Machine Learning-Based Prediction of Distribution Network Voltage and Sensors Allocation

Preprint

Alvaro Furlani Bastos,[1] Surya Santoso,[1] Venkat Krishnan,[2] and Yingchen Zhang[2]

[1] *University of Texas at Austin*
[2] *National Renewable Energy Laboratory*

To be presented at the 2020 IEEE Power and Energy Society General Meeting (IEEE PES GM)
Montreal, Canada
August 2–6, 2020

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-5D00-75247
March 2020

Machine Learning-Based Prediction of Distribution Network Voltage and Sensors Allocation

Preprint

Alvaro Furlani Bastos,[1] Surya Santoso,[1] Venkat Krishnan,[2] and Yingchen Zhang[2]

[1] *University of Texas at Austin*
[2] *National Renewable Energy Laboratory*

Suggested Citation
Furlani Bastos, Alvaro, Surya Santoso, Venkat Krishnan, and Yingchen Zhang. 2020. *Machine Learning-Based Prediction of Distribution Network Voltage and Sensors Allocation: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-5D00-75247. https://www.nrel.gov/docs/fy20osti/75247.pdf.

Conference Paper
NREL/CP-5D00-75247
March 2020

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Electricity (DE-OE0000876). The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

Machine Learning-Based Prediction of Distribution Network Voltage and Sensor Allocation

Alvaro Furlani Bastos and Surya Santoso, *The University of Texas at Austin, Austin, TX, USA*,
Venkat Krishnan and Yingchen Zhang, *National Renewable Energy Laboratory, Golden, CO, USA*

***Abstract*—Increasing penetration levels of fast-varying energy resources might negatively affect power system operation. At the same time, sensor deployment throughout distribution networks improves system awareness and enables the development of new and advanced voltage control solutions. Such control techniques rely on accurate prediction in anticipation of voltage violation scenarios. This paper analyzes various approaches to voltage prediction in a distribution system, and it is shown that combining multiple techniques into a single regressor improves its predictive power. Moreover, a two-step regressor is proposed in which initial predictions based on a global regressor are refined by local regressors; in this case, prediction errors decrease significantly. Additionally, a clustering approach is employed to perform sensor allocation so that only the most influential buses are selected for monitoring without diminishing prediction accuracy.**

***Index Terms*—distributed generation, ensemble regressor, machine learning, sensor allocation, voltage prediction**

I. Introduction

Increasing penetration levels of distributed energy resources (DERs) have led to significant changes in power system behavior, especially because of the widespread deployment of photovoltaics in distribution networks [1]. This makes power systems more dynamic because consumers are now able to act as active users rather than only passive buyers [2]. Consequently, distribution voltage regulation has become more challenging because of the fast fluctuations in DER outputs, which commonly result in voltage violations (either less than 0.95 pu or greater than 1.05 pu) [3], [4]. In scenarios of such fast generation variability, traditional voltage control techniques result in suboptimal performance.

Mitigating voltage violations in a modern electric power grid requires distribution system operators to adopt an active management role [2]. Use of historical data and load/DER forecasts would enable them to predict the near-future behavior of the grid, so that preemptive control actions could be planned in advance to prioritize and better coordinate efforts against voltage violations [4], [5]. Multiple factors contribute to the uncertainty in this voltage prediction process, such as load size, DER generation, tap setting of step-voltage regulators, switched capacitor banks, and network topology changes.

This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Electricity (DE-OE0000876). The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Large-scale deployment of smart meters in distribution networks presents new opportunities for voltage prediction and regulation under high DER penetration levels. Therefore, this paper presents a data-driven approach to predict voltage behavior in distribution systems. Multiple machine learning techniques and power/voltage measurements at distribution buses are used for learning a mapping function between load sizes and voltage magnitudes without full knowledge of the circuit topology. These prediction techniques have been analyzed both individually and as an ensemble model in which multiple techniques are combined into a single predictor to exploit their strengths. Also, a two-step prediction approach is proposed to enhance voltage prediction accuracy for conditions near ANSI voltage limit violations. Their prediction performance is assessed through traditional metrics as well as an improved metric that assigns increased importance to the scenarios where a voltage violation occurs. Additionally, strategic buses for monitoring are identified such that high prediction accuracy is achieved. Section II presents the problem formulation, including a description of required inputs and data set generation. Section III presents the results for a case study where multiple strategies for voltage prediction are compared.

