Title

Short-Term Deterministic Forecasting of Individual Household Electricity Consumption Using the Hungarian Algorithm

Authors

McClone, Graham
Botman, Lola
Khurram, Adil
et al.

Publication Date

2024-05-09

DOI

10.1109/td47997.2024.10555926

Short-Term Deterministic Forecasting of Individual Household Electricity Consumption Using the Hungarian Algorithm

Graham McClone, Lola Botman, Adil Khurram, Bart De Moor, and Jan Kleissl

***Abstract*—This work proposes a new approach for improving one day ahead point forecasting of stochastic individual household electricity consumption. The focus is tackling the double peak penalty effect and improving peak predictions. Each prediction is generated by comparing a household's energy usage of the seven days leading up to the target day with all seven day periods from all households in the dataset. The households with the closest consumption patterns are then used to create the forecast. The proposed method selects nearest neighbors in a similar manner as in the kNN algorithm. However, it utilizes the Hungarian algorithm to extend this approach to allow for comparisons between consumption values that occur at different times. A case study using an open dataset composed of electric consumption data from 100 Irish households demonstrates that this method improves performance of RMSE over kNN by up to 4.5% and 10.6% for persistence forecasting.**

***Index Terms*—Hungarian Algorithm, Point Forecasting, Electricity Consumption, kNN**

I. Introduction

Short-term forecasting of individual household electricity consumption is an exceedingly challenging task that is relevant for a variety of applications including grid congestion management and infrastructure investments [1]. Due to the stochastic behavior of individual households and burst-like patterns in high resolution data, point forecasting of household electricity consumption profiles is challenging. A critical challenge is the double peak penalty effect [2] that arises when the magnitude of a peak is correctly forecasted but the time when the peak occurs is shifted by a few time steps. Using regular error metrics, in this case, results in large forecast errors as the magnitude of the peak is counted twice, once at the forecasted time and then again at the time of actual occurrence. Because of this, most methods will tend to under-predict peaks or avoid predicting peaks at all. However, in congestion forecasting, a peak prediction of the correct magnitude but with a small delay is more useful than no peak being predicted.

G. McClone and A. Khurram are with the Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, 92037 USA (UCSD); email:{gmcclone,akhurram}@ucsd.edu.

L. Botman and B. De Moor Fellow IEEE, SIAM, IFAC are with the research group STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics within the department of Electrical Engineering (ESAT), KU Leuven, Belgium; email: {lola.botman,bart.demoor}@kuleuven.be.

J. Kleissl is the Director of the Center for Energy Research in the Department of Mechanical and Aerospace Engineering at UCSD, email: jkleissl@ucsd.edu.

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program, from the KU Leuven Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), from the European Research Council under the European Union's Horizon 2020 research and innovation programme (ERC Adv. Grant, grant agreement N°885682) and a FWO travel grant for long stay abroad (V408523N).

This work proposes the novel use of the Hungarian algorithm [3] to improve forecasting of household energy consumption. The proposed method uses the minimum linear assignment of a cost matrix which is constructed using the squared difference between the latest observed data and data from the historical database, at multiple time steps. The assignment provides a metric that is used to select the most similar households to perform a day ahead prediction.

A. Prior Work

Load forecasting has been studied for many years and can be classified based on forecast horizon, i.e., how far in the future the prediction spans, and spatial granularity, i.e., the aggregation level of the load, ranging from individual households to countrywide aggregated loads. Individual household energy consumption forecasting has been receiving more attention recently given accessibility to smart meter data. With increasing levels of penetration of renewable energy resources, short-term energy consumption forecasting is becoming an essential part of future grid planning and operation [4].

The authors of [5] perform large-scale aggregated short-term load forecasting using multiple linear regression for big data. In [6], the authors perform an aggregated short-term load forecast for heterogeneous buildings. However, the aggregated forecasts do not provide information about individual behavior which is critical in congestion planning.

In [7], the authors discuss the need for more research into individual household consumption, and compare a variety of methods. The existing methods that forecast individual household electricity consumption are data intensive approaches that often require additional exogenous data such as weather [8] or household attributes, e.g., square footage, number of rooms, or heating types. The authors of [9] performed short-term nodal load forecasting with the hybrid use of three machine learning methods that require exogenous data inputs. In [10], the authors developed an energy consumption forecast of multi-family residential buildings using support vector regression. Multi-family residential buildings tend to have larger consumption profiles with reduced stochasticity compared to individual households, as multi-family dwellings average out human behavior more than individual households.

