Time-Series Prediction: Application to the Short-Term Electric Energy Demand

Alicia Troncoso Lora[1],
Jesús Manuel Riquelme Santos[2], José Cristóbal Riquelme[1],
Antonio Gómez Expósito[2], and José Luís Martínez Ramos[2]

[1] Department of Languages and Systems, University of Sevilla, Spain
[2] Department of Electrical Engineering, University of Sevilla, Spain
{ali,riquelme}@lsi.us.es, {jsantos,age,camel}@us.es

Abstract. This paper describes a time-series prediction method based on the kNN technique. The proposed methodology is applied to the 24-hour load forecasting problem. Also, based on recorded data, an alternative model is developed by means of a conventional dynamic regression technique, where the parameters are estimated by solving a least squares problem. Finally, results obtained from the application of both techniques to the Spanish transmission system are compared in terms of maximum, average and minimum forecasting errors.

1 Introduction

Accurate prediction of the future electricity demand constitutes a vital task for the economic and secure operation of power systems. In the short, medium and long terms, generation scheduling comprises a set of interrelated optimization problems strongly relying on the accuracy of the forecasted load. The same happens in new competitive electricity markets, where the hourly bidding strategy of each partner is significantly conditioned by the expected demand. Consequently, it is crucial for the electric industry to develop appropriate forecasting techniques.

Existing forecasting methods for the estimation of electric demand can be broadly classified into two sets, namely: classical statistical methods [1, 2] and automated learning techniques. Statistical methods aim at estimating the future load from past values. The relationship between the load and other factors (temperature, etc) are used to determine the underlying model of the load time series. The main advantage of these methods lies in its simplicity. However, owing to the nonlinear nature of such a relationship, it is difficult to obtain accurate enough and realistic models for classical methods.

In the last few years, machine learning paradigms such as Artificial Neural Networks (ANN) [3–5] have been applied to one day-ahead load forecasting. The ANNs are trained to learn the relationships between the input variables (past demand, temperature, etc.) and historical load patterns. The main disadvantage of the ANN is the required learning time.

R. Conejo et al. (Eds.): CAEPIA-TTIA 2003, LNAI 3040, pp. 577–586, 2004.

Recently, classification techniques based on the k nearest neighbors (kNN), have been successfully applied in new environments outside traditional pattern recognition such as medical diagnosis tools, game theory expert systems or time series forecasting. Several papers have been published on the application of those techniques to forecast the electricity market price [6, 7], providing competitive results, but its application to the next-day load forecasting problem has not been yet tested.

This paper describes a time-series prediction method based on the kNN technique. The proposed methodology is applied to the 24-hour load forecasting problem, and the results obtained from its application to the Spanish case are analyzed. Then, based on available data, an alternative model is built up by a conventional dynamic regression technique. Finally, results obtained by both techniques are compared in terms of maximum, average and minimum forecasting errors.

2 Problem Statement

The one day-ahead load forecasting problem aims at predicting the load for the twenty-four hours of the next day. To solve this problem two schemes can be considered:

1. Iterated Scheme: This scheme forecasts the load for the next hour and the value obtained is used for the prediction of subsequent hours. The process is repeated until the 24-hour load forecasting is obtained. The iterated prediction has the disadvantage that the errors get accumulated particularly during the last hours of the prediction horizon.
2. Direct Scheme: Under this scheme, the load for the entire 24-hour period is forecasted from past input data. The direct prediction does not take into account the relationships between the load for one hour and the load for the next hour.

Test results have shown in average the same accuracy for both schemes. Thus, the direct scheme has been adopted.

2.1 Proposed Approach

In this section, a kNN approach [8] for next day hourly load forecasting is described. kNN algorithms are techniques for pattern classification based on the similarity of the individuals of a population. The members of a population are surrounded of individuals which have similar properties. This simple idea is the learning rule of the kNN classifier. Thus, the nearest neighbors decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. Unlike most statistical methods, which elaborate a model from the information available in the data base, the kNN method considers the training set as the model itself. A kNN algorithm is characterized by issues such as number of neighbors, adopted distance, etc.

