc IEEE 75(12):1558–1573
Hahn H, Meyer-Nieberg S, Pickl S (2009) Electric load forecasting methods: Tools for decision making. Eur J Oper Res 199(3):902–907
Hayes B, Gruber J, Prodanovic M (2015) Short-term load forecasting at the local level using smart meter data. In: 2015 IEEE Eindhoven PowerTech. IEEE, New York. pp 1–6
Hippert HS, Pedreira CE, Souza RC (2001) Neural networks for short-term load forecasting: A review and evaluation. IEEE Trans Power Syst 16(1):44–55
Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Humeau S, Wijaya TK, Vasirani M, Aberer K (2013) Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. In: 2013 Sustainable Internet and ICT for Sustainability (SustainIT). IEEE, New York. pp 1–6
Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practice. OTexts, Monash University, Australia
Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
Kong W, Dong ZY, Jia Y, Hill DJ, Xu Y, Zhang Y (2017) Short-term residential load forecasting based on lstm recurrent neural network. IEEE Trans Smart Grid 10(1):841–851
Kuster C, Rezgui Y, Mourshed M (2017) Electrical load forecasting models: A critical systematic review. Sustain Cities Soc 35:257–270
Kyriakides E, Polycarpou M (2007) Short term electric load forecasting: A tutorial. In: Chen K, Wang L (eds). Trends in Neural Computation. Springer, Berlin. pp 391–418
Lusis P, Khalilpour KR, Andrew L, Liebman A (2017) Short-term residential load forecasting: Impact of calendar effects and forecast granularity. Appl Energy 205:654–669
Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and machine learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3):1–26
Makridakis S, Spiliotis E, Assimakopoulos V (2018) The M4 Competition: Results, findings, conclusion and way forward. Int J Forecast 34(4):802–808. https://doi.org/10.1016/j.ijforecast.2018.06.001
Massana J, Pous C, Burgas L, Melendez J, Colomer J (2016) Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. Energy Buildings 130:519–531
Mirowski P, Chen S, Ho TK, Yu C-N (2014) Demand forecasting in smart grids. Bell Labs Tech J 18(4):135–158
Penya YK, Borges CE, Fernández I (2011) Short-term load forecasting in non-residential buildings. In: IEEE Africon'11. IEEE. pp 1–6
Sevlian R, Rajagopal R (2018) A scaling law for short term load forecasting on varying levels of aggregation. Int J Electr Power Energy Syst 98:350–361
Shcherbakov MV, Brebels A, Shcherbakova NL, Tyukov AP, Janovsky TA, Kamaev VA (2013) A survey of forecast error measures. World Appl Sci J 24(24):171–176
Shi H, Xu M, Li R (2017) Deep learning for household load forecasting—a novel pooling deep rnn. IEEE Trans Smart Grid 9(5):5271–5280
Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222
Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., Red Hook. pp 2951–2959
Yildiz B, Bilbao JI, Sproul AB (2017) A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renew Sust Energ Rev 73:1104–1122

Publisher's Note