II. Problem Formulation

This study analyzes a 69-bus, single-phase network derived from a portion of a Pacific Gas and Electric Company distribution system; it is a radial system with nine lateral branches [6]. The variables of interest are net load sizes (features) and bus voltage magnitudes (targets), where net load size is defined as the difference between power consumption and distributed generation at the corresponding bus.

The input data set contains 20,000 distinct scenarios created in OpenDSS by modifying the load size in each bus. For each scenario, load sizes are updated to $P' = k_1 P$ and $Q' = k_2 Q$, where P and Q are the nominal active and reactive load sizes, respectively. The multiplicative factors k_1 and k_2 are independently drawn from a uniform distribution between -1 and 2; this range of values allows us to represent distributed generation growth above the power consumption (i.e., reverse power flow situations), load variations, or a combination thereof. Note that load power factor assumes a different value in each simulation scenario. Moreover, load sizes are modified independently from other buses in the network to also include extreme low-probability cases that are typically not found within historical load archives.

Data analysis is implemented in Python, including preprocessing steps such as features standardization. The input data

set is randomly split into training and test sets containing 75% and 25% of the number of scenarios, respectively. These subsets are fixed throughout this study to ensure a fair comparison among various voltage prediction techniques.

Various performance metrics are analyzed for each prediction technique; they are based on prediction residuals, i.e., the difference between observed and predicted voltage values. Traditional metrics include mean absolute error (MAE), root mean square error (RMSE), and maximum absolute error (MaxAE) [7]. In these metrics, the importance assigned to each residual depends on only its own magnitude, so that the cost of predictions is uniform across the domain of the target variable [8]. However, even though scenarios with extreme voltage values (less than 0.95 pu and greater than 1.05 pu) are infrequent, their prediction accuracy is more important than the scenarios with voltage around 1.0 pu because extreme cases are the ones where voltage control will need to be deployed. Therefore, performance metrics that incorporate such differentiated importance must also be used for model evaluation, where a higher cost of prediction is associated with rarity [9]. One such metric is the weighted mean absolute error (wMAE), defined as [10]:

$$\phi(y_i) = \left(1 + \exp\left[-s(\Delta y_i - c)\right]\right)^{-1} \quad (1)$$

$$\text{wMAE} = \left[\sum\nolimits_i \phi(y_i)\right]^{-1} \sum\nolimits_i \phi(y_i)|\widehat{y}_i| \quad (2)$$

where y_i and $\widehat{y}_i$ are the i-th observed and predicted voltage values, respectively; the sigmoid $\phi(y_i)$ is the relevance function (s and c control its shape and center, respectively); and $\Delta y_i = |y_i - 1.0|$. The sigmoid tuning parameters are set to $s = 100$ and $c = 0.05$, so that $\phi(0.95) = \phi(1.05) = 0.5$ and $\phi(y_i)$ is nearly 0 for y_i close to 1.0 pu.

III. Results

A. Single Regressors

Bus voltages are initially predicted through a multiple-input, multiple-output linear regression (LR), where predictors are the net load sizes in all 69 buses. The presence of strong multicollinearity between these predictors is assessed through the variance inflation factor (VIF). The maximum value obtained for this factor is 2.86; thus, there is no need to remove any of the predictors (predictor removal is recommended only if its VIF is greater than 5 [11]).

LR is straightforward (no tuning of hyper-parameters) and easily interpretable (the importance of each feature is directly related to the magnitude of its coefficient). Fig. 1 depicts the range of absolute residuals for each bus. Note that this simple approach does not perform well for all buses, where residuals larger than 0.01 pu are observed; however, it provides a baseline accuracy for comparison with other prediction methods. The worst performance accuracy is observed for buses *b61* through *b65*; thus, for simplicity, only bus *b65* (the worst-case scenario) is analyzed in this paper.