In the kNN method proposed in this paper, each individual is defined by the 24 demand values corresponding to a whole day. Thus, the kNN classifier tries to find the daily load curve which is "similar to" the load curve of previous days.

The basic algorithm for predicting the electric energy demand of a given day $d+1$ can be written as follows:

1. Calculate the distances between the load of day d, D_d, and that of preceding points $\{D_{d-1}, D_{d-2}, ...\}$. Let v_1,...,v_k be the k nearest days to the day d, sorted by descending distance.
2. The prediction is:

$$\widehat{D}_{d+1} = \frac{1}{\alpha_1 + ... + \alpha_k} \sum_{l=1}^{l=k} \alpha_l \cdot D_{v_l+1} \tag{1}$$

where

$$\alpha_j = \frac{d_w(D_d, D_{v_k}) - d_w(D_d, D_{v_j})}{d_w(D_d, D_{v_k}) - d_w(D_d, D_{v_1})} \tag{2}$$

Notice that $0 \leq \alpha_j \leq 1$, i.e., the weight is null for the most distant day and is equal to one for the nearest day.

The two former steps are repeated for every day of the forecasting horizon, taking into account that true values replace forecasted ones for past days.

Therefore, the prediction aims at estimating the load for tomorrow from a linear combination of the loads corresponding to those days that follow the k nearest neighbors of today. The weights adopted for this averaging process reflect the relative similarity of the respective neighbor with the present day.

If the k nearest neighbors of D_d are $[D_{v_1}, ..., D_{v_k}]$, the set of points $[D_{v_1+1}, ..., D_{v_k+1}]$ will usually contain the k nearest neighbors of D_{d+1}, at least for noise-free time series.

Figure 1 geometrically illustrates the idea behind the kNN classifier when the considered number of neighbors is equal to one. Today's hourly load and tomorrow's unknown load are represented by circumferences. The four black

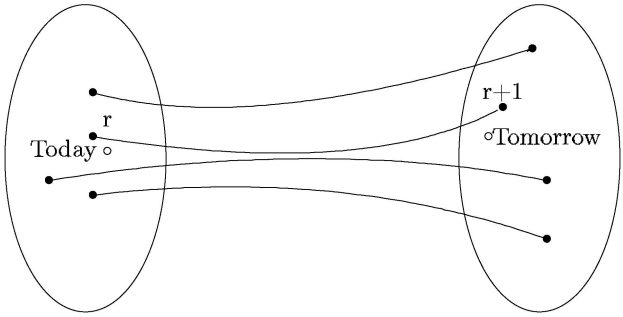


Fig. 1. kNN Learning Rule.

points are neighbors of today's load, point r being the nearest neighbor. Then, a possible estimation for tomorrow's load is the load of the day $r + 1$.

The classical kNN algorithm would resort to the actual nearest neighbors of day $d + 1$ for the prediction of D_{d+1}, but this is not possible because D_{d+1} is unknown in advance. The kNN proposed in this paper is a modification by which the k nearest neighbors of day d, whose demand is available, are adopted instead.

Some important parameters defining the kNN classifier are:

1. Choice of a metric: A time series Y can be considered as a point in a n-dimensional space. Given a sequence query, q, a sequence of Y, z, of the same length as q is searched, such that the distance between both sequences is minimum. The choice of the metric to measure the similarity between two time series depends mainly on the specific features of the considered series. The most common metric is the square of the Euclidean distance, although other metrics can be used [9, 10].
2. Number of neighbors: Accuracy of the forecasted load can be influenced by this parameter. In practice, the optimal value of k is usually small for noise-free time series, since a small number of different k values needs to be considered to find the optimal value. In this paper, k is determined by minimizing the relative, absolute and square mean errors for the training set.

2.2 Numerical Results

The kNN described in the previous section has been applied in several experiments to obtain the forecast of Spanish electric energy demand. The period January 2000-May 2001 has been used to determine the optimal number of neighbors and the best metric to measure the similarity between two curves.

The period June-November 2001 has been subsequently chosen as a test set to check the forecasting errors and to validate the proposed method.

