USE OF LINKY SMART METER DATA TO ENHANCE THE DIVERSITY FACTOR ASSESSMENT IN REAL NETWORKS

Guilherme Ramos Milis, Christophe Gay, Marie-Cécile Alvarez-Hérault, Bruno Gourguechon, Raphaël Caire

USE OF LINKY SMART METER DATA TO ENHANCE THE DIVERSITY FACTOR ASSESSMENT IN REAL NETWORKS

Guilherme RAMOS MILIS
Enedis – France
guilherme.ramos-milis@enedis.fr

Christophe GAY
Enedis – France
christophe.gay@enedis.fr

Marie-Cécile ALVAREZ-HERAULT
G2Elab - France
marie-cecile.alvarez@g2elab.grenoble-inp.fr

Bruno GOURGUECHON
Enedis – France
bruno.gourguechon@enedis.fr

Raphaël CAIRE
G2Elab - France
raphael.caire@g2elab.grenoble-inp.fr

ABSTRACT

In this paper, we propose a study of the use of Linky smart meter data to enhance the Diversity Factor (DF) assessment in real networks. The literature focus on how the number of customers per secondary station affects the diversity factor and it distribution. One of our objectives is to bring into discussion other indicators, some local, some general, which could affect DF variations, as we dispose of real French network measurements. The other objective is to develop a model to predict the DF based in these sets of indicators. We validated the interest of our model results by comparing them to one naïve substation-peak sizing method, obtaining ~50% result improvement on some specific test cases. Thanks to this work, we are able to bring into discussion the importance of smart meter data for network planning.

INTRODUCTION

Electricity networks are going through transformations in depth due to the interconnection of renewable energies and other distributed energy resources surveyed through smart meters. One clear example is when gasoline vehicles commercialization will be stopped in Europe. The latter scenarios show that half of the French vehicle fleet shall be converted to electric by 2035, which might result in a massive grid connection of charging infrastructures. From the perspective of DSOs, these changes could represent a large investment cost to reinforce the existing network and/or design new ones. The coincident peak load at the MV/LV substation level is an important value for distribution network planner. A better assessment of this value allows a better estimate of the hosting capacity and therefore, at first glance, to avoid/delay/minimize network investments. A traditional way to estimate the coincident peak load is the use of the Diversity Factor (DF). DF reflects the fact that neighbourhood customers do not consume with the same profile over time. It can be calculated at the secondary substation as the ratio between the sum of individual peak loads and the peak-aggregated load [1], as shown by equation 1:

$$DF = \frac{\sum(individual\ peak\ loads)}{Peak\ aggregated\ load} \quad \text{(Eq. 1)}$$

Based on this definition, the literature focuses on analysing the distribution of DF from random samples of customers groups. The work presented in [2] supports that DF is normally distributed. Using a different normality indicator [3] test, the results indicate that DF follows a gamma distribution and depends on the customers group size. Meanwhile, the authors in [4] use a generalized extreme value distribution to model the coincident peak load as a random variable that depends of mean energy consumption. However, in French scenario, the massive roll out of Linky smart meter allows us to recover the individual peak loads daily. Complementary, Enedis disposes of some grid data measurements at the secondary (MV/LV) substation level, such as load aggregated time series. These two-information make us able to compute the real Diversity Factor for several substations. Since all our electricity data come from known French networks, we are able to study how different indicators, some local (temperature, geographical area), and some general (type of clients) affect the DF. This brings into the discussion other variables that influence DF besides the number of customers per secondary substations, consequently, expanding our vision about the subject. Our research aims to increase the understanding of how the local and general factors influence DF and to develop a model to predict it based on these indicators.

DATA

Data Recovery

The data collection period is from beginning of July 2021 to end of June 2022. A full year period allows us to have a better representation of the weather variations as the French load is classically largely thermal-dependent. All the data used in this work comes from Enedis databases and are divided into three categories:

Electricity Data

Load time series from 89 secondary substations are selected. These substations are located in different regions of France and cover well the country's climatic diversity. They serve 12352 customers, which the individual peak loads and energy consumption measured by Linky Smart Meter are collected.

Customer Data
The 12352 customers are grouped into categories : Basic Residential, Time of Use (TOU) Residential, Basic Professional, TOU Professional, Public Light and Industrial. The associated contractual power value subscription is also known.

Meteorological Data
Temperature, nebulosity from météo-france stations, climate zone classification established by 2012 thermal regulation norm [7] and types of territory established by Enedis (Urban, Rural, Semi-Urban and Dense) are collected.