The authors of [11] perform short-term residential load forecasting using a nonlinear auto regressive method with an exponential weight decay function. The data requirements for many of these methods limit their scalability and broad applicability. Additionally, the models proposed are local, i.e., one model is trained per household, this also limits scalability and prevents forecasting for household where no or few historical data is available [12].

The Hungarian algorithm is well known and regularly used for various applications, for example, comparing forecasted individual electric vehicle profiles [13]. Haben et al. [2] utilized the algorithm to create a new error metric that is more robust than standard metrics to address the double penalty effect. The new metric allows peaks to shift within a given time window, to find the best match in magnitude between the forecast and the true data.

However, in [2], the Hungarian algorithm shifts electricity consumption peak values to evaluate forecasts only. The authors of [14] implement a graph-based algorithm that uses a similar approach to compare consumption profiles as in [2], but use the Hungarian algorithm only to compare daily profiles. The algorithm proposed in [14] is designed to reduce running time of the Hungarian algorithm but is only demonstrated with daily comparisons.

B. Present Work and Novelty

This work proposes a novel utilization of the Hungarian algorithm for short-term deterministic predictions of individual household electricity consumption based on historical consumption. The Hungarian algorithm compares arrays in a manner similar to the k-Nearest Neighbors (kNN) algorithm, but allows for shifted time comparisons. For example, given two time series, the kNN algorithm only compares data at 1300h whereas the Hungarian algorithm allows comparison between data at 1300h and 1330h as well. This time shifted comparison avoids the double penalty effect and can match peak magnitudes more accurately. The proposed approach is referred to as the Shifted Peaks (SP) method, in which (i) household consumption patterns are compared via the Hungarian algorithm, and then, (ii) the most similar consumption patterns are selected to build predictions [15].

SP uses only household electricity consumption data and does not require additional data such as weather or household attributes and reduces forecast error metrics over kNN by 4.5%-9.3%. SP is a global technique which utilizes data from all households to make predictions.

The novelty in this work is the utilization of the Hungarian algorithm for day ahead deterministic electricity consumption forecasting.

The rest of this paper is organized as follows. Section II describes each of the forecasting techniques in this study. Section III describes the data and parameters. Section IV presents the results of a case study implementation. Section V is the conclusion.

II. Methods

The proposed SP forecasting method is compared against two benchmarks: (i) The baseline approach called persistence forecasting as described in Section II-A, and (ii) a kNN based forecasting model, detailed in Section II-B. The proposed SP approach is described in Section II-C. In all methods, the forecast is updated daily with the horizon of 24h, time resolution of 0.5h, and 0h lead time. Finally, the error metrics used for the results analysis are included in Section II-D.

A. Persistence Forecasting

The persistence model uses a household's prior day's electricity consumption as a prediction for the current day's electricity consumption behavior. The persistence forecasting relationship is depicted in Equation (1):

$$\hat{y}_d = y_{d-1}, \tag{1}$$

where $\hat{y}_d$ represents predicted electricity consumption of day d and y_{d-1} is the electricity consumption of the previous day, $d-1$.

B. kNN Forecasting

The kNN forecasting approach consists of two steps. First, kNN is used to identify similar time series by making element-wise comparisons between two time series (e.g., comparing the value at 1300h with the other time series' value at 1300h) based on a chosen similarity measure. Second, the most similar time series, i.e., households, are used to make predictions.

There are two inputs to the kNN forecasting approach. The first input, Z, is the vector of the consumption values of the target household spanning seven days prior to the day to be predicted and is defined as,

$$Z = \begin{bmatrix} z_1 & z_2 & \dots & z_M \end{bmatrix}^\top \in \mathbb{R}^M, \tag{2}$$

where M is total the number of time steps in a seven day period. The second input, X, is called historical data and can be considered as the training dataset of size $M \times N$:

$$X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_N \end{bmatrix} \in \mathbb{R}^{M \times N}. \tag{3}$$

The matrix X is constructed using a rolling window technique with a contiguous seven-day window such that each $\mathbf{x}_j$ is defined as,

$$\mathbf{x}_j = \begin{bmatrix} x_{1,j} & x_{2,j} & \dots & x_{M,j} \end{bmatrix}^\top \in \mathbb{R}^M, \tag{4}$$

where $j \in \{1, ..., N\}$. Each $\mathbf{x}_j$ is obtained by shifting the window one day at a time across the entirety of the households' consumption data. The training dataset X includes data from all households. The kNN algorithm then compares the input, Z, of the target household with each column $\mathbf{x}_j$ of X.

