An Hybrid Approach to Prediction of Electric Load with MARS and Kohonen Maps

Francisco Ortega[1], María Teresa Rodríguez[2], César Menéndez[2], Nieves Roqueñi[1], Vicente Rodríguez[1], Valeriano Álvarez[1], Gemma Martínez[1], Joaquín Villanueva[1] and José Manuel Mesa[1]

[1] Universidad de Oviedo, Project Engineering Area, Independencia,13,
33004 Oviedo, Spain
{mailto:fran@api.uniovi.es, nievesr@api.uniovi.es, montequi@api.uniovi.es, valer@api.uniovi.es, gemma@api.uniovi.es, balsera@api.uniovi.es,
mesa@api.uniovi.es
http://www.api.uniovi.es
[2] Applied Mathematics Area, Independencia,13,
33004 Oviedo, Spain
mayte@api.uniovi.es
cesarm@scig.uniovi.es

Abstract. This paper presents the analysis done by the API-Uniovi group to the problem of load forecasting in Slovakia in January 1999. Due to the available data, the predictor is designed after a decomposition of the different variables of influence. In general, patterns of demand behavior are modeled with Unsupervised Kohonen Neural Networks and forecasting is done with Multidimensional Adaptive Regression Spines. Temperature is modeled with piecewise linear interpolation and forecasting is compensated with the influence of holidays on the days after and before. Results were tested with real data from Dec 1998 with a very promising level of success.

1. Description of the problem: API-Uniovi Impression

In the framework of the EUNITE project, this competition presents the practical problem of the electric load forecasting with a very limited set of data from the temporal and spatial point of view.

The availability of two complete years with temperature and half-hour loads, makes difficult the use of time series that would need at least seven complete years (from our point of view), due to the rotation of days. The knowledge of the temperatures of two previous years has contributed to evaluate the effort to do in temperature prediction.

Furthermore the criteria of maximum daily load instead of mean or accumulated daily load introduces a new limitation for the model as its variability is bigger and prediction has more risk.

With these conditionings, applying previous experience of API-Uniovi group in AI modeling and forecasting and considering also the unfortunately limited effort that is possible to dedicate to this research, we have decided to face the problem from an hybrid point of view, developing a composed model integrating unsupervised neural network, Case Based Reasoning and multivariate Statistical Techniques, which present the best combination of modeling effort and result. It will also be a way to make a blind validation of a forecasting method here developed as a modification of MARS (APIMARS).

2. Global Approach

From the literature and previous works developed by the team for Hidrocantábrico (Spanish electric company) it is clear that there are some factors influencing the consumption of energy, although their specific influence is unknown. The factors considered here are:

Table 1. Main variables affecting electricity demand

Variable	Type	Correlation
Energy consumption in previous periods	Linear	Positive
Temperature outside	Random/Cyclic	Negative
Hour.	Cyclic	Sinoid
Day of the week	Cyclic	Sinoid
Month	Cyclic	Sinoidal
Existence of holidays	Linear	Negative
Economical development of region	Potential	Positive

Those factors will be studied in the following paragraphs and the way they will be introduced in the forecasting will be explained.

2.1. Prediction of temperature.

Existing data shows extreme differences of energy consumption according to temperature. Then it is important to introduce in the forecasting model this effect, determining with the maximum accuracy the temperature of January 1999. Unfortunately temperature is one of the most difficult things to predict as it depends of hundreds of factors necessarily unavailable in this competition and mostly in the real world. Even with updated information, the most sophisticated techniques, equipment and meteorological data, meteorologist cannot produce accurate temperature predictions for a month without high degree of uncertainty.

Certainly there are some repeatable aspects as the annual cycle and differences due to geographical latitude and longitude or over-sea level, but that is not enough to determine temperatures for a given day. We are not considering here a further problem as it is the introduction of temperature in the models supposed known, as it will be discussed lately.

The representation of hourly temperatures versus time (Fig. 1) shows that repeatability of the term behaviour but, if we look in detail, that effect is not applicable to precise prediction of days. The approach of the temperature evolution with a sinoidal curve shows maximum y minimum values acting as an evolving curve.