Diversity Factor Computation

The DF is then computed for all period, on a daily basis, using the formulation presented in [1]. With our data, the equation can be writen as follows:

$$DF = \frac{\sum Linky\ peak}{Peak\ load\ at\ MV/LV\ substation} \quad \text{(Eq. 2)}$$

Figure 1 shows the results of this computation for a given secondary substation every day and reflects the fact that the DF is not constant throughout the year. It is possible to note seasonal and daily variations. Thus, it is most suitable for our analysis to have a year collection period.

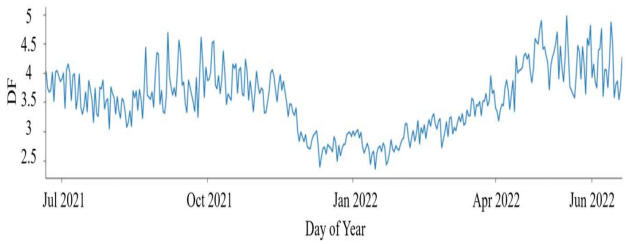

Figure 1: DF values computed for the entire period

A histogram showing all DF values for all the select substations is provided in Figure 2:

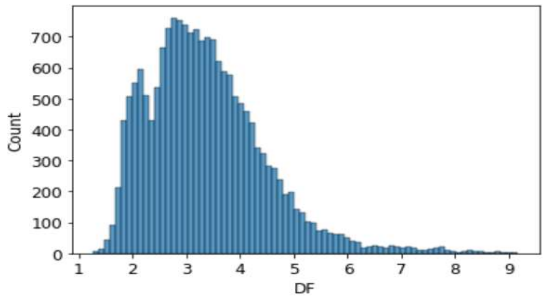

Figure 2: DF histogram for the 89 secondary substations

It is possible to observe that there are very few large DF values, which correspond to the discussions in literature [2-4]. It is possible to observe some values between six and nine with support the fact that a better assessment of DF can allow a better estimation of hosting capacity. Indeed, the higher the DF is the more room we have for renewable connections.

Indicators Definition

Using the collected data, a set of indicators listed in Table 1 is created for each substation to observe the differences between them. In total, there are 56 indicators constructed from our data collection.

Categories	Quantity
1 - Customer categories rate	One indicator for each customer category (*)
2 - Subscription power rate	One indicator for each subscription value (**)
3 - Daily total individual peak load by customer categories rate	(*)
4 - Daily total individual peak load by subscription power rate	(**)
5 - Daily total energy consumption by customer categories	(*)
6 - Daily total energy consumption by subscription power rate	(**)
7 - Thermo-sensitivity	Two indicators
8 - Daily mean temperature	One indicator
9 - Daily maximum and minimum temperature	Two indicators
10 - Daily mean nebulosity	One indicator
11 - Type of territory	Four indicators
12 - Climate zone classification	Eight indicators
13 - Weekday or Weekend	One indicator

Table 1: Set of indicators

MODEL PREDICTION

Figure 3 shows a flowchart of all the steps presented by this section.

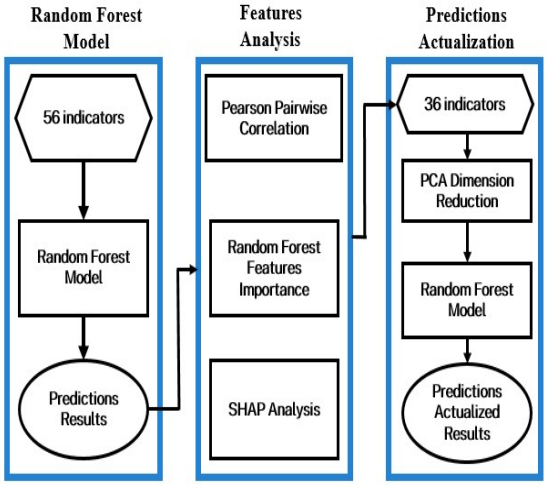

Figure 3: Model prediction flowchart

There are three main steps; the first one is training our model using the indicators, the next one is to analyse the results to understand the features importance, and the last one is to retrain our model based on our analysis

Random Forest Model

Random Forest is the supervised learning method chosen to predict the DF from our selected set of indicators. This traditional algorithm generates a number of selected trees based on the data and combines the output of all the trees [8]. In such manner, it reduces overfitting problem in decision trees and reduces the variance. The focus of this study is to find a powerful but explainable model, to show the importance of daily factors in DF variations. Thus, the random forest fits well with our objectives.

