

# Measures for combining prediction intervals uncertainty and reliability in forecasting

Vânia Almeida<sup>1</sup> and João Gama<sup>1,2</sup>

<sup>1</sup> LIAAD / INESC TEC, University of Porto, Portugal  
vania.g.almeida@inescporto.pt,

<sup>2</sup> Faculty of Economics, University of Porto, Portugal  
jgama@fep.up.pt

**Abstract.** In this paper we propose a new methodology for evaluating prediction intervals (PIs). Typically, PIs are evaluated with reference to confidence values. But, these values cannot be considered individually. Higher probability values are associated to too wide intervals, that convey little information and are of no use for decision making. We propose that the comparison should take into account the error distribution (predictions out of the interval) and the maximum mean absolute error (MAE) allowed by the confidence limits. This paper presents a neural network based method for 24-hour-ahead load forecast. It is implemented using customer load data. PIs are compared using two different strategies: (1) dual perturb and combine (DPC) algorithm and (2) conformal prediction. We demonstrated that depending of the real scenario (e.g. time of day) different algorithms perform better. The main contribution is the identification of high uncertainty levels in forecasting. This information can guide the decision makers to avoid the selection of risky actions under uncertain conditions. In contrast, lower errors mean that decisions can be made more confidently with less chance of confronting a future unexpected condition.

**Keywords:** Load forecasting, prediction intervals, neural networks, conformal prediction, uncertainty assessment

## 1 Introduction

In time series forecasting, most research focuses around producing and evaluating point forecasts. Point forecasts are a topic of first-order importance, being easy to compute and understand. However, predictions intervals (PIs) are assuming increasing importance comparatively to these conventional techniques. By definition, a PI is an estimate of an interval in which a future observation will fall, with a certain probability called confidence level.

Similarly to point forecasts, error measures play an important role in calibrating or refining a PI model [1, 2]. Typically, PIs evaluation is focused on the calibration of confidence intervals that indicates the probability for correct predictions. But, confidence values cannot be considered individually. Higher

probability values are associated to intervals that can include extreme prediction errors. And so, a too wide PI conveys little information and is of no use for decision making. Sharpness and resolution are also considered as added value, i.e. the average size and the variability of intervals, respectively. The literature offers some metrics for PIs evaluation. However, a reliable representation based on the error distribution has not yet been studied.

This paper aims to develop a useful methodology for evaluating PIs. The prediction errors are computed as the distance to the upper and lower bounds. Additionally, the Mean Absolute Error (MAE) range is computed, considering that the prediction values are contained within the lower and upper prediction bounds. This value represents the range of 'acceptable' errors, and it is correlated with the interval width. Since many of intervals are asymmetric, and the forecast is not the midpoint of the estimated interval [3], the evaluation of the cost associated to the underestimation or overestimation is also analysed. For the purpose of experimental evaluation, a case study in load forecasts is presented. This is a challenging topic where PIs assume major importance. In order to increase sustainability and optimize resource consumption, electric utilities are constantly trying to adjust power supply to the demand. However, more than providing accurate forecasts, reliable interval predictions are fundamental.

The paper is organized as follows. Section 2 describes related work. The proposed PIs evaluation methodology is formulated in Section 3. Case study description and results are presented in Section 4. Finally, Section 5 concludes this paper and provides guidelines for future work.

## 2 Related Work

### 2.1 Models used in this study

Several strategies can be used to provide PIs. Two strategies are adopted: (1) dual perturb and combine (DPC) algorithm [4] which produces PIs based on the perturbed predictions, and (2) conformal prediction (CP), one of the most promising strategies used to determine precise levels of confidence. Other strategies can be used, such as detailed in [5, 6].

**Dual Perturb and Combine Method** [4] is an efficient method that allows the reduction of the variance exhibited by neural networks (NNs), but also the estimation of the confidence values associated to the predictions. It consists of perturbing each test example several times, adding white noise to the attribute values, and predicting each perturbed version of the test examples. The final prediction is obtained by aggregating all the predictions, implemented as follows:

1. For each input variable in the test set  $x$ ,  $k$  perturbations are performed,  $i = 1, \dots, k$ .

$$x_i = x + \delta_i \quad (1)$$

with  $\delta_i$  white noise  $N(0, \sigma_i^2)$ , where  $\sigma_i$  and  $k$  are user-defined parameters.

