INTRODUCING GUARD SMART METERS: VOLTAGE PREDICTIONS AND ITS IMPLICATIONS FOR SMART LV GRID OPERATION

Iker GARCIA RIBOTE
OCT[1] – Spain
iri@ormazabal.com

Roberto SANTANA
UPV/EHU[2] – Spain
roberto.santana@ehu.eus

Patrick MULROY
OCT[1] - Spain
pmu@ormazabal.com

Luis DEL RIO ETAYO
OCT[1] - Spain
lre@ormazabal.com

[1] *Ormazabal Corporate Technology (Ormazabal Analytics)*
[2] *University of the Basque Country*

ABSTRACT

With the advent of new loads and generation on the low voltage grid, voltage fluctuation has increased, especially in active distribution grids with a high penetration of distributed resources and a large deployment of electric vehicles. All this leads to greater uncertainty in future consumption patterns, giving greater importance to predictive models capable of adapting in real time to unexpected variations. Given the current measuring and communication infrastructure of smart meters in Spain, it is not feasible to request real-time data from all consumers. However, it is possible to communicate on-line with a few, which we call guard (or sentinel) smart meters. The results present (I) a novel methodology to select the sentinel meters and (II) their integration in predictive models based on neural networks, improving their ability to adapt and anticipate future states. For the analyses, it has been used (i) data from three groups of meters belonging to the living lab of a DSO, (ii) data from the head-end feeders of the transformer substation and (iii) data from the nearest weather station.

INTRODUCTION

Smart grids play a critical role in the efficient use of energy resources in today's societies. Electric vehicles (EVs) and distributed generation (DG), along with other emerging technologies such as heat pumps, are increasing the complexity and requirements of distribution networks. Along with proper grid reinforcement, it is essential to efficiently manage these new technologies, for which smart charging and flexibility mechanisms play a critical role.

Within the framework of planning and flexible network operation, data analysis techniques are key, especially in the field of consumption forecasting, which allows anticipation and adaptation to the state of the network [1]. The importance of Neural Networks in terms of predictions is well known.

In the literature, different prediction approaches based on Neural Networks can be found, among others, auto-regressive predictive models trained with historical data [2], voltage estimators trying to infer the network model using known or other external variables [3], etc.

However, knowing the expected future development of power grids (EVs, DG, heat pumps, roof PVs, etc.); the voltage fluctuation of consumers may vary in a way that is not expected for these type of models. It is therefore important to have models that use data in "real" time or as close to real-time as possible, so that they can adapt to these variations.

Despite the advanced metering infrastructure (AMI) for low voltage (LV) networks existing in Spain, it is well known that all smart meters (SM) cannot communicate their status in real time due to communication saturation in the grid, however, it is possible to request the status of a few of them, which we call **guard smart meters** or **sentinel smart meters**.

This paper proposes a potentially novel methodology for sentinel smart meter selection, consisting of a modification of the well-known "Maximum Relevance and Minimum Redundancy" (*mRMR*) feature selection method [4] [5]. Regarding consumers voltage prediction, Convolutional Neural Network (CNN) is proposed. CNN combines the real-time data of voltage and current at the Secondary Transformer Substation (TS) and voltage data of sentinel meters, predicting the voltage of the rest of the meters in the network.

The predictions obtained reduce the error to less than half a volt in some cases, improving errors by up to three times compared to predictions obtained without the use of guard meters. In addition, the ability to adapt to external variations, such as, for example, transformer tap changes, or meteorological and socio-cultural events, is demonstrated.

Real historical data from three groups of smart meters have been used. In order to analyze the impact of external variables in the model, data from the corresponding TS feeders and from the nearest weather station have also been used.

GUARD SMART METER SELECTION

The difficulty in selecting sentinel meters lies in finding a few that yield trained models capable of inferring the state of the rest of the meters. The similarity with problems such

as feature selection is obvious, for example, additivity is neither fulfilled here, i.e., the *n* best sentinels individually do not have to be the best *n* as a group.

A well-known methodology for feature selection is *mRMR* [4] [5]. Although in the literature it is generally used for discrete variables [6] [7], we also find examples for continuous variables [4]. In addition, the most studied deployments focus on a single target variable, however, in the case carried out here; there are multiple target variables, in particular, all the meters not selected as sentinels, therefore, some modifications are necessary

Commonly, to measure the relevance of a possible predictor variable, the *F-statistic* of the linear regression between the candidate variable and the target variable is used. For estimating redundancy, the Pearson correlation coefficient (ρ) between the candidate variable and the variables already selected is frequently used. The method to combine both indicators varies, but the main methods are *difference* and *quotient*. After computing a score that integrates the relevance and redundancy, the variables that obtain the maximum value are progressively selected.