Figures 2a) and 2b) show the influence of the number of neighbors on the relative, absolute and square mean errors for the considered training set, when the metric adopted to evaluate the similarity between a previous day and the historical data, is the Euclidean and Manhattan distance, respectively.

From these figures, the following can be stated for the particular time series under study:

1. The optimal number of neighbors is six using the Euclidean distance and thirteen using the Manhattan distance. Consequently, this number depends of the chosen distance.
2. This parameter is independent of the objective error function to minimize.

Test results have shown the same average error for the training set when both distances are considered. Thus, the Manhattan distance is the only one considered in the sequel.

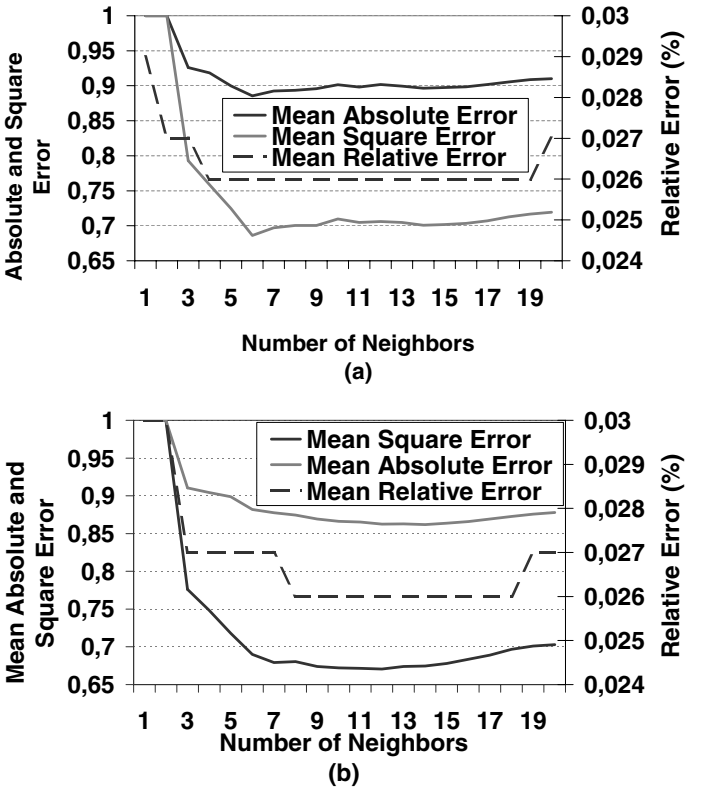


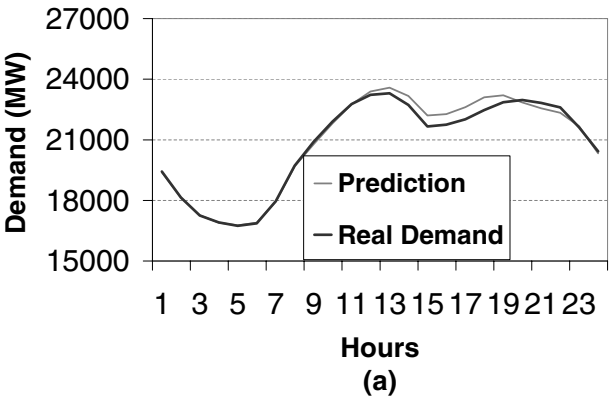
Fig. 2. Optimal No. of Neighbors with **a)** Euclidean distance, **b)** Manhattan Distance.

Figure 3a) shows the hourly average of both the actual and forecasted load for the working days from June 2001 to November 2001. The mean forecasting error, 2.3%, is of the same order as that of alternative techniques considered earlier, like ANNs [3–5].

Figure 3b) presents the hourly average absolute error of the forecasted load for the Autumn and Summer seasons. Note that the forecasting errors are larger during valley hours, i.e., hours with lower demand (1am-5am). However, it is more important to obtain a good prediction during peak hours (10am-2pm and 6pm-10pm) since the electric energy is more expensive at those periods.

Figure 4 presents the forecasted load on Monday August 6 and Thursday July 17, which are the days leading to the largest and smallest average relative errors respectively, along with the actual load for those days. It can be observed that the worst day corresponds with the first Monday of August, when most Spaniards start their Summer holidays.