Subsequently, various linear and nonlinear regressors are analyzed: AdaBoost (AB), bagging (BG), Bayesian ridge (BR), ElasticNet (EN), extra-trees (ET), gradient boosting (GB),

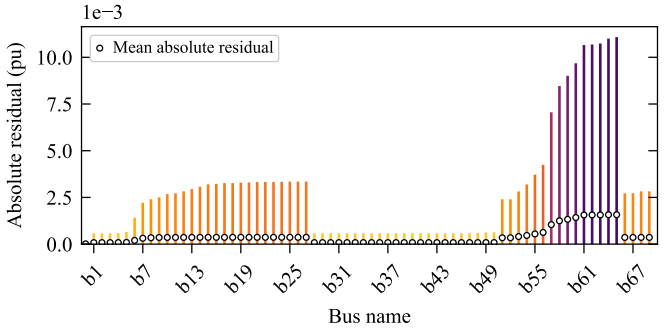

Fig. 1. Range of absolute residuals for voltage magnitudes at each bus obtained through linear regression for all scenarios in the test set.

TABLE I. Performance Metrics for Each Single Regressor

Regressor	Bias	MAE	wMAE	RMSE	MaxAE	Ω_{best} (%)
EN	0.034	1.560	2.178	1.984	10.83	8.38
BR	0.034	1.563	2.180	1.986	10.66	7.22
LR	0.034	1.563	2.180	1.986	10.65	8.78
LL	0.063	1.661	2.290	2.161	13.53	22.44
GB	0.103	1.788	1.909	2.246	11.43	26.94
ET	0.151	3.711	3.903	4.662	20.83	8.48
BG	0.169	4.057	4.264	5.089	20.99	5.58
RF	0.171	4.082	4.298	5.131	21.12	5.82
AB	-0.639	5.526	6.213	6.933	31.42	6.36

Note: units for the performance metrics are 10^{-3} pu.

Lasso-Lars (LL), and random forest (RF). These regressors have hyper-parameters that must be properly set to achieve near-optimal prediction accuracy. Their proper values are obtained through fivefold cross-validation and an exhaustive grid search over user-specified parameter values for each regressor.

Once the hyper-parameters values are obtained, each regressor is fitted on the training data set. Prediction accuracy measures for the test data set are presented in Table I, where Ω_{best} corresponds to the percentage of test samples where each regressor yields the lowest absolute residual among all nine regressors. Note that although some regressors are rather accurate on the average sense, all of them produce residuals larger than 0.01 pu for at least one test scenario. On the other hand, even regressors with a poor overall performance provide the best prediction accuracy for at least a few test samples.

A comparison of prediction residuals for a pair of regressors is presented in Fig. 2. This plot depicts the absolute residuals for all 5,000 test samples obtained through LR and GB. In this pairwise comparison, LR provides a lower absolute prediction residual for 53.34% samples, whereas GB is more accurate for the remaining 46.66% test samples.

B. Ensemble Regressor

As observed in Table I, each regressor provides the best voltage prediction for at least a few test samples; however, it is not possible to determine a priori which regressor would result in the lowest residual for a given unseen input sample. A solution would be to find an optimal combination of regressors that performs well for most cases on average; such an approach is called ensemble prediction.

In ensemble learning, the outputs of multiple learners are combined to create a final output; each single learner, referred

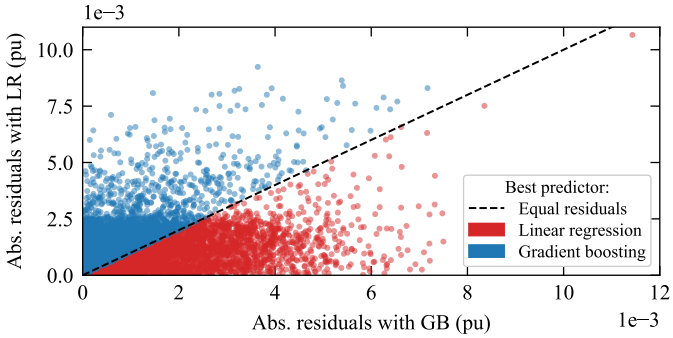

Fig. 2. Comparison of absolute residuals obtained through linear regression and gradient boosting regressor.