Let Ω be the total set of variables, given m variables (m time series, one per smart meter, i.e., $|\Omega| = m$), then, given a candidate variable x_i, a set S of variables already selected and a target variable y, (1) and (2) represent the *mRMR F-test correlation difference* and *F-test correlation quotient* respectively.

$$f^{FCD}(x_i) = F(y, x_i) - \frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|, \quad (1)$$

$$f^{FCQ}(x_i) = F(y, x_i) / [\frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|], \quad (2)$$

where $\rho(x_s, x_i)$ is the Pearson correlation coefficient, and $F(y, x_i)$ is the F-statistic. However, in the problem discussed here, there will not be a single target variable y, but there will be a set of target variables Y, with cardinal $|Y| = |\Omega| - |S| - 1$. Consequently, a small adjustment is necessary to accommodate multiple target variables.

$$\bar{f}^{FCD}(x_i) = \frac{1}{|Y|}\sum_{y_i \in Y} F(y_i, x_i) - \frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|, \quad (3)$$

$$\bar{f}^{FCQ}(x_i) = \frac{1}{|Y|}\sum_{y_i \in Y} F(y_i, x_i) / [\frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|]. \quad (4)$$

To deal with the problem of multiple targets, the F statistic of a multivariate linear regression could also have been used instead of the *mean of one-to-one F-statistics*; however, the method employed (*one-to-one*) is postulated as the best way to maintain the nature of the initial idea.

Since the *forward* method has been used, the first sentinel is selected only on the basis of relevance. Other strategies such as the *backward* method can be found in the literature [7].

Standard *mRMR* methods for continuous variables use the *F-statistic* to measure relevance; however, the magnitude of variability of F is very different from ρ. Some papers propose a way to normalize the values [8] [9] [10], or other relevance measures such as *R-value* [11] [12]. In this paper, it is proposed to use the *R-squared* (R^2) measure, since it is the natural alternative to F and its values are already constrained, without depending on the number of observations or degrees of freedom of the linear regression.

$$\bar{f}^{R^2CD}(x_i) = \frac{1}{|Y|}\sum_{y_i \in Y} R^2(y_i, x_i) - \frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|, \quad (5)$$

$$\bar{f}^{R^2CQ}(x_i) = \frac{1}{|Y|}\sum_{y_i \in Y} R^2(y_i, x_i) / [\frac{1}{|S|}\sum_{x_s \in S} |\rho(x_s, x_i)|]. \quad (6)$$

In this paper, two other simpler methods are also proposed: (i) *random* selection of sentinel meters and (ii) selection of sentinels by voltage level (V). For the second method (ii), we calculate the average voltage of the candidate meters, and we select (and eliminate from the list of candidates) the one with the voltage closest to the median, the one with the maximum voltage, and the one with the minimum voltage consecutively. This loop will be interrupted depending on the number of sentinels we want to select.
In order to compare the presented methods, the following methodology is proposed:

1. The time series with 15-minute voltage measurements of 32 monophasic SM for a period of approximately two years have been used.
2. The selection of 1 to 5 sentinels is proposed consecutively. For the random method, 20 random subsets are selected for each number of sentinels.
3. Data are divided into train-tests with a 90-10 ratio.
4. Each sentinel selection method is applied on the train set. After that, 10 CNN are trained for each case.
5. Predictions are made on the test set and quality is measured with the mean absolute error (MAE). The median of the 10 predictions is calculated. In the *random* case, it is also necessary to compute the median of the 20 subsets.

The use of the MAE error measure is proposed since, in general, traditional error measures, such as mean absolute percentage error (MAPE), cannot reasonably quantify individual load forecasting performance. Some papers

already propose the use of the MAE or modifications of the MAPE [13] [14].

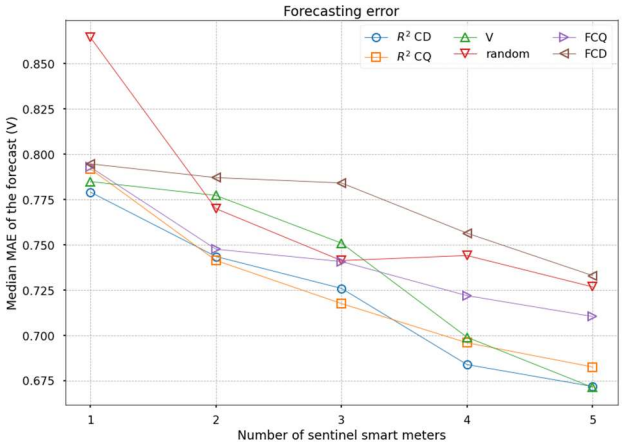

Figure 1: Median MAE over the test set for the predictions corresponding to the 10 CNN for each number of sentinel smart meters and method.