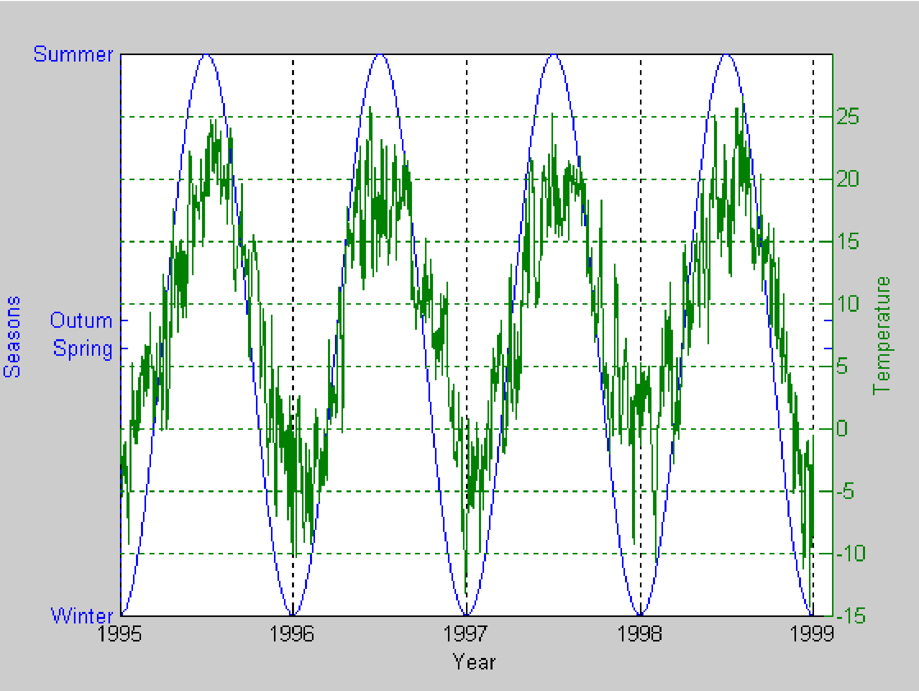

Fig. 1. Evolution of temperature during the four previous year. The blue line is a sinoidal approximation. It can be observed as temperatures in winter 1998 are higher than in previous winters and in the existing data of winter'98.

The observation of January 1998, in theory the most accurate idea for 1999 temperatures, shows an abnormal behaviour, very different of previous values of 1995, 1996 and 1997. 1998 seems to be an especially hot year in the region under study and it will necessarily have influence on prediction in two levels:

1. Temperature prediction could be excessively high for real 1999 temperatures.
2. January 1998 will probably have an abnormally low consumption of temperature producing an alteration in the very short time series.

From this observation and the impossibility to determine a more accurate temperature forecast for 1999, the API-UNIOVI decision was to:

- Avoid the use of serial analysis based in January 1998.
- Introduce the temperature of December 1998 as the most promising estimator for the first day of January 1999.
- Consider the averaged value of temperatures of January in previous years as the best estimation for Jan 15th 1999 as the importance of the information from Dec'98 is minimal after so many days. A linear evolution was established from Jan1 to Jan 15 as the prediction of temperature for the first days.
- Consider the averaged value of temperatures for February as the best estimation for Feb 14th '99. Then an interpolation between the mean of January and mean of February is used to predict the temperatures of the rest of January'99.

As result of these decisions, the temperatures to introduce in the model are represented by the following curve (Fig. 2):