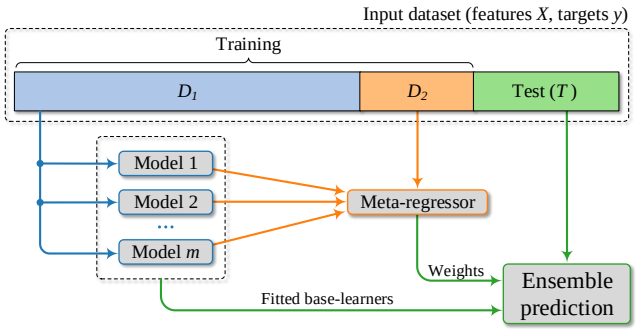

Fig. 3. Concept of ensemble learning and prediction for regression models.

to as a base learner, is trained on the same input data set. The goal is to achieve a higher generalization performance such that the composite model is more accurate than any of the base learners [12]. The usually lower prediction error of ensemble methods is a result of the expansion of the hypothesis space as well as reduction of the base learners' bias and overfitting through averaging [13]. Such an approach allows us to obtain strong learners (high accuracy) from a collection of weak learners, which are more easily obtained in practice. Although increased computational complexity could be presented as a drawback of ensemble methods, note that the base learners can be run independently from each other, and parallel computation in a distributed framework alleviates this limitation.

Highly accurate ensemble predictors require a diverse set of base learners, where they exhibit complementary strengths and their weaknesses are offset by the strengths of other base learners. An approach to obtaining such model diversity consists of generating m base learners through various machine learning algorithms, resulting in an ensemble method known as *stacking* [14]. Then, the ensemble prediction $\widehat{y}$ is computed as a weighted average, as shown in (3):

$$\widehat{y} = \sum_{k=1}^{m} \widehat{\alpha}_k \widehat{y}_k \tag{3}$$

where $\widehat{\alpha}_k$ and $\widehat{y}_k$ are the weight and prediction related to the k-th base learner, respectively. The weights can be interpreted as the importance of each base learner in the final prediction; they also represent a trade-off between model accuracy and the base learners' correlation (the goal is to assign large weights to a subset of base learners that are simultaneously accurate and uncorrelated to each other).

The stacking procedure is illustrated in Fig. 3. The training data set is split into two subsets: the first (D_1) is used for training the m base learners, whereas D_2 and the base learners' predictions are inputs to a meta-regressor that determines the optimal weights. Such a split prevents overfitting of the ensemble regressor, which would be biased toward overfitted base learners if the weights were determined through D_1. These weights are obtained through nonnegative least-squares optimization, as proposed in [15] and shown in (4):

$$\begin{aligned} &\widehat{\alpha} = \text{argmin}_{\alpha} \left\| A\alpha - y^{(D_2)} \right\|^2 \\ &\text{s.t. } \alpha_k \geq 0 \text{ for } k = 1, \ldots, m, \text{ and } \textstyle\sum_{k=1}^{m} \alpha_k = 1 \end{aligned} \tag{4}$$

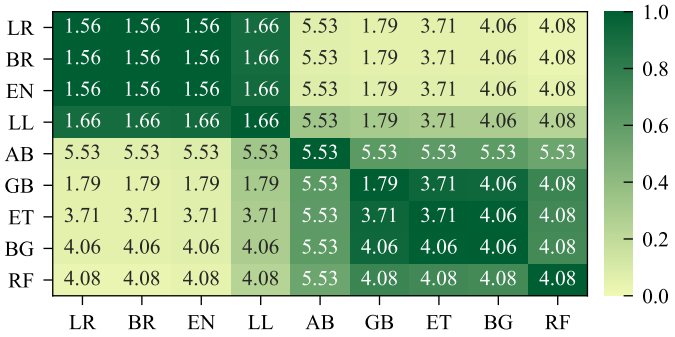

Fig. 4. Heat map depicting the correlation coefficients between predicted voltages for each pair of regressors. The number in each cell represents the largest MAE for the corresponding pair of regressors.

where $A = [\widehat{y}_1, \widehat{y}_2, \cdots, \widehat{y}_m]$ are the predictions for each base learner using D_1 and $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_m]$. The nonnegativity constraint is needed to achieve good generalization performance. Moreover, it allows us to interpret the ensemble model as an *interpolating* predictor (i.e., between minimum and maximum base learners predictions), and it leads to a tractable quadratic programming problem [11].

The regressors presented in Table I are adopted as base learners for a voltage ensemble regressor. As before, 25% of the input data set is reserved for testing; moreover, D_1 and D_2 correspond to 75% and 25% of the training data set, respectively. The weights obtained for each regressor are as follows: 0.5519 for linear regression, 0.4481 for gradient boosting regressor, and zero for all other regressors.