As Figure 1 shows, the R^2CD method achieves the best results, followed by the R^2CQ and V methods, improving considerably with respect to the original methods found in the literature (*FCD*, *FCQ*). Therefore, the methods proposed have turned out to be successful, thus providing new methodologies for feature selection (sentinel smart meter selection in this study).

Note that the median magnitude of the MAE over all predictions of all target variables is less than 0.9V, reaching less than 0.675V in the best case. It is also important to highlight the consistency of the results, having trained 50 CNN for each method, except for the random method, where 1000 have been trained, adding up to 1250-trained and tested neural networks.

It is worth mentioning that the number of target variables changes for each number of sentinels, i.e., for the case of 1 sentinel, the voltage is inferred for the other 31 meters, for the case of 2 the other 30, etc. Therefore, the median is calculated on a set of MAEs of different size for each case. This makes the comparison not entirely objective, but illustrative.

IMPACT OF EXTERNAL VARIABLES

The objective is to assess the impact of variables external to the smart meters on the model, for this purpose, measurements of the corresponding feeder at the head-end TS and meteorological (MET) data have been used.

The possible influence of the TS feeder data seems obvious; in the case of weather data, it may help the model to anticipate the influence of appliances such as heating, ventilation and air conditioning (HVAC). In this TS there is no influence of EV, DG, etc.

Since data must coincide in time, the available dataset has been reduced to one year. As a result, the quality of the predictions worsens; however, they are still sufficient to compare the relative impact.

The external variables used are: (TS) head-end voltage and current and (MET) temperature, wind, humidity, pressure, type of sky and whether the day was a holiday or not.

Again, the selection from 1 to 5 sentinels consecutively has been iterated, it has been considered all possible combinations with external variables, training 10 CNN for each case, and the median MAE has been calculated.

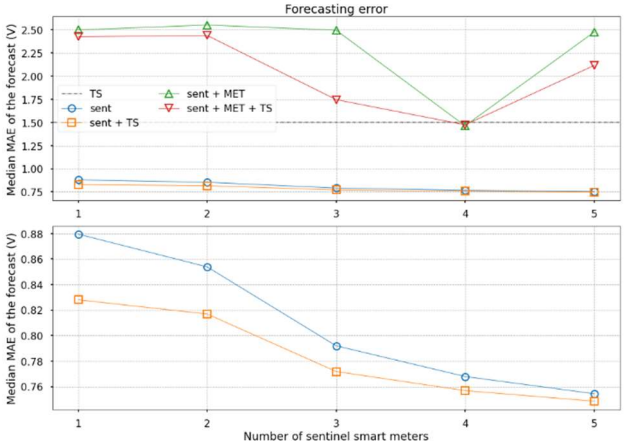

Figure 2: Median MAE over the test set for the predictions corresponding to the 10 CNN for each number of sentinel smart meters and used external variables. The bottom graph shows the zoom on the two best curves.

As shown in Figure 2, introducing MET data significantly worsens the results. This is mainly due to two factors: (i) the available data had a frequency of 30 minutes and was interpolated to 15 minutes and (ii) the data corresponds to a weather station at a distance of more than 5km from the TS. Therefore, we are putting noise into the model. As can be observed in the bottom graph, including the feeder data improves the predictions.

Therefore, the following (*sent + TS*) is postulated as the best methodology: (i) select the sentinel meters using the R^2CD method and (ii) combine the data from these sentinels and the voltage and current at the TS to generate a neural network-based model capable of estimating the voltage of the remaining meters.

MAIN TESTS

In this section, *sent + TS* methodology will be put into practice with 5 sentinel smart meters. For this purpose, three different groups of sentinel smart meters belonging to different phases (R, S, and T) have been selected. Groups *1*, *2* and *3* consist of 26, 24 and 32 monophasic meters respectively.

The following steps have been followed: (i) first, data has been separated into train-test with a 90-10 ratio, (ii) on the train set, the R^2CD method is applied to select the 5 sentinels, (iii) then the neural network is trained, (iv) finally the voltage of the unselected meters is predicted over the test set and the MAE is measured.

The nature of the method lies in exploiting the similarity between groups of monophasic consumers belonging to the same feeder and phase [15]. CNN receives as input a vector of size 7 (the voltage of the 5 sentinels plus the 2 header values –current and voltage- at instant *t*), and returns as output a vector of size *r* (the voltage at instant *t* for the rest of the meters: sizes 21, 19 and 27 for groups *1*, *2* and *3* respectively). Figure 4 shows the process schematically for group *3*.