The data set of indicators is split in 80% of train and 20% of test set based on the quantity of MV/LV substations to avoid prediction bias. This division generates 71 MV/LV substations in train set and 18 MV/LV substations in test set. Using all the indicators introduced in the section above, the model predictions have a symmetric mean percentage error (SMAPE) of 16 %. Figure 4Figure 4 presents a scatterplot showing the Predictions as functions of DF values. It is possible to observe that the model has difficulties in predicting higher DF values. This is logical because there are few values in this range in our data collection, so the model will have less training examples. In addition, we could observe that the highest errors in our predictions are related to a few substations where the presence of industrial customers is important.

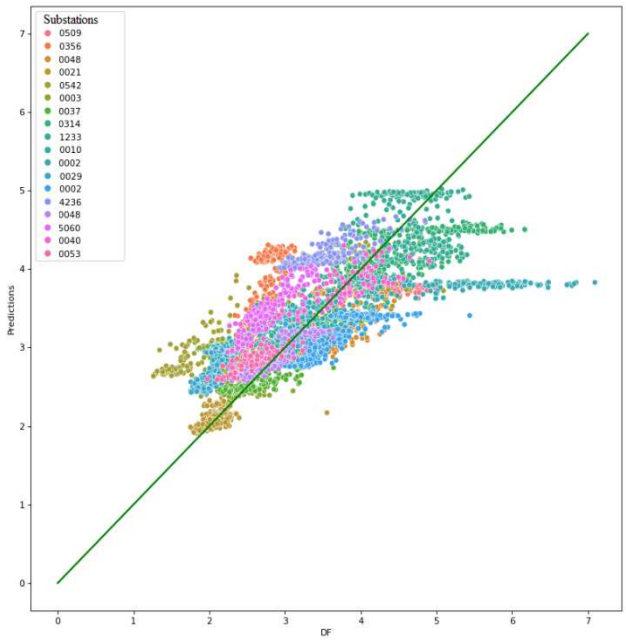

Figure 4: Scatterplot of Prediction x DF values

Features Analysis

We note that some of our indicators seem to express the same type of information. It was possible to confirm this by using Pearson pairwise correlation method. The indicators related to the customer categories rate are strongly positively correlated. The same was observed for the indicators related to the subscription power rate. Therefore, we prefer to keep only the daily peak rate information in both cases. Twelve features were discarded. The next step is to look the model features importance. The random forest toolbox of scikit-learn library offers the possibility to analyse the importance of features. This analyse is known as Gini Importance [9]. It is defined as the total decrease in node impurity weighted by the probability of reaching that node averaged over all trees of the ensemble. The result of this analysis shows that eight indicators can be discarded because their importance for model results are low.

Thenceforth, we perform a Shapley addictive explanation (SHAP) analysis to interpret the contribution of each indicator to the models output. SHAP is a method based in cooperative game theory, where it computes the contribution of each feature in order to explain the prediction [10]. Figure 5 shows how the average daily temperature contributes to the variation in DF for all substations of our test set. It is possible to observe that the higher temperatures contribute to increase the diversity factor. Otherwise, when temperature decreases it contributes to the decrease of DF.

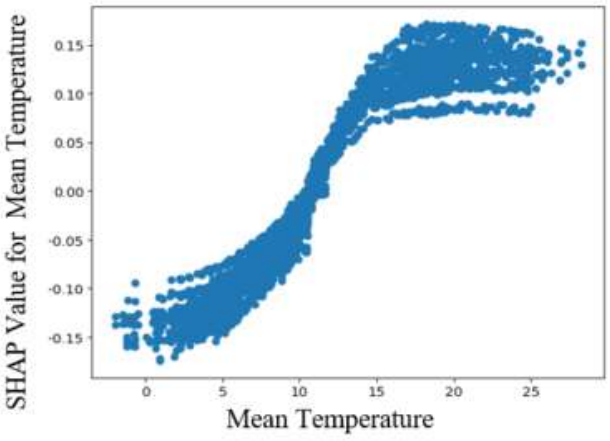

Figure 5: Contribution of average temperature to variations in DF

This can be explained by the fact that part of low voltage consumption is temperature-dependent. In France, households are equipped with heating systems and water boilers that tend to be switch on at similar schedules of the day to the residential customers. In this way, the individual peak demand and the aggregate peak demand are close to each other, reducing the diversity factor. Figure 6 presents the daily total peak rate by TOU Residential customer type. TOU Residential are those clients that have an energy tariff based on the time of day when energy is consumed. There are the “full hours” where the tariff is more expensive and the “off-peak hours” where the tariff is cheaper. Normally, for these customers the consumption is

higher during "off-peak hours". It can be seen that if a large part of the consumption is due to these clients it contributes to lowering the DF, which corresponds to the fact that they have their peaks close in time.