Simultaneously analyzing the base learners' accuracy and their pairwise correlation helps explain why some of them are not part of the ensemble regressor (i.e., their weight is zero). In Fig. 4, each cell's color represents the correlation coefficient between absolute residuals, whereas the annotated number is the maximum MAE between the corresponding pair of base learners. Large weights should be assigned to a subset of base learners that are accurate (low MAE values) and weakly correlated to each other (light green cells). Note that the pair LR/GB satisfies both of these conditions. Moreover, although LR, BR, EN, and LL have the lowest MAE values, only LR is selected for the ensemble regressor; this is because these four regressors are strongly correlated to each other (dark green cells), thus including two or more of them does not provide additional predictive information to the ensemble regressor.

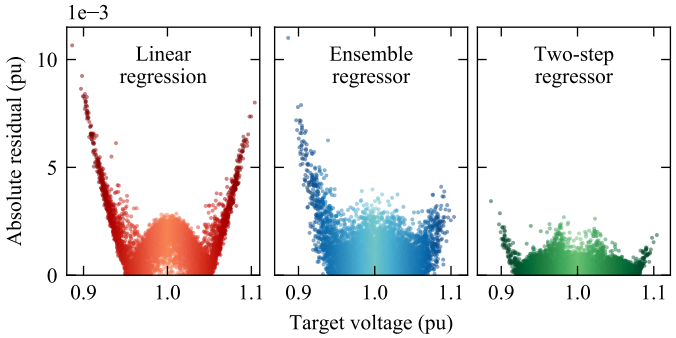

Fig. 5. Distribution of absolute prediction errors versus target voltages for linear regression, ensemble regressor, and two-step regressor.

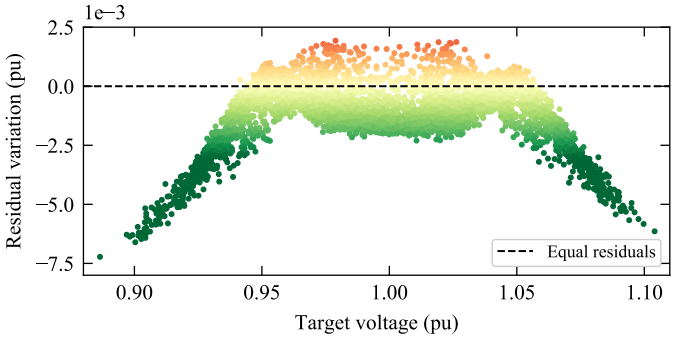

Fig. 6. Residual variation between global and two-step regressors.

Further, a nonzero weight for GB shows that it might be useful to include some less accurate base learners in the ensemble regressor as long as their strengths are complementary to the other selected base learners (i.e., the correlation coefficient between their residuals should be low). As shown in Fig. 2, LR and GB result in low residuals for distinct subsets of the test samples.

Compared to the ElasticNet regression, which is the base learner with the lowest prediction error, the performance metrics for the ensemble regressor improved as follows: MAE, wMAE, and RMSE decreased by 25.9%, 30.4%, and 22.5%, respectively. However, the maximum absolute residual is still larger than 0.01 pu (more specifically, MaxAE = 0.011 pu).

C. Exploring Prediction Errors

The behavior of the voltage residuals is analyzed in Fig. 5, where the absolute residuals are plotted against the target voltage for linear regression and the ensemble model. In both cases, the absolute residuals increase as the target voltage values deviate from 1.0 pu, indicating a systematic bias in these prediction models.

There are two possible causes for the occurrence of this strong prediction bias. First, the input data set is not uniformly distributed across the full range of target voltages (from 0.88 pu to 1.12 pu), thus very low and very high voltage values are underrepresented. Such an imbalance could affect the prediction performance because training of prediction models will be more influenced by the most frequent target values [16]. Further, a global model might not be able to capture the voltage behavior for the entire range of target values.

D. Improved Voltage Predictor: Two-Step Regressor

Given the observations from the previous section, an improved voltage predictor is proposed in this paper. The range of target values is split into three subranges: less than $(1.0-\eta)$ pu, from $(1.0-\eta)$ to $(1.0+\eta)$ pu, and greater than $(1.0+\eta)$ pu, where η is a user-defined parameter. A separate model is fitted for each subrange such that each subrange of extreme voltage values is covered by a different model, and this set of local models is more likely to accurately capture the voltage behavior than a single global model.