The distance ($d^j_{i,l}$) is used to compare entries z_i in Z with $x_{i,j}$ in X,

$$d^j_{i,l} = (z_i - x_{l,j})^2. \tag{5}$$

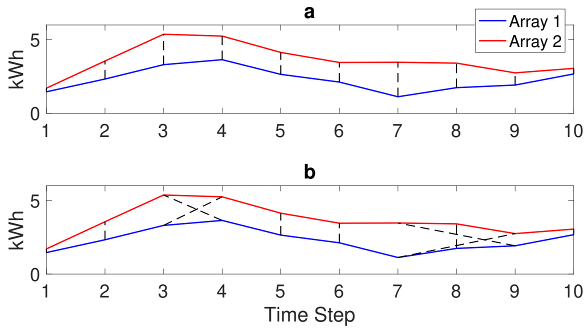

Fig. 1: Comparison of two arrays using (a) kNN with a Euclidean distance of 4.63 kWh, and (b) the Hungarian algorithm with 2 time shifts with a Euclidean distance of 4.50 kWh. Dashed lines connect values that are compared. The Hungarian algorithm minimizes the distance between the arrays by using temporal shifts.

In kNN, the distance $d^j_{i,l}$ is computed for each z_i and $x_{l,j}$, for all, $i = 1, \ldots, M$ and $l = i$ since time shifts are not considered, resulting in the following cost matrix,

$$\mathrm{kNN}_j = \begin{bmatrix} d^j_{1,1} & 0 & \cdots & 0 \\ 0 & d^j_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d^j_{M,M} \end{bmatrix}. \tag{6}$$

The scalar cost $C_{\mathrm{kNN},j} = \sqrt{\sum_{i=1}^{M} d^j_{i,l}}$ is used to obtain the k nearest neighbors to Z from X. The forecast is then obtained by taking the element wise average of the selected nearest neighbors. Figure 1(a) illustrates the one-to-one element-wise comparison between two vectors of 10 time steps with kNN.

This kNN forecasting method has three parameters that can be tuned: (i) the number of time steps used for the array comparison, i.e., M ; (ii) the distance metric used to determine the similarity between two arrays, i.e., the Euclidean distance; and (iii) the number of nearest neighbors.

C. Shifted Peaks Forecasting

Shifted Peaks uses the same inputs as kNN but includes time shifted comparisons in the cost matrix. Specifically, the cost matrix (SP_j) includes off-diagonal entries given by,

$$\mathrm{SP}_j = \begin{bmatrix} d^j_{1,1} & d^j_{1,2} & \cdots & d^j_{1,M} \\ d^j_{2,1} & d^j_{2,2} & \cdots & d^j_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ d^j_{M,1} & d^j_{M,2} & \cdots & d^j_{M,M} \end{bmatrix}. \tag{7}$$

In Equation (7), the i-th row represents the comparison between z_i of Z and the time shifted entries of $x_{l,j}$ of $\mathbf{x}_j$, with $l = 1, \ldots, M$.

The Hungarian algorithm [3] solves a linear assignment problem that minimizes over the sums of the distances in SP_j, such that each element of Z is paired to a unique element of $\mathbf{x}_j$. The scalar cost $C_{\mathrm{SP},j}$ is the sum of those distances. In kNN, the cost is computed by using only the diagonal of the matrix. Selecting elements off the diagonal to minimize the cost means allowing shifts between the two arrays. The off-diagonal values in Equation (7) represent the shifted comparisons illustrated with two time step shifts in Figure 1(b).

In the proposed method, the SP_j matrix is not fully populated, as this implies the capability to minimize the cost by shifting a consumption value an unreasonable number of time steps. For example, a Tuesday at 1100h would be able to shift to a Thursday at 1300h. We introduce a new parameter in order to limit the shifting time frame to between 0.5-2 hours such that the general timing of a peak and its magnitude are accounted for, keeping its usefulness for congestion forecasting. The number of sub and super diagonals that are in the matrix is equal to the number of time steps the algorithm is allowed to shift on each side of the current time step.