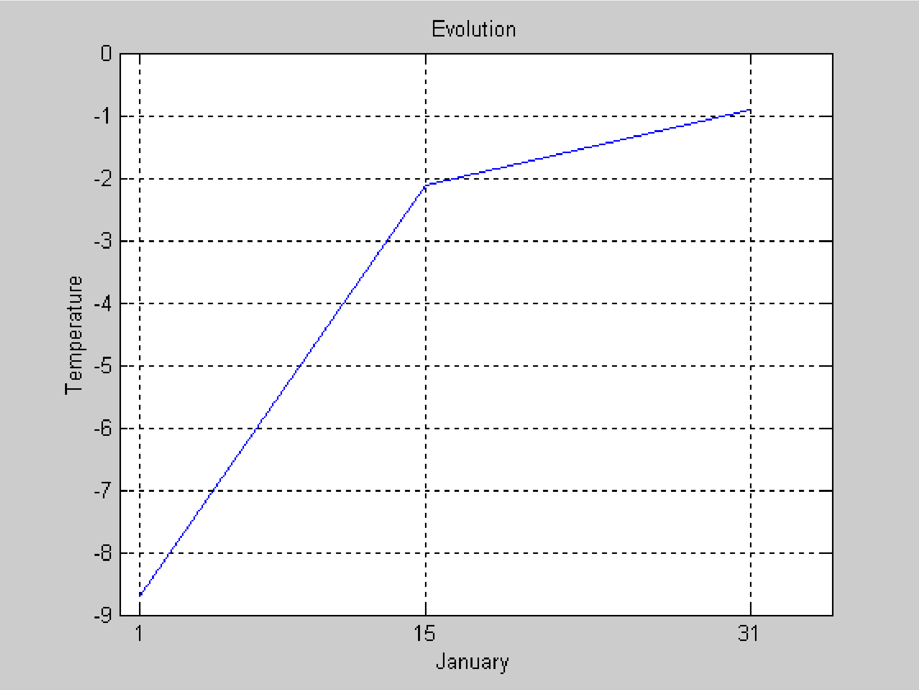

Fig. 2. Estimation of temperature for January 1999, based in averages of January and February as well as in December temperature. It is composed of two linear interpolations.

In any case as the temperature is considered unpredictable accurately, the predictor must be based in the lowest possible level in the temperatures.

Prediction was tested with the known data with very promising results as shown in figure 3.

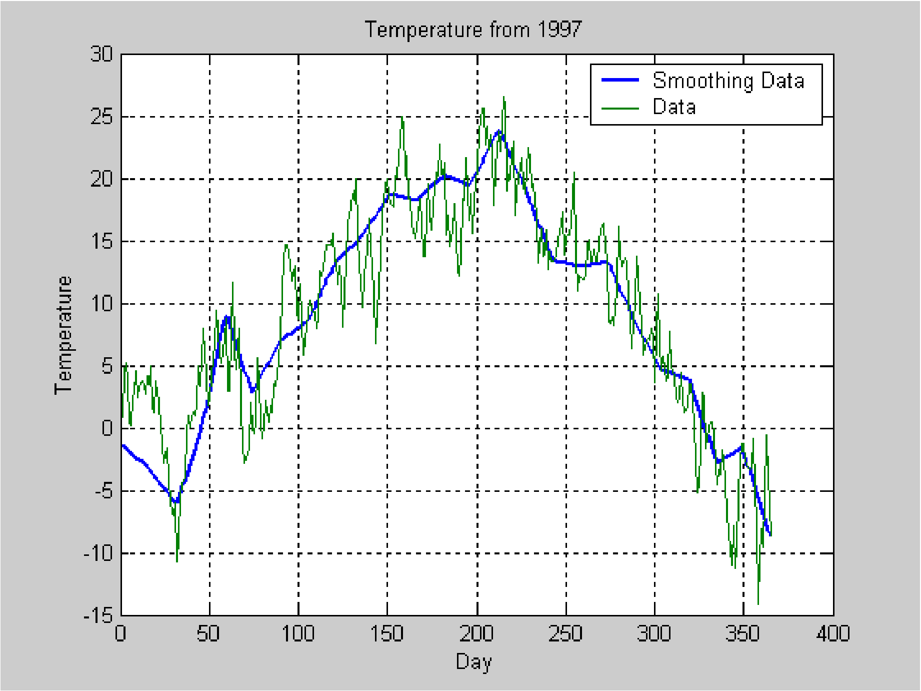

Fig. 3. Application of the temperature prediction interpolation to 1997 temperature. It can be observed as predictions follows the main variations of real data .

2.2. Effect of economical development

As a country develops the industrialisation creates new sources of consumption and comfort levels require new charges for individuals as much at work as at home. Also the development introduces new environmental rules and traditional energy sources are replaced by electricity. That growing is produced in a continuous way and it is accumulative, although influenced by economical cycles. It is then expected that every year the demand of electric energy grows continuously and every year (globally considered) must be higher than the year before.

Local effects as the previously mentioned abnormal temperatures of January 1998 could hide that effect. The study of accumulated data between 1997 and 1998 does not reflect the growing of the economy. We consider that it is due to a probable recession in the industry during that year and perhaps 1998. As there are not predicted indicators for 1999, it is considered that the influence of the economy will be inherent in the data used for training from 1997 and 1998, so no special action is designed for it.

2.3. Effect of the hour

In the data it is possible to observe different loads with time. In general midday is the moment of highest consumption. An analysis of the weekly patterns found in the last paragraph should show bigger variability in those patterns produced with maximum values than those with mean values, but differences are smaller than expected. Standard deviation of every pattern in every case (mean and max values) are quasi identical.

Moreover first studies of variable importance done with Principal Component Analysis produce small values for hour and a reduced range of variability.