2.  $k$  predictions  $\hat{y}_i$  are obtained, and the final prediction  $\hat{y}$  is:

$$\hat{y} = \frac{\sum \hat{y}_i}{k} \quad (2)$$

3. The lower and upper bounds are defined as:  $[\min(\hat{y}_i), \max(\hat{y}_i)]$ .

**Conformal Prediction** uses the past experience to determine precise levels of confidence in new predictions, assuming that the data is identically and independently distributed (*i.i.d.*). CPs have been developed based on several algorithms, such as Support Vector Machines [7], k-Nearest Neighbors [8] or Neural Networks Regression [9]. In this paper a neural networks regression based on inductive conformal prediction (NNR-ICP) is implemented as proposed by Papadopoulos and Haralambous [9]:

1. The training and the calibration set are represented as:  
 Training:  $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$  where  $m < l$   
 Calibration:  $\{(x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \dots, (x_l, y_l)\}$  with  $k=l-m$  elements
2. A nonconformity score is associated with every pair  $(x_{m+i}, y_{m+i})$  in the calibration set. It evaluates how strange the pair is for the trained NNR rule, being defined as:

$$\alpha_i = |\hat{y}_{m+1} - y_{m+1}| \quad (3)$$

where  $\hat{y}_{m+1}$  is the predicted value.

3. Assuming *i.i.d.* distribution, these  $\alpha$ 's are sort in descending order:

$$\alpha_{m+1}, \dots, \alpha_{m+k} \quad (4)$$

4. Finally, the lower and upper bounds are computed according to:

$$(\hat{y}_{l+1} - \alpha_{m+s}, \hat{y}_{l+1} + \alpha_{m+s}) \quad (5)$$

where  $s = \delta(k + 1)$

Assuming that a confidence level,  $1 - \delta$ , is given a priori, where  $\delta > 0$  is a small constant (*e.g.* 5%). It means that for a  $\alpha = 0.05$  and a confidence of 95%, the interval width is given by  $\alpha_{m+0.05(k+1)}$ , where  $k$  is the calibration set length.

## 2.2 PIs metrics

The literature offers a variety of methods for the evaluation of the performance of point prediction methods, *e.g.* mean square error (MSE), mean absolute error (MAE) or mean absolute percentage error (MAPE). However, there is no well-established error measure dedicated to PIs assessment. Typically, the evaluation is only made based on the PI coverage probability (PICP), that can be interpreted as the probability that target values will be covered by the interval bounds. It is defined as:

$$PICP = \frac{1}{N} \sum_{i=1}^N c(i) \quad (6)$$

where  $N$  is the number of samples and  $c(i) = 1$ , if  $\hat{y}(i) \in [L(i), U(i)]$ ,  $L(i)$  is the lower bound, and  $U(i)$  is the upper bound, otherwise  $c(i) = 0$ . Ideally, the PICP should be as close as possible to its nominal value  $(1 - \alpha)\%$ , the confidence level for which PIs have been constructed. However, without considering its length the PI evaluation sound more subjective than objective. Therefore, it is essential the computation of width-based indices. Typically, the intervals are normalized to the number of interval, PI normalized average width (PINAW).

$$PINAW = \frac{1}{N} \sum_{i=1}^N (U(i) - L(i)) \quad (7)$$

Typically, very narrow PIs with a low coverage probability are not very reliable. On the other hand, very wide PIs with a high coverage probability are not very useful to use practically. The combination of both PI aspects can be performed by the use of different criteria, like coverage-length-based criterion (CLC) [10].

$$CLC = NPINAW(1 + e^{-\eta(PICP - \mu)}) \quad (8)$$

where  $\mu$  and  $\eta$  are two controlling parameters. The CLC tries to compromise between informativeness and correctness of a PI [6]. PIs should be as narrow as possible from the informativeness perspective. However, the narrowness tends to result in a low coverage probability.

### 3 PIs evaluation

#### 3.1 min-max error

We propose to compute errors with reference to the lower bound (min error) and upper bound (max error). They result from the distance from the predicted value to the respective limit ( $min_{dist}$ ) or ( $max_{dist}$ ). If the predictions are within PI limits,  $min_{dist}$  is negative and  $max_{dist}$  is positive.