Next, as in the kNN approach, the nearest neighbors with the lowest cost are used to make predictions. The day following each selected set of seven days $\mathbf{x_j}$ is averaged, element-wise to make predictions of the target day.

SP has four parameters that can be tuned: (i) the number of time steps used for the comparison, i.e., M; (ii) the distance metric used to determine the similarity between households; (iii) the number of nearest neighbors; and (iv) the size of the time frame in which peak shifting is allowed, we quantify this time frame in terms of time steps, i.e., 1, 2, 3 or 4 time steps corresponding to a time frame of 0.5 h, 1 h, 1.5 h and 2 h respectively.

D. Error metrics

Conventional error metrics are used to evaluate the forecasts. RMSE and mean absolute error (MAE), as defined in [16], are computed between the true consumption and the predicted consumption, for each method discussed in Sections II-A through II-C. RMSE and MAE are evaluated for varying values of k between 0 and 50 for both kNN and SP and for the SP method allowing for time frame parameter shifts of 1 to 4 time steps.

III. Case Study

A. Data

The data used in this work is provided by the Commission for Energy Regulation (CER) of Ireland. It is available via the Irish Social Science Data Archive (ISSDA) [17]. This dataset consists of the electrical consumption values of over $5,000$ businesses and households in Ireland between July, 2009 and December, 2010 recorded at half-hourly intervals. There are thus 48 data points in one day of recording. The non-residential profiles and profiles with incomplete data records have been removed from the dataset consistent with [1] and [18]. A GitHub repository for processing this data is available at [19]. A total of 100 randomly selected households are used for forecasting in this work.

TABLE I: RMSE and MAE results for each forecasting method described in Section II. SP and kNN are evaluated for a varying number of nearest neighbors. SP is evaluated for a varying number of allowed time shifts, e.g., SP 1 indicates that a shift of one time step is allowed. The best performing method per number of nearest neighbors emboldened. The best performing method overall is highlighted in gray.

(a) RMSE

	Persistence	kNN	SP 1	SP 2	SP 3	SP 4
1 NN	0.743	0.888	0.886	0.832	0.807	0.805
5 NN		0.750	0.734	0.718	0.723	0.709
10 NN		0.736	0.721	0.693	0.701	0.697
20 NN		0.713	0.716	0.681	0.679	0.677
50 NN		**0.695**	**0.703**	**0.674**	**0.668**	**0.664**

(b) MAE

	Persistence	kNN	SP 1	SP 2	SP 3	SP 4
1 NN	0.373	**0.429**	**0.429**	**0.405**	**0.392**	**0.392**
5 NN		0.507	0.492	0.477	0.476	0.467
10 NN		0.526	0.507	0.483	0.485	0.482
20 NN		0.526	0.521	0.486	0.483	0.480
50 NN		0.526	0.528	0.494	0.484	0.479

The case study individually forecasts 14 sequential days of 30 households, training on historical data from 100 households. The training data for the initial forecasted day is composed of the first 39 days of the dataset for all households. The training data then increases by one day when forecasting the next day, (e.g., from 39 days to 40 days when forecasting the second day). The testing data is composed of the individual forecasted days between 40 to 53 for the 30 forecasted households. The case study uses a small subset of the data available. An increase in the amount of data used should increase SP and kNN accuracy. Implementing SP with a larger subset of the data available is an area of future research.

Each weekly consumption profile is scaled between $[0, 1]$ using MinMax normalization as to allow comparison across the entire recording period and across all households regardless of the magnitude of the consumption, which might be dependent on household sizes or weather data.

B. Method parameters

SP was implemented with time shifts of 1, 2, 3, and 4 time steps, i.e., a time frame of half an hour up to 2 hours, in order to determine the number of shifts that produces the most accurate forecast.

IV. Results and Discussion

Table I(a) and (b) respectively depict the RMSE and MAE results averaged over the test set. We let two parameters vary: k the number of nearest neighbors for both the kNN forecasting and SP model, and the number of time shifts allowed. The RMSE for kNN and SP models decreases as the number of nearest neighbors increases, while the MAE increases. SP 4 with 50 nearest neighbors reduces RMSE by 10.6% from persistence forecasting and a 4.5% reduction is achieved compared to kNN with 50 nearest neighbors for all households in the study. Figure 2 illustrates these trends and shows that SP 3 and SP 4 significantly improve forecast error over kNN for all nearest neighbor cases.