Figure 5 presents the forecasted load for the two weeks leading to the largest and smallest average errors, along with the actual load for those weeks. Those weeks correspond with Tuesday September 11-Monday September 17 and Monday October 22- Friday October 26, respectively. Note that the week with higher



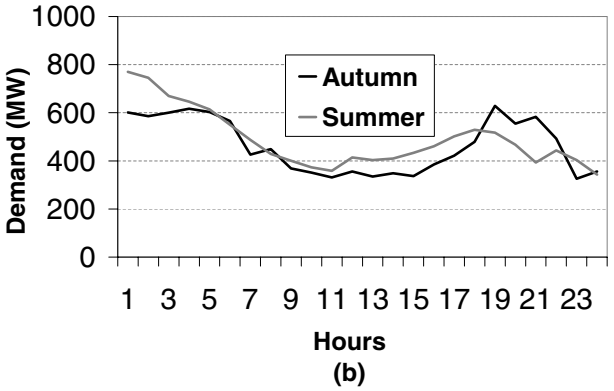


Fig. 3. a) Hourly average of forecasted and actual demand; **b)** Hourly average absolute error of the forecasted values.

Table 1. Daily Errors corresponding to the best and worst weekly Predictions.

	Daily errors (%)					Weekly error (%)
October 22 - October 26	3.13	1.25	1.11	0.08	0.07	1.4
September 11 - September 17	5.88	6.73	1.18	1.13	4.55	3.9

prediction errors is the one when the terrorist assault to the New York twin towers took place. Errors corresponding to September 12 arise as a consequence of the atypical behavior of the previous day.

The average errors for the five working days of the best and worst weeks appear in Table 1. The weekly mean errors are 1.4% and 4%, respectively.

3 Dynamic Regression

In this section a Dynamic Regression (DR) model is developed for the prediction of the hourly electricity demand. Under this approach, the demand at hour t,

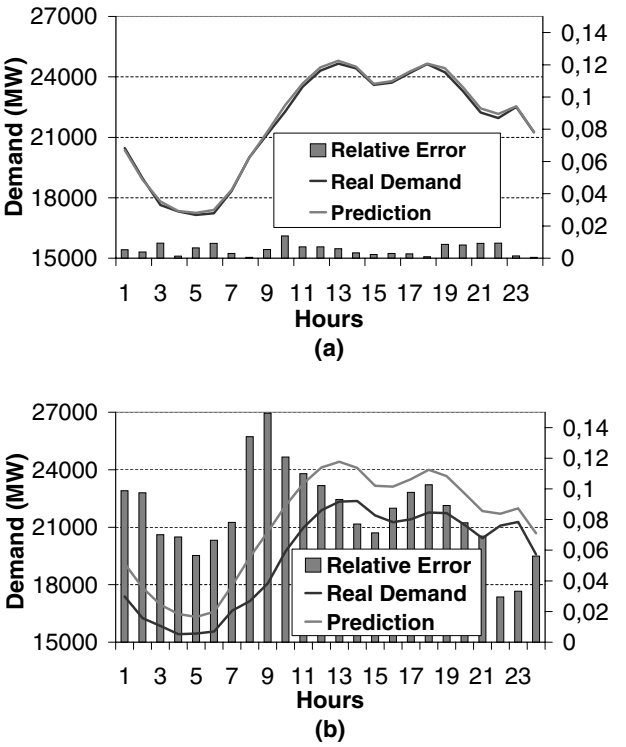


Fig. 4. a) Best Daily Prediction; **b)** Worst Daily Prediction.

D_t, is estimated from the demand at hours $t-1$, $t-2$,..., etc. First of all, a correlation study is performed on D_t, D_{t-1},... in order to determine which of the past demand values are most influential on the present demand.

Figure 6a) presents the average correlation coefficient between the present demand and past demand values for the period January 2000-May 2001. Notice that this coefficient presents a periodicity corresponding to a day. The main conclusion is that the highest correlation with the demand at a given hour takes place at the same hour of previous days. Furthermore, such a correlation decreases as the number of past hours increases.