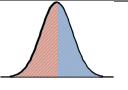
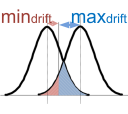
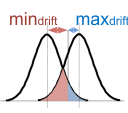
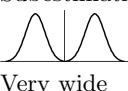
$$min_{dist} = \hat{y}_{min} - y \quad (9)$$

$$max_{dist} = \hat{y}_{max} - y \quad (10)$$

Errors values are represented in histograms. The bins associated to the prediction outside of the respective bound are identified as:

- Bins placed in  $x > 0$  in  $min_{dist}$  histogram
- Bins placed in  $x < 0$  in  $max_{dist}$  histogram

**Table 1.** Main properties of min-max distribution.

Interval	min-max distance	min-max error	min-max drift
 Very narrow	$min_{dist} \approx max_{dist}$	$min_{area} \approx max_{area}$	$\approx 0$
 Overestimation	$min_{dist} < max_{dist}$	$min_{area} < max_{area}$	$min_{drift} < max_{drift}$
 Subestimation	$min_{dist} > max_{dist}$	$min_{area} > max_{area}$	$min_{drift} > max_{drift}$
 Very wide	$min_{dist} \approx max_{dist} \approx \infty$	$\approx 0$	$\approx \infty$

The key points are: (1) the area associated to the error bins (min-max error), and (2) the drift from zero (min-max drift). PIs can range from point predictions (very narrow intervals,  $\hat{y}_{max} = \hat{y}_{min}$ ) to very wide intervals. The key properties are presented in Table 1, assuming an *i.i.d* distribution.

### 3.2 Mean absolute error of the interval

The calculation of MAE for a point forecast is relatively simple. It involves summing the magnitudes (absolute values) of the errors to obtain the total error and then dividing the total error by N. In the case of a PI, a single measure is not possible, since the prediction belongs to a range of values. We propose to compute the maximum MAE allowed for an interval (considering a hit,  $\hat{y} \in [\hat{y}_{min}, \hat{y}_{max}]$ ). To facilitate, the central value of the interval is taken as the forecast value:

$$\hat{y} = \frac{\hat{y}_{max} - \hat{y}_{min}}{2} \quad (11)$$

So, MAE is computed as:

$$MAE = \frac{1}{n} \sum_{i=1}^n \left| y - \frac{\hat{y}_{max} - \hat{y}_{min}}{2} \right| \quad (12)$$

The MAE limits are verified when the real value is  $y = \hat{y}_{max}$  or  $y = \hat{y}_{min}$ . In this case, MAE ranges between:

$$\frac{1}{n} \sum_{i=1}^n \left| \hat{y}_{min} - \frac{R}{2} \right| \leq MAE \leq \frac{1}{n} \sum_{i=1}^n \left| \hat{y}_{max} - \frac{R}{2} \right| \quad (13)$$

considering  $|\hat{y}_{max} - \frac{R}{2}| > |\hat{y}_{min} - \frac{R}{2}|$ , where  $R$  is the interval width. MAE range is the absolute value of the difference between limits.

## 4 Case Study and Experimental Setup

### 4.1 Data

The dataset includes historical data from 1 April to 31 November 2014, collected in the Customer Load Active System Services (CLASS) Project run by the UK Distribution Network Operator Electricity North West Limited <sup>3</sup>. The data consist of 30 MV substations. Each one is treated individually, being that 70% of data used for learning the global model, and the remaining 30% for prediction. All of the experiments were repeated 5 times.

### 4.2 Horizon Forecasting

NNs are one of the most popular options in the electric load forecasting [11]. The predictive model for the next day (hourly predictions) is a feed-forward neural network. The choice of the network topology and inputs was motivated by previous work [12–14]. It is constituted by:

*Inputs:*

- 24 values of the load curve  $[L(d-1)1, L(d-1)2, \dots, L(d-1)24]$  of day  $d-1$  (day before the forecasting day  $d$ ).
- Day of week, entered as two different variables, in the form of sines and cosines, by means of  $\sin[(2\pi d)/7]$  and  $\cos[(2\pi d)/7]$ , for each one of the days: Sunday( $d=0$ ), Monday( $d=1$ ), Tuesday( $d=2$ ), Wednesday( $d=3$ ), Thursday( $d=4$ ), Friday( $d=5$ ), Saturday( $d=6$ ).