The selection in this case is to avoid the variable hour and to produce the forecasting directly on the maximum consumption variable. We consider than the information lost is compensated with the reduction of complexity of the solution's space. That simplification is essential as the number of available data is small and there is a certain risk to be affected by a problem of over fitting and/or "empty space".

2.4. Effect of the day of the week

The relation between the energy consumption and the day of the week is evident. There is a cyclic effect with a period of seven days, very clear in any case independently of the month. Sunday is the day with the smallest load and Saturday is very low loaded too. Monday is not completely loaded, especially due to the inactivity of the night hours. Friday is nearly higher and it is very difficult to distinguish attending on

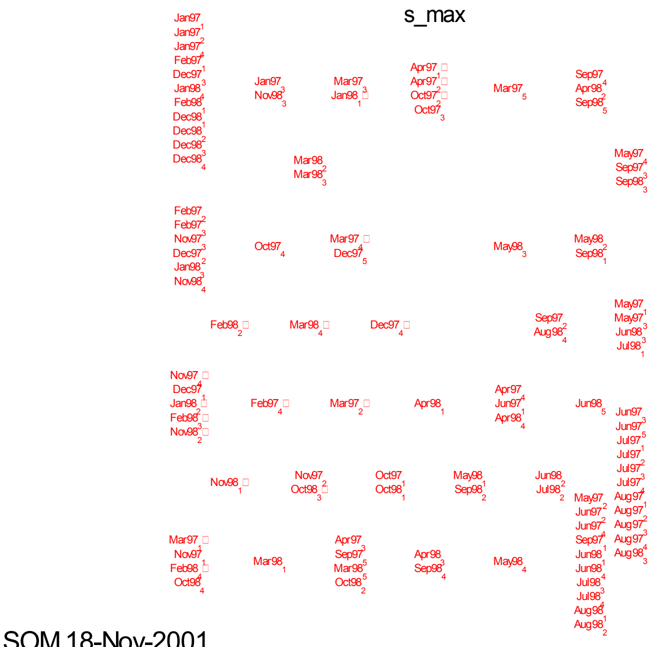

Fig. 4. Self Organising Map of the weekly consumption patterns. It can be observed as the weeks are grouped by terms, but also with some dispersion due to temperatures.

The first analysis was to detect if that cycle is repeated in every week with similar characteristics, i. e. if the shape of the consumption curve is similar every week across the year. In that case the prediction of a representative day of the week could be enough to determine the prediction, extending the main value according to the pattern of weekly consumption.

Examples show big differences in different weeks, not necessarily related to month, although the influence is important. So we decided to use some unsupervised neural networks trying to find relationships between patterns. After testing other clustering algorithms and several topologies, the final configuration is shown in next figure.

SOM was prepared with 14 inputs, including the seven maximum values of consumption and the seven temperatures of those days. Figure 5 shows the characteristics of the neuron representative and the ability of separation of the network. .

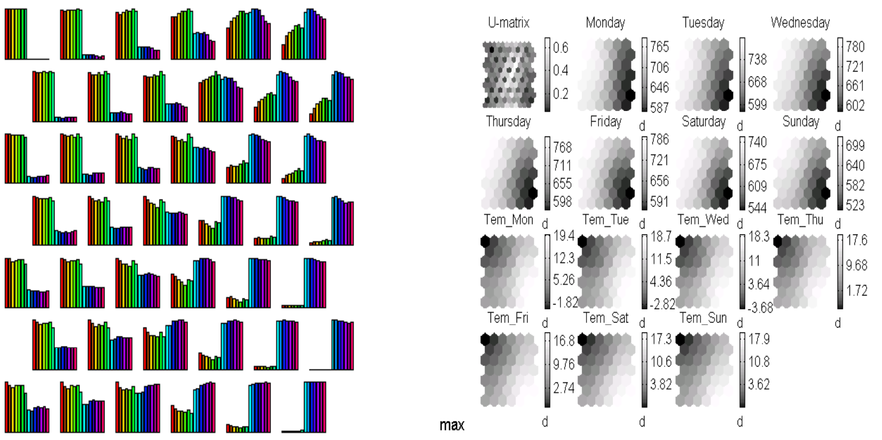

Fig. 5. Self Organising Map characteristics. Left figure shows the values of every variable in the class representative and right figure shows the plane components.

It is easy to observe how the January patterns, especially those from 1997, are concentrated in one neuron with other patterns of December of February. Other group hits the third neuron at the left. Observing the component it can be seen that differences are due to changes of temperature, causing different electricity demands. A deeper study shows that third neuron patterns are produced for hotter temperatures (near +3C) and neuron 1 represent temperatures of –3C approximately.