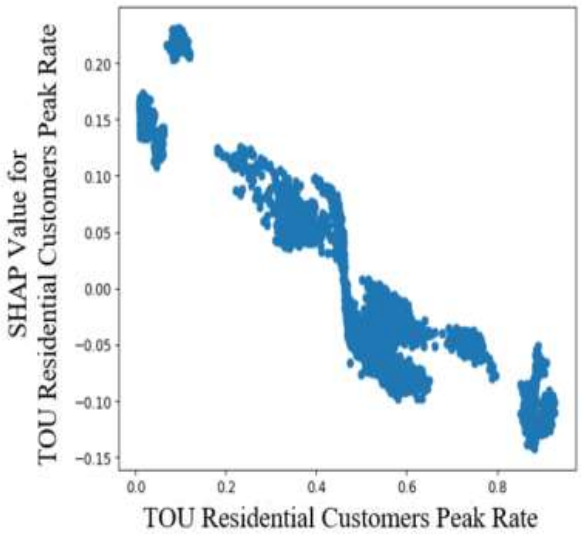

Figure 6: SHAP visualization - contribution of TOU Residential peak rate to variations in DF

After the analysis, the following indicator categories were selected: 3, 4, 5, 6, 8, 9, 10 and 12. This leaves us with 36 indicators.

Prediction Result Actualization

Now, we are able to re-compute the model predictions using the features analyses results. Principal Components Analysis (PCA) [11] is used in order reduce the data set dimensions. PCA is a method that returns the sequence of linear combinations with the greatest variance of variables in a dataset. For this purpose, PCA uses a vector space transformation to reduce the dimensionality of a large data set. Eight components are keep after the PCA. The same substations test set was used to see the impact caused by the changes in our input. Thus, it is possible to compare the models results. The model is then re-trained and predictions give us a SMAPE of 12.9 %. This increase in model performance could be related to the fact that we have small data set with a large number of features. Therefore, there could have some overlearning even for a random forest model.

MODEL ASSESSMENT

The further step to measure the quality of our prediction results is to compare them to other methods. This section presents a comparison between results of a naïve substation-peak sizing method and our model. To do that, we use our DF Prediction to compute the MV/LV substation peak, as expressed by Equation 1. Naïve method is describe by Equation 3:

$$SP = WC * \sum(Psubscriptions) \quad \text{(Eq. 3)}$$

Where, *SP* is the substation peak, *WC* is a weighting coefficients based on the total number of customers in the MV/LV substation and *Psubscriptions* is power value subscriptions. WC values are compute to have the minimal SMAPE between SP and the real peak values for the select number of customer's interval, as shown by Equations 4 and 5:

$$WC_{error} = SMAPE(Real\ Peak \mid SP) \quad \text{(Eq. 4)}$$

$$WC \rightarrow minimization(WC_{error}) \quad \text{(Eq. 5)}$$

This optimization problem is solve using python's scipy library optimize.minimize(method = 'Nelder-Mead'). Table 2 shows the WC values due to the number of clients in the substation.

Number of customers	WC Value
< 50	0.166
50 – 99	0.151
100 – 149	0.118
150 – 249	0.116
> 249	0.104

Table 2: WC values

Figure 7 presents the results of this comparison for our 18 substations test set:

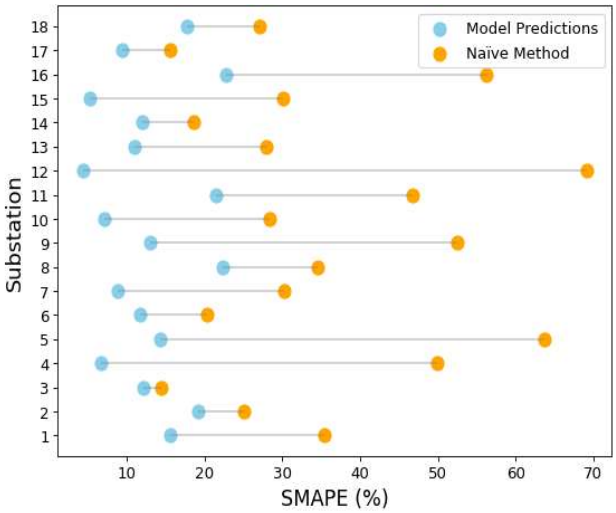

Figure 7: Comparison results

The prediction model has better results in all 18 substations. The average difference between the prediction model results and the naïve method results is 20 %, for this test set. To further investigate our comparison, we decided to extend it to 100 random test sets of 18 substation from our 89 substation tests. Figure 8 shows the average results per test set.