*Output:* 24 values of the load curve  $[Ld1, \dots, Ld24]$

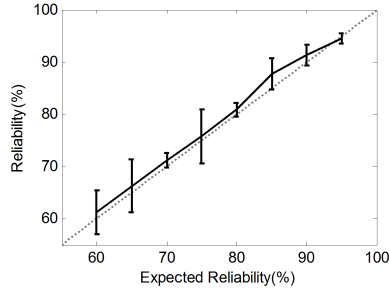
### 4.3 Calibration

The NNR-ICP model is calibrated through Equation 4. In Figure 1 the expected and the observed reliability values are plotted, considering all the substations. As shown, the calibration fit is linear with predictions falling near to the line of equality for the predicted and expected values. It can be concluded that NNR-ICP is well calibrated in the case of the database considered in this study.

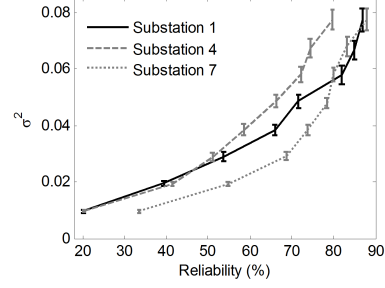
In opposition to the NNR-ICP model, the DPC method is not calibrated a priori. The jit added to the input variables follows  $x_i \approx N(0, \sigma_i^2)$ ,  $i = 1, \dots, 10$ . The calibration curve is shown in Figure 2. Results are presented individually for each substation, considering 5 independent trials. Calibration curves for three of the substations are depicted, evidencing the inter-substation variability.

In figure 3 we compare the NNR-ICP and DPC prediction regions at different significant levels. As expected, as the confidence level increases, the corresponding interval width is enlarged. Prediction intervals are different. NNR-ICP produces symmetric intervals, while DPC intervals are asymmetric.

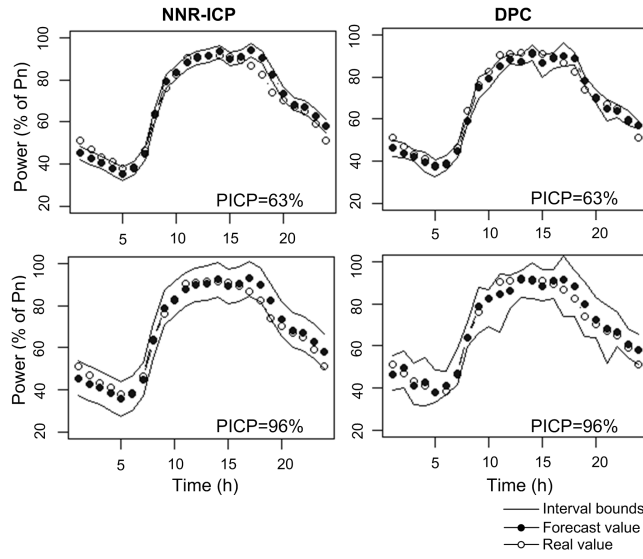
<sup>3</sup> <https://www.enwclass.nortechonline.net/data#substation-group/31>



**Fig. 1.** NNR-ICP calibration considering all the substations.



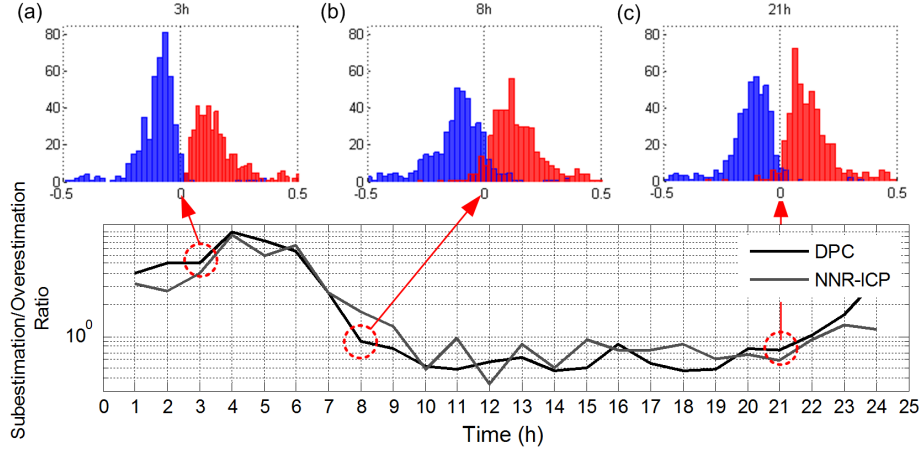
**Fig. 2.** DPC calibration for three different substations.