Then, voltage prediction is performed in a two-step process:

1) Predict voltage using a global regressor (single model).
2) Based on the initial prediction, select the corresponding local regressor and perform the voltage prediction again.

Because of its simplicity and fast computation, linear regression has been chosen as both the global and local regressor. The parameter η is determined through cross-validation, with test values varying from 0.0025 pu to 0.05 pu in steps of 0.0025 pu; its optimal value is $\eta^* = 0.025$ pu.

Also compared to the ElasticNet regression, the performance metrics for this two-step regressor improved as follows: MAE, wMAE, and RMSE decreased by 69.6%, 79.4%, and 69.6%, respectively; and the maximum absolute residual is only 0.0034 pu (a 68.6% reduction). The absolute prediction errors are shown in Fig. 5. Note that these improvements are significantly better than the ones achieved with the ensemble regressor. Fig. 6 depicts the change in absolute residuals between global and two-step regressors for each test sample (green and red dots portray a decrease and increase in the prediction error, respectively). The two-step regressor improved voltage prediction for 81.72% of the test samples, notably for the scenarios with voltage magnitude far from 1.0 pu.

E. Sensor Allocation

The voltage prediction procedures in the previous sections assume that all buses in the distribution network are continuously monitored; however, deploying measurement devices to only a smaller subset of buses is more realistic. Such a sensor allocation approach represents a trade-off between network observability and fewer measurements at strategic locations.

As a result of the power grid structure, electrically neighboring buses are likely to present very similar voltage behavior; hence, only a few of them would need to be monitored to achieve a high level of network observability.

In this study, similarity between buses is assessed through complete linkage hierarchical clustering. Each bus is initialized as an individual cluster, and the closest pair of clusters is merged at each iteration until only one cluster remains (which contains all buses) [17]. In this iterative process, the distance between clusters is given by the longest distance between two samples in each cluster, such that clusters are merged together only if *all* samples within both clusters are close to each other. Moreover, the distance between samples is measured in terms of the correlation coefficient of their voltage profiles.

The number of clusters plays a central role in data clustering. Two approaches are considered to determine the optimal

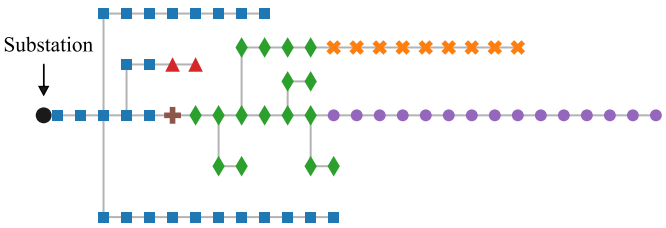

Fig. 7. Clustering of buses based on their voltage profiles. Each combination of color and marker style represents a distinct cluster.

number of clusters for the voltage data set in this study. First, the silhouette score measures the similarity of an object to its own cluster (cohesion) compared to other clusters (separation) [18]. Another approach is gap statistics, which compares the clustered data set to an appropriate reference null distribution, which contains no discernible clusters [19]. Both methods suggest splitting this distribution network into six clusters, as shown in Fig. 7. As expected, neighboring buses belong to the same cluster even though information about their geographic location is not used by the clustering algorithm.

After clustering, buses are ranked according to their importance to the predictor through backward stepwise feature selection. This iterative process is initiated with the full set of buses. At each iteration, the least important bus is removed and the predictor is refitted on the smaller set of buses, where bus importance is given by the absolute value of the corresponding linear regression coefficient [11]. Then, the following procedure is used to select buses for monitoring:

1) Select the most important bus from each cluster for monitoring.
2) Add more buses (in descending order of importance) to the predictor until the desired accuracy is achieved or there is no further improvement in the prediction error.

Fig. 8 illustrates the MAE variation as the number of monitored buses changes (other performance metrics exhibit a similar behavior). In addition to the proposed sensor allocation procedure, this figure shows the performance metrics for an alternative approach: buses are selected in descending order of nominal load size. For both procedures, the combination of multiple local regressors outperforms the global regressor approach. Moreover, note that prediction errors do not decrease significantly until at least one bus from each cluster has been selected. Finally, although sensor allocation based on load size only is simpler, more buses would need to be monitored to achieve high prediction accuracy.