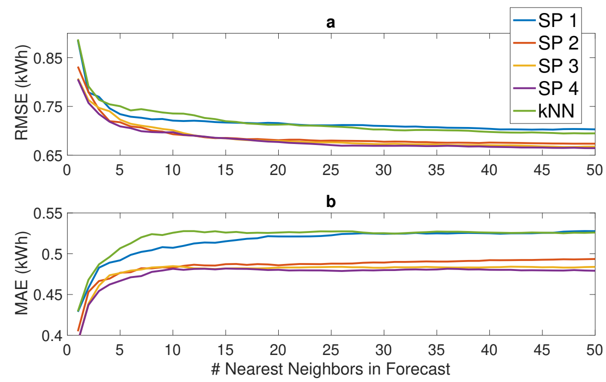

Fig. 2: (a) RMSE and (b) MAE for kNN and SP methods as a function of number of nearest neighbors ranging from $k = 1 : 50$.

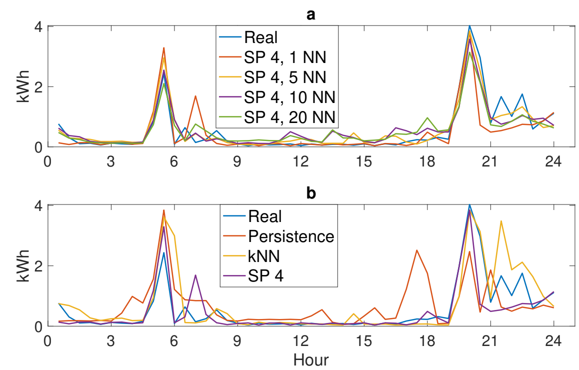

Fig. 3: (a) SP forecast with 4 time step shifting comparing k =1, 5, 10 and 20 versus real data. (b) Persistence, kNN and SP with 4 time step shifting and k =1 versus real data.

Figure 3(a) compares SP 4 with $k = 1, 5, 10, 20$ nearest neighbors for one of the days in the test set. It can be seen from figure 3(a) that peak values decrease in magnitude with an increase in the number of nearest neighbors, due to averaging. Figure 3(b) compares persistence, kNN and SP 4 forecasts with $k = 1$ for the same day in the test set. Finally, figure 3(b) illustrates that SP 4 predicts electrical consumption behavior more accurately than kNN or persistence methods.

Figure 4 illustrates how SP 4 performs best for RMSE and persistence performs best for MAE. In figure 4, the RMSE and MAE are plotted for each individual household with the persistence model on x-axis and SP 4 on y-axis. Households represented by points above the green line suggest higher error for SP than persistence, while below the line suggests the opposite. Points that are further away from the green line represent households in which one method strongly out-performed the other. The SP and kNN forecasts can be improved by applying weights to the nearest neighbors prior to averaging and by including additional household profile data which will allow these methods to identify more similar behavior. Weighting of nearest neighbors and utilizing more data will be explored in future work.

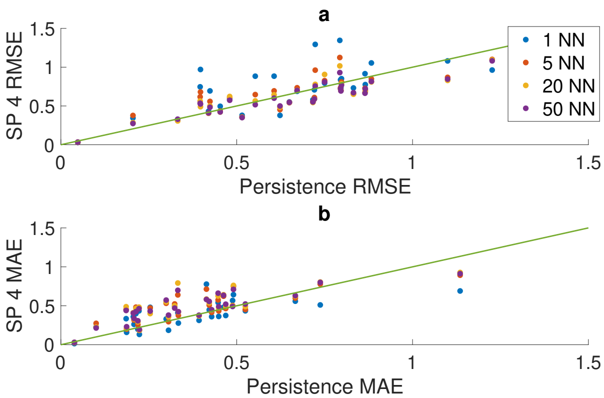

Fig. 4: Scatter plot of the (a) RMSE and (b) MAE of the 30 individual household forecast performance metrics comparing SP with 4 shifting time steps and nearest neighbor values of $k = 1, 5, 20, 50$ vs forecasts with the persistence model.

V. Conclusion

This work proposes the novel use of the Hungarian algorithm to forecast one day ahead household electricity consumption. The proposed SP method is novel in that no prior works have implemented the Hungarian algorithm to forecast electricity consumption. In a case study forecasting electricity consumption behavior for 14 individual days across 30 households, SP with 4 allowable time step shifts (2 hour) and 50 nearest neighbors reduced RMSE from persistence forecasting by 10.6% and 4.5% from forecasting with kNN.