**Fig. 3.** PIs for the NNR-ICP and DPC methods at different confidence values.

#### 4.4 min-max evaluation

An unreliable PI can lead to the underestimation or overestimation of the real value. In load forecast, the trend to underestimate or overestimate depends of the location, season, or even time of day. The min-max error is computed as described in Section 3. Three different situations are depicted in Figure 4 (upper panel), representing the forecast errors along the day. In (a) the overlapped red/blue area is minimized due to the low number of predictions that fall out of the interval. Along the day, the error increases to maximum values (b), and in (c) turns to decreases. At lower panel, the underestimation/overestimation ratio for both algorithms at confidence level 82% is presented. It is visible, that during the periods  $[0h - 8h]$  and  $[21h - 24h]$  the load predicted values tend to

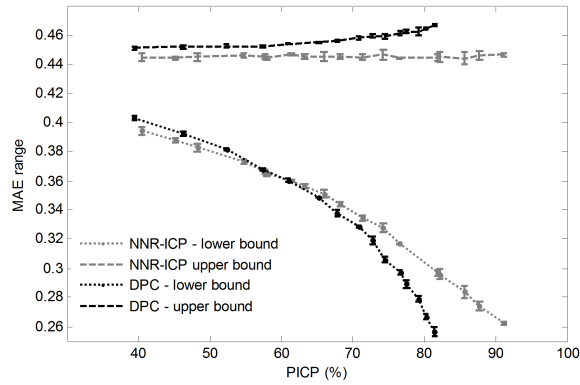


**Fig. 4.** Forecast errors (upper panel) and underestimation/overestimation errors along day (lower panel) for both algorithms.

overestimate the real value, while during the period  $[9h - 21h]$  the predicted values are underestimated.

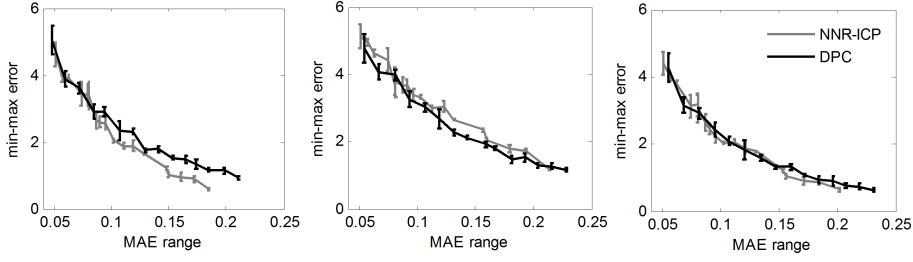
#### 4.5 PIs comparison

The MAE limits for the NNR-ICP and DPC algorithms at different confidence levels are presented in Figure 5. The interval width increases at faster or lower speeds depending of the method used to built it. DPC method often leads to PIs whose lengths are significantly larger than PIs constructed using the NNR-ICP method. The paid cost is a higher MAE range and less informative intervals. This effect is visible for PICP values above 50%.



**Fig. 5.** MAE limits for the NNR-ICP and DPC algorithms.





**Fig. 6.** Min-max error vs. MAE at three different periods: (1)  $[1 - 4h]$ , (2)  $[11 - 14h]$  and (3)  $[21 - 24h]$ .

Finally, we compare the min-max error *vs.* MAE for three different periods: (1)  $[1 - 4h]$ , (2)  $[11 - 14h]$  and (3)  $[21 - 24h]$ . A model is better as the min-max error is minimized. In (a) NNR-ICP presents a superior performance. In (b) the min-max error is minimized for the DPC method. This means, that PIs obtained with DPC are more robust during this period of day. In (c) both methods are comparable. Min-max error measured during the period  $[11 - 14h]$  is superior. This period is associated to higher forecast uncertainty. The selection of the DPC algorithm at this schedule is justified due to their wider and asymmetric interval bounds.