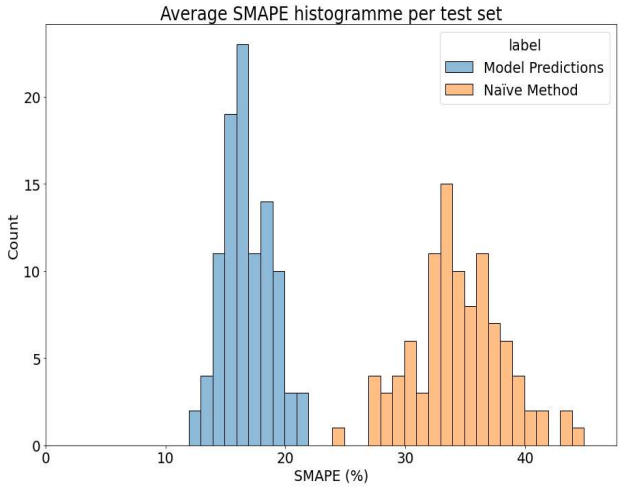

Figure 8: Average SMAPE histogram per test set

The histogram shows the difference between the model predictions results and the naïve method results. In average, the prediction model results have ~50% less error than the naïve method results, for these 100 test sets. If we look at all substations individually, it is possible to observe that for some substations the model predictions have an error in the magnitude of 40% to 60%. Again, this can be explained by the absence of higher DF values in our data set. Random forest models are not capable to extrapolate; they only make predictions based on previously observed values. This comparison helps to understand the quality of predictions in a real life scenario.

CONCLUSION

In this paper, we use Linky smart meter data to increase the assessment of Diversity Factor variations. Using real network data, it was possible to investigate and understand how local and global indicators affect the DF and use this knowledge to develop a supervised learn model to predict this value. Our results and comparison make easy to visualise the importance of smart meter data to the energy planning. We were able to obtain a model with around 17% error reduction on our specific test set, in the estimation of peak-aggregated load. Provide a better assessment of Diversity Factor in LV grids can allow a better estimation of hosting capacity and thus, minimizing / delaying / avoiding network investments.

ACKNOWLEDGMENTS

This study is part of a thesis work on the use of Linky smart meter data to improve load knowledge in LV grids conducted by Enedis, G2Elab and MIAI.

REFERENCES

[1] A. Sargent, R. P. Broadwater, J. C. Thompson and J. Nazarko, 1994, "Estimation of diversity and kWHR-to-peak-kW factors from load research data ", *IEEE Transactions on Power Systems.* vol. 9(3), 1450-1456.

[2] J. Nazarko, R. P. Broadwater and N. I. Tawalbeh, 1998, "Identification of statistical properties of diversity and conversion factors from load research data" *MELECON '98. 9th Mediterranean Electrotechnical Conference*, vol.1, 217-220.

[3] V. P. Chatlani, D. J. Tylavsky, D. C. Montgomery and M. Dyer, 2007, "Statistical Properties of Diversity Factors for Probabilistic Loading of Distribution Transformers" *39th North American Power Symposium*, 555-561.

[4] J. Lee, M. Bariya, and D. Callaway, 2022. “Peak Load Estimation with the Generalized Extreme Value Distribution” *Berkeley Education Technical Report*

[5] M. Nijhuis, N. Vermeltfoort, and R. Bernards, 2019, "Applying smart meter data to low voltage network planning ", *Proceedings CIRED conference*, AIM, vol.1, 210-220

[6] J. Dickert and P. Schegner, 2010, "Residential load models for network planning purposes," *2010 Modern Electric Power Systems*, pp. 1-6.

[7] Réglementation Thermique 2012 : un saut énergétique pour les bâtiments, 2011 [Online]. Available : https://www.ecologie.gouv.fr/sites/default/files/RT%202012%20-%20un%20saut%20%C3%A9nerg%C3%A9tique%20pour%20les%20b%C3%A2timents%20neufs%20-%20Avril%202011.pdf

[8] M. R. Segal, 2004, “Machine learning benchmarks and random forest regression”.

[9] A. L. Boulesteix, S. Janitza, J. Kruppa and I. R. König, 2012, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics" *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discover,* pp. 493-507.

[10] S. M. Lundberg and S. I. Lee, 2017, “A unified approach to interpreting model predictions”, *Advances in neural information processing systems*, 30.

[11] R. Bro and A. K. Smilde, 2014, “Principal component analysis” *Analytical methods*, 2812-2831.