SP has lower data requirements than existing approaches in the literature as it does not require exogenous data such as weather or household attributes. It also has the major advantage of being global, allowing it to leverage the information in the consumption patterns of other households. Future work will expand upon [1] and [15] by implementing the SP method for probabilistic forecasting.

References

[1] L. Botman, J. Lago, T. Becker, O. M. Agudelo, K. Vanthournout, and B. De Moor, "A scalable method for probabilistic short-term forecasting of individual households consumption in low voltage grids," in *2023 IEEE PES Grid Edge Technologies Conference Exposition (Grid Edge)*, 2023, pp. 1–5.

[2] S. Haben, J. Ward, D. Vukadinovic Greetham, C. Singleton, and P. Grindrod, "A new error measure for forecasts of household-level, high resolution electrical energy consumption," *International Journal of Forecasting*, vol. 30, no. 2, pp. 246–256, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207013001386

[3] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109

[4] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on lstm recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841–851, 2019.

[5] A. Y. Saber and A. K. M. R. Alam, "Short term load forecasting using multiple linear regression for big data," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.

[6] A. Bellahsen and H. Dagdougui, "Aggregated short-term load forecasting for heterogeneous buildings using machine learning with peak estimation," *Energy and Buildings*, vol. 237, p. 110742, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778821000268

[7] Y. Peng, Y. Wang, X. Lu, H. Li, D. Shi, Z. Wang, and J. Li, "Short-term load forecasting at different aggregation levels with predictability analysis," in *2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 2019, pp. 3385–3390.

[8] M. Savi and F. Olivadese, "Short-term energy consumption forecasting at the edge: A federated learning approach," *IEEE Access*, vol. 9, pp. 95 949–95 969, 2021.

[9] D. Lu, D. Zhao, and Z. Li, "Short-term nodal load forecasting based on machine learning techniques," *International Transactions on Electrical Energy Systems*, vol. 31, no. 9, p. e13016, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/2050-7038.13016

[10] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy," *Applied Energy*, vol. 123, pp. 168–178, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261914002013

[11] F. Abbas, D. Feng, S. Habib, U. Rahman, A. Rasool, and Z. Yan, "Short term residential load forecasting: An improved optimal nonlinear auto regressive (narx) method with exponential weight decay function," *Electronics*, vol. 7, no. 12, 2018. [Online]. Available: https://www.mdpi.com/2079-9292/7/12/432

[12] M. Grabner, Y. Wang, Q. Wen, B. Blažič, and V. Štruc, "A global modeling framework for load forecasting in distribution networks," *IEEE Transactions on Smart Grid*, pp. 1–1, 2023.

[13] G. McClone, A. Ghosh, A. Khurram, B. Washom, and J. Kleissl, "Hybrid machine learning forecasting for online mpc of work place electric vehicle charging," *IEEE Transactions on Smart Grid*, pp. 1–1, 2023.

[14] N. Charlton, D. V. Greetham, and C. Singleton, "Graph-based algorithms for comparison and prediction of household-level energy use profiles," in *2013 IEEE International Workshop on Inteligent Energy Systems (IWIES)*, 2013, pp. 119–124.

[15] L. Botman, J. Lago, T. Becker, K. Vanthournout, and B. De Moor, "A global probabilistic approach for short-term forecasting of individual households' electricity consumption," *Submitted to IEEE Transaction on Smart Grids*, 2023.

[16] W. H. Stephen Haben, Marcus Voss, *Core Concepts and Methods in Load Forecasting*, 1st ed. Springer Cham, 2023.

[17] Commission for Energy Regulation (CER), "Cer smart metering project-electricity customer behaviour trial, 2009-2010," 2012, [dataset]. 1st Edition. Irish Social Science Data Archive. SN: 0012-00.

[18] M. Arpogaus, M. Voš, B. Sick, M. Nigge-Uricher, and O. Dürr, "Probabilistic short-term low-voltage load forecasting using bernstein-polynomial normalizing flows," in *ICML 2021, Workshop Tackling Climate Change with Machine Learning, June 26, 2021, virtual*, 2021. [Online]. Available: https://www.climatechange.ai/papers/icml2021/20/paper.pdf

[19] M. Arpogaus, "Short-term probabilistic load forecasting using conditioned bernstein-polynomial normalizing flows." [Online]. Available: https://github.com/MArpogaus/stplf-bnf