## 5 Conclusions

A new methodology for evaluating PIs is addressed. From the methodological aspect, we have adopted two innovative approaches. One is the exploration of prediction errors distribution (the min-max error), while the other is the quantification of the MAE values associated to the interval bounds. The comparison consists in analyzing the contrast of the maximum MAE allowed, and the minimization of the min-max error for these values. It aims to guide model selection for PIs with the shortest length and the lowest error dispersion. Higher errors are an indication of the presence of higher levels of uncertainty in forecasts. This information can guide the decision makers to avoid the selection of risky actions. In contrast, lower errors mean that decisions can be made more confidently with less chance of confronting an unexpected condition in the future.

Firstly, the calibration issue was addressed. Results indicate that NNR-ICP is well calibrated in the case of the database considered in this study. In the case of DPC (non calibrated method) the inter-substation variability is evident.

The min-max error depends of the time of day for both algorithms. Additionally, we can conclude that during periods of day associated with higher levels of uncertainty, DPC tends to have a better performance. This is justified due to their wider and asymmetric interval bounds, in comparison to the shorter and symmetric NNR-ICP limits. However, NNR-ICP tends to have a superior performance.

*Acknowledgments* . This work was supported by NORTE-07-0124-FEDER-000056 financed by ON.2—O Novo Norte, under the National Strategic Reference Framework, through the Development Fund, and by national funds, through FCT. And, by European Commission through MAESTRA (ICT-2013-612944).

## References

1. P. F. Christoffersen, “Evaluating Interval Forecasts,” *International Economic Review*, vol. 39, pp. 841–62, November 1998.
2. J. Armstrong and F. Collopy, “Error measures for generalizing about forecasting methods: Empirical comparisons,” *International Journal of Forecasting*, vol. 8, no. 1, pp. 69 – 80, 1992.
3. M. OConnor, W. Remus, and K. Griggs, “The asymmetry of judgemental confidence intervals in time series forecasting,” *International Journal of Forecasting*, vol. 17, no. 4, pp. 623 – 633, 2001.
4. P. Geurts and L. Wehenkel, “Closed-form dual perturb and combine for tree-based models,” in *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005* (L. D. Raedt and S. Wrobel, eds.), vol. 119 of *ACM International Conference Proceeding Series*, pp. 233–240, ACM, 2005.
5. H. Quan, D. Srinivasan, and A. Khosravi, “Short-term load and wind power forecasting using neural network-based prediction intervals,” *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, pp. 303–315, Feb 2014.
6. A. Khosravi, S. Nahavandi, D. C. Creighton, and A. F. Atiya, “Comprehensive review of neural network-based prediction intervals and new advances,” pp. 1341–1356, 2011.
7. C. Saunders, A. Gammerman, and V. Vovk, “Transduction with confidence and credibility,” in *In Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 722–726, 1999.
8. K. Proedrou, I. Nourtdinov, V. Vovk, and A. Gammerman, “Transductive confidence machines for pattern recognition,” in *in ECML 2002*, pp. 381–390, Springer, 2001.
9. H. Papadopoulos and H. Haralambous, “Reliable prediction intervals with regression neural networks,” *Neural Networks*, vol. 24, no. 8, pp. 842 – 851, 2011. Artificial Neural Networks: Selected Papers from {ICANN} 2010.
10. H. Quan, D. Srinivasan, and A. Khosravi, “Uncertainty handling using neural network-based prediction intervals for electrical load forecasting,” *Energy*, vol. 73, no. 0, pp. 916 – 925, 2014.
11. H. Hippert, C. Pedreira, and R. Souza, “Neural networks for short-term load forecasting: a review and evaluation,” *Power Systems, IEEE Transactions on*, vol. 16, pp. 44–55, Feb 2001.
12. J. Gama and P. Rodrigues, “Stream-based electricity load forecast,” in *Knowledge Discovery in Databases: PKDD 2007*, vol. 4702 of *Lecture Notes in Computer Science*, pp. 446–453, Springer Berlin Heidelberg, 2007.
13. P. P. Rodrigues and J. a. Gama, “A system for analysis and prediction of electricity-load streams,” *Intell. Data Anal.*, vol. 13, pp. 477–496, Aug. 2009.
14. L. Hernandez, C. Baladrn, J. M. Aguiar, B. Carro, A. Snchez-Esguevillas, and J. Lloret, “Artificial neural networks for short-term load forecasting in microgrids environment,” *Energy*, vol. 75, no. 0, pp. 252 – 264, 2014.