IV. Conclusion

This paper demonstrated the prospect of accurately predicting voltage magnitudes throughout a distribution network. It was also shown that selecting a few strategic buses for monitoring yields high network observability. This data-driven prediction approach can be used proactively for operational purposes by supporting enhanced voltage regulation. In such online applications, voltage predictions are updated as near-real-time data from sensor streaming become available. Also, prediction models can be updated periodically (such as

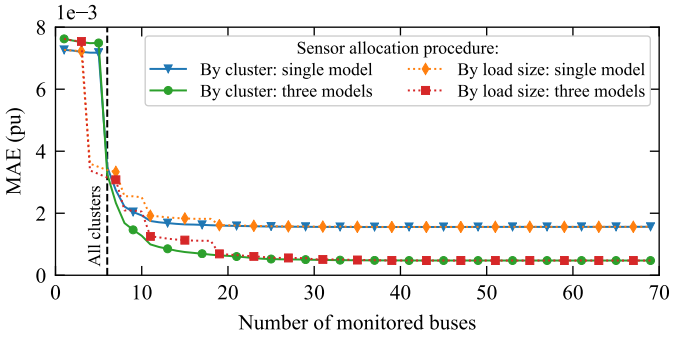

Fig. 8. Relationship between number of monitored buses and MAE.

weekly) with the most recent sensor data, which are likely to better represent the current system's behavior. Additionally, variability in the input data sets can be enriched by including uncertainty factors, such as topology changes, and discrete controller and switch settings.

References

[1] A. Abdel-Majeed, S. Tenbohlen, D. Schollhorn, and M. Braun, "Development of state estimator for low voltage networks using smart meters measurement data," in *Proc. IEEE Grenoble Conf.*, June 2013, pp. 1–6.

[2] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3875–3883, Sept. 2017.

[3] Q. Hou, E. Du, N. Zhang, and C. Kang, "Impact of high renewable penetration on the power system operation mode: A data-driven approach," *IEEE Trans. Power Syst.*, Jul. 2019 (Early Access).

[4] R. Yang, H. Jiang, and Y. Zhang, "Short-term state forecasting-based optimal voltage regulation in distribution systems," in *Proc. IEEE Power Energy Soc. Innov. Smart Grid Tech. Conf.*, April 2017, pp. 1–5.

[5] M. Mokhtar et al., "Predicting the voltage distribution for low voltage networks using deep learning," in *Proc. Int. Conf. Innov. Smart Grid Tech.*, Bucharest, Romania, Oct. 2019, pp. 1–6.

[6] M. E. Baran and F. F. Wu, "Optimal capacitor placement on radial distribution systems," *IEEE Trans. Power Del.*, vol. 4, no. 1, pp. 725–734, Jan. 1989.

[7] J. Zhang et al., "A suite of metrics for assessing the performance of solar power forecasting," *Solar Energy*, vol. 111, pp. 157–175, Jan. 2015.

[8] L. Torgo and R. P. Ribeiro, "Precision and recall for regression," in *Discovery Science*. Springer Berlin Heidelberg, 2009, pp. 332–346.

[9] L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, "SMOTE for regression," in *Progress Artif. Intell.* Springer, 2013, pp. 378–389.

[10] S. C. Tamayo, J. M. F. Luna, and J. F. Huete, "On the use of weighted mean absolute error in recommender systems," in *Workshop on Recommend. Utility Evaluation*, Dublin, Ireland, Sept. 2012, pp. 24–26.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, 2nd ed., ser. Springer Series in Statistics. Berlin, Germany: Springer, 2008.

[12] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. London, UK: Chapman & Hall/CRC, 2012.

[13] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Berlin, Germany, 2000, pp. 1–15.

[14] G. Tsoumakas, I. Partalas, and I. Vlahavas, *An Ensemble Pruning Primer*. Berlin, Germany: Springer, 2009, ch. 1, pp. 1–13.

[15] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, July 1996.

[16] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, June 2004.

[17] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2nd ed. London, UK: Pearson, 2006.

[18] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comp. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[19] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. R. Statist. Soc. B*, vol. 32, no. 2, pp. 411–423, 2001.