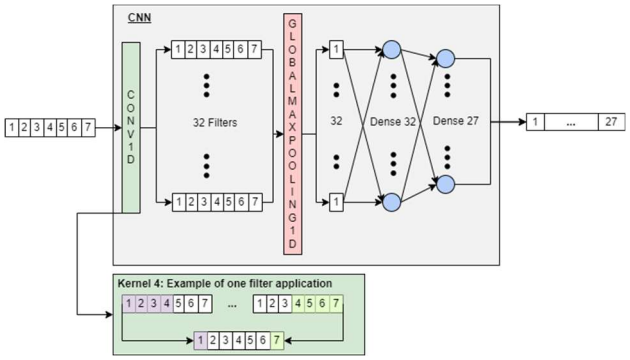

Figure 3: Schematic representation of the used CNN architecture, specifically for case *3* (27 target variables).

Given that predictions are made for multiple users, it is not feasible to study the forecasts one by one; however, a simple way to assess the results is to analyze the distribution of the errors.

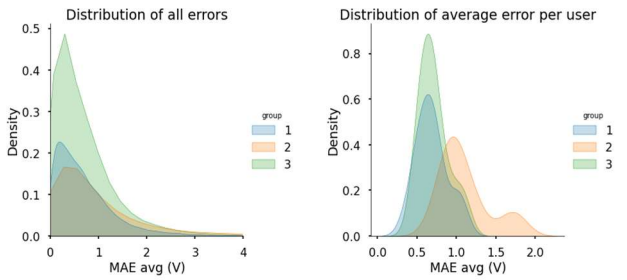

Figure 4: Distribution of all errors (left) and distribution of average error per smart meter (right).

The blue, orange and green curves represent the distribution of errors for groups of 21, 19 and 27 meters respectively. In the left chart of Figure 4, we can see how the distribution of all the errors for the three groups follow *chi-square* distributions, with most of the errors concentrated between 0V and 1V.

However, the chart on the right shows how the distributions of the mean error per user follow a shape more similar to the *normal* distribution. This makes sense, since, according to the *central limit theorem*, the greater the number of independent random variables combined, the closer to a normal distribution. Group *2* appears to have a somewhat different distribution, more similar to a *bimodal* one.

Using the Mann-Whitney U null hypothesis test between the distribution of errors of the groups one-to-one, it has been verified with 99% confidence that the average errors per user distribution of group *2* follows a significantly different distribution from groups *1* and *3*.

Table 1: Mann-Whitney U Test one-to-one p-values

	Groups *1-2*	Groups *2-3*	Groups *1-3*
p-value	7.34×10^{-6}	2.67×10^{-6}	0.37

The main reason is that the predictions for 3 of the 19 meters of group *2* is significantly worse. However, the fact that the group *2* is smaller also has an impact, as stated in [16], where the authors claim that the accuracy of predictions improves as the group size increases.

In any case, these are good results, with most errors concentrated in 0-1V and with average error per user contained in 0.5-1V. It should be noted that in some cases even the uncertainty range of smart meters measurements is higher. Finally, for illustrative purposes, Figure 5 depicts one of the best predictions obtained for a consumer of group *1*.

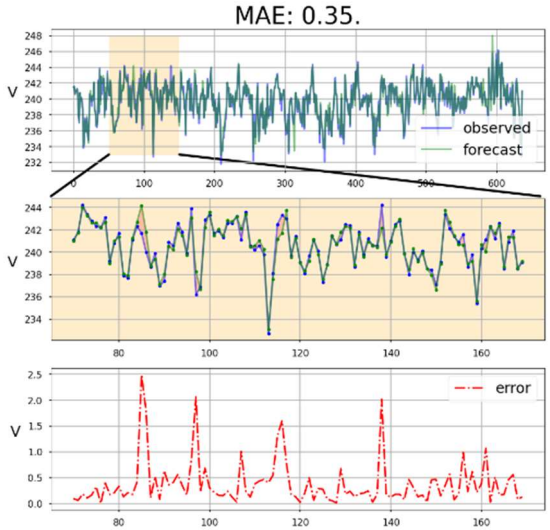

Figure 5: Prediction on a consumer of group *1*.

CONCLUSION

The *mRMR* methods with the necessary adaptations (R^2CD and R^2CQ) are postulated as the best for the selection of guard meters. Convolutional neural networks are able to

efficiently infer the associations between sentinel meters and feeders with respect to the rest of the consumers.