In view of the conclusions obtained from the correlation study, the following model is proposed:

$$\widehat{D}_t = a_0 D_{t-1} + a_1 D_{t-24} + a_2 D_{t-48} + a_3 D_{t-72} + a_4 D_{t-96} + a_5 D_{t-120} \quad (3)$$

Experimental results suggest that it is not worth including extra terms like D_{t-23}, D_{t-25} in the model, as the average forecasting error for the tested period remains unaffected.

The model parameters a_i are obtained from the solution of the following least squares problem:

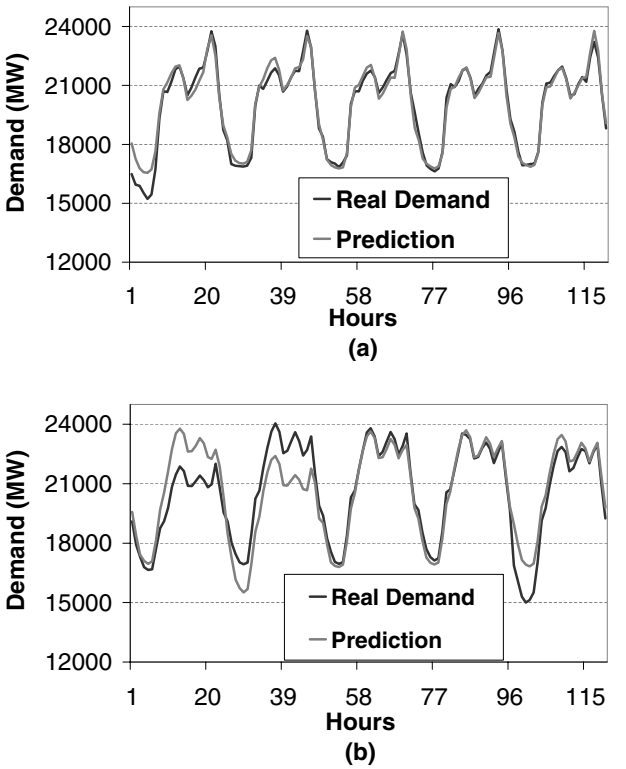


Fig. 5. a) Best weekly prediction; **b)** Worst weekly prediction.

$$\sum_t (D_t - \widehat{D}_t)^2 \tag{4}$$

where $\widehat{D}_t$ is defined by (3).

Those parameters can be computed only once, from the training set, or they can be updated daily.

3.1 Results

The regression model described above has been also applied to the load forecasting problem in Spain. Working days between January 2000 and May 2001 have been employed to determine the model parameters, yielding $a_0 = 0.31$, $a_1 = 0.37$, $a_2 = 0.08$, $a_3 = 0.01$, $a_4 = -0.03$ y $a_5 = 0.25$. The small values of a_2, a_3 and a_4 suggest that, when predicting D_t, the most influential values are D_{t-1}, D_{t-24} and D_{t-120}, i.e., the previous hour, the same hour of the previous day and the same hour and same day of the previous week respectively.

Figure 6b) shows the evolution of the model parameters when they are updated every day for the period June-November 2001. It can be observed that

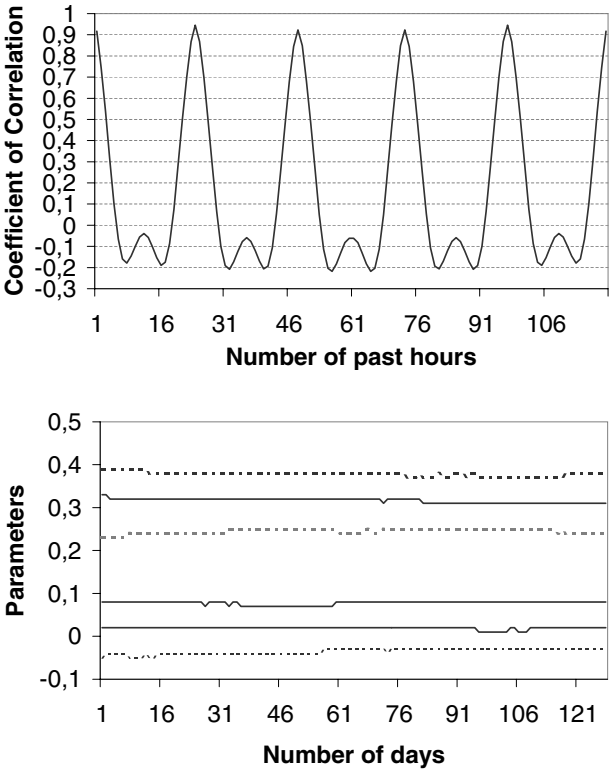


Fig. 6. a) Correlation Coefficients; b) Evolution of parameters a_i.

Table 2. Comparison of both methods.

	S.D.	Relative Error Mean (%)	Absolute Error Mean(MW)	Max. Error Daily (%)	Max. Error Hourly (%)
kNN	0.015	2.3	471	8.5	16.6
DR	0.019	2.82	572	9.2	19.3

these coefficients remain essentially constant, which means that no new demand patterns are added to the data base each time a day is included.

Finally, table 2 presents the standard deviation, the average absolute and relative forecasting errors, and the maximum daily and hourly errors obtained from the application of both the kNN method and the DR model to the period considered.

4 Conclusions

In this paper, a method based on the k Nearest Neighbors is proposed for the prediction of time series. The method is applied to the Spanish short-term electric

load forecasting problem and the resulting errors for a six-month period are analyzed. Then a dynamic regression model whose parameters are obtained by solving a least squares problem is developed for the same application. Comparing the results provided by both approaches leads to the conclusion that the kNN classifier is more accurate for the load forecasting problem than the conventional regression method.

Acknowledgments

Thanks are due to the Spanish CICYT (TIC2001-1143-C03-02, DPI2001-2612) and Junta de Andalucía (ACC-1021-TIC-2002) for sponsoring this project.

References

1. A. D. Papalexopoulos and T. C. Hesterberg: A Regression-Based Approach to Short-Term System Load Forecasting. IEEE Trans. on Power System, Vol. 5, pp. 1535-1547. 1990.
2. F. J. Nogales, J. Contreras, A. J. Conejo and R. Spínola: Forecasting Next-Day Electricity Prices by Time Series Models. IEEE Trans. on Power System, Vol. 17, pp. 342-348. 2002.
3. A. S. Alfuhaid and M. A. El-Sayed: Cascaded Artificial Neural Network for Short-Term Load Forecasting. IEEE Trans. on Power System, Vol. 12, pp. 1524-1529. 1997.
4. J. Riquelme, J.L. Martínez, A. Gómez and D. Cros Goma: Load Pattern Recognition and Load Forecasting by Artificial Neural Networks. International Journal of Power and Energy Systems, Vol. 22, pp. 74-79. 2002.
5. R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, V. Orsolini Cencellli: A Neural Network Based Technique for Short-Term Forecasting of Anomalous Load Periods. IEEE Transaction on Power Systems, Vol. 11, pp. 1749-1756. 1996.
6. A. Troncoso Lora, J. C. Riquelme Santos, J. M. Riquelme Santos, J. L. Martínez Ramos, A. Gómez Expósito: Electricity Market Price Forecasting: Neural Networks versus Weighted-Distance k Nearest Neighbours. DEXA Database Expert Systems and Applications, Aix Provence, 2002.
7. A. Troncoso Lora, J. M. Riquelme Santos, J. C. Riquelme Santos, A. Gómez Expósito, J. L. Martínez Ramos: Forecasting Next-Day Electricity Prices based on k Weighted Nearest Neighbours and Dynamic Regression. IDEAL Intelligent Data Engineering Autamitized Learning, Manchester, 2001.
8. B.V. Dasarathy : Nearest neighbour (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press, 1991.
9. R. D. Short, K. Fukunaga: The Optimal Distance Measure for Nearest Neighbour Classification. IEEE Transaction on Information Theory, 1981.
10. K. Fukunaga, T. E. Flick: An Optimal Global Nearest Neighbour Metric. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1984.