FUTURE WORK

An open line of research would be to use the connection hierarchy of the power grid, which has already been used for distribution load forecasting [17], to exploit the selection of sentinel meters. The association between sentinel selection method and CNN model improvement remains to be studied. Finding the cut-off point in the number of sentinels that contribute significantly to the CNN model can be critical. Another crucial aspect, outliers, remains to be studied. The usefulness of the model may be questioned if the errors are concentrated in extreme voltage values, which are, in short, where flexibility decisions have more weight. To this end, it is proposed to combine this model with a time series classification model capable of predicting future outliers. Work is currently in progress.

ACKNOWLEDGEMENTS

R. S. acknowledges support by the Basque Government (IT1244-19 and project KK-2020/00049 through the ELKARTEK program), and the Spanish Ministry of Economy and Competitiveness MINECO (projects TIN2016-78365-R and PID2019-104966GB-I00).

REFERENCES

[1] M. Bitos, D. Mills and S. R. Ashton, "Utilising forecasting time series data and flexibility services to manage distribution networks," *CIRED,* vol. 1, no. 0757, pp. 2760 - 2763, 20-23 Sept 2021.

[2] M. Bassi, L. F. Ochoa and T. Alpcan, "Calculating voltages without electrical models: smart meter data and neural networks," *CIRED,* vol. 1, no. 0785, pp. 1547-1551, 20-23 Sept 2021.

[3] A. Boyd, H. Sun, M. Black and S. Jesson, "Short-term load forecasting using artificial neural networks and social media data," *CIRED,* vol. 1, no. 0631, pp. 2695 - 2699, 20-23 Sept 2021.

[4] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology,* vol. 03, no. 02, pp. 185-205, 2005.

[5] Z. Zhao, R. Anand and M. Wang, "Maximum Relevance and Minimum Redundancy feature selection methods for a marketing machine learning platform," *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA),* pp. 442-452, 2019.

[6] P. Bugata and P. Drotar, "On some aspects of minimum redundancy maximum relevance feature selection," *Science China Information Sciences,* vol. 1, no. 63, pp. 1869-1919, 2019.

[7] P. Hanchuan, L. Fuhui and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, 2005.

[8] J. Che, Y. Yang, L. Li, X. Bai, S. Zhang and C. Deng, "Maximum relevance minimum common redundancy feature selection for nonlinear data," *Information Sciences,* Vols. 409-410, pp. 68-86, 2017.

[9] V. La, N. Thang and Y.-K. Lee, "An Improved Maximum Relevance and Minimum Redundancy Feature Selection Algorithm Based on Normalized Mutual Information," in *2010 10th IEEE/IPSJ International Symposium on Applications and the Internet*, 2010, pp. 395-398.

[10] M. García-Torres, F. Gómez-Vela, B. Melián-Batista and J. M. Moreno-Vega, "High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach," *Information Sciences,* vol. 326, pp. 102-118, 2016.

[11] I. Jo, S. Lee and S. Oh, "Improved Measures of Redundancy and Relevance for mRMR Feature Selection," *Computers,* vol. 8, no. 2, p. 42, 2019.

[12] J. Lee, N. Batnyam and S. Oh, "RFS: Efficient feature selection method based on R-value," *Computers in Biology and Medicine,* vol. 43, no. 2, pp. 91-99, 2013.

[13] C.-N. Yu, P. Mirowski and T. K. Ho, "A Sparse Coding Approach to Household Electricity Demand Forecasting in Smart Grids," *IEEE Transactions on Smart Grid,* vol. 8, no. 2, pp. 738-748, 2017.

[14] P. Li, B. Zhang, Y. Weng and R. Rajagopal, "A Sparse Linear Model and Significance Test for Individual Consumption Prediction," *IEEE Transactions on Power System,* vol. 32, no. 6, pp. 4489-4500, 2017.

[15] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau and D. Ernst, "Phase Identification of Smart Meters by Clustering Voltage Measurements," in *2018 Power Systems Computation Conference (PSCC)*, 2018, pp. 1-8.

[16] P. Goncalves Da Silva, D. Ilić and S. Karnouskos, "The Impact of Smart Grid Prosumer Grouping on Forecasting Accuracy and Its Benefits for Local Electricity Market Trading," *IEEE Transactions on Smart Grid,* vol. 5, no. 1, pp. 402-410, 2014.

[17] X. Sun, P. B. Luh, K. W. Cheung, W. Guan, L. D. Michel, S. S. Venkata and M. T. Miller, "An Efficient Approach to Short-Term Load Forecasting at the Distribution Level," *IEEE Transactions on Power Systems,* vol. 31, no. 4, pp. 2526-2537, 2016.