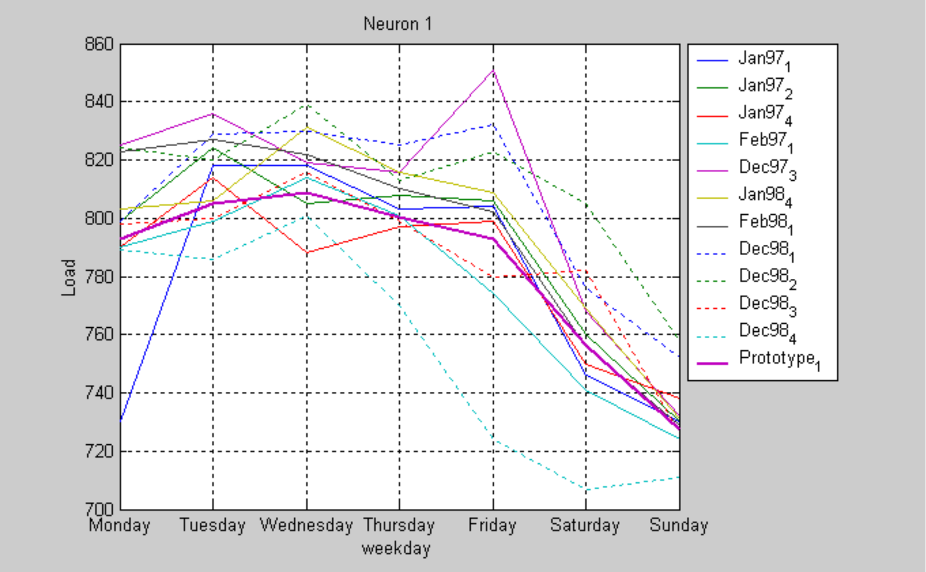

Fig. 6. Maximum consumption of electricity patterns for neuron 1. The thicker line represents the neuron value taken as default..

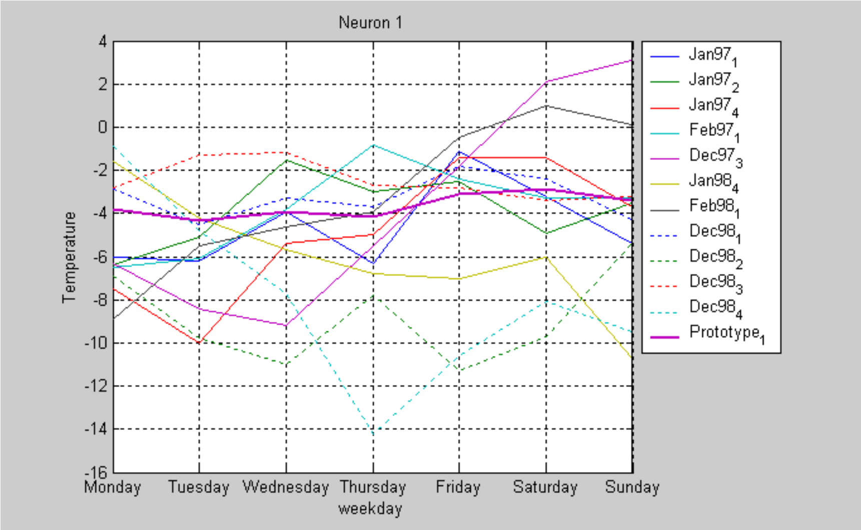

Fig. 7. Temperature patterns for neuron 1. The thicker line represents the neuron value taken as default..

It can be observed how weeks of January 1997 are grouped with some of 1998, but the rest of 1998 are in other neurones. That confirms our impression that January 1998 was too hot.

With those results we have decided to use that homogeneous pattern for the prediction of weekly values and to produce a predictor only for one day of the weeks, generally Mondays.

2.5. Effect of the month

Consumption is very dependent of the month. This is due to different causes, being one of them the temperature studied previously, but also the light hours, the vacation periods in summer, etc.

That cyclic effect is one of the most important for the prediction. After many testes and the observation that there are some weekly cycles as indicated previously, it was decided to prepare a different model for every day of the week, using for example only Mondays for training.

Although MLP NN were tested, finally a version of MultiAdaptive Regressive Splines were used, called APIMARS. J. Friedman developed MARS in 1988 and they have been used for multidimensional fitting in several applications successfully.

As in the rest of the fitting methods, the aim of MARS is to model the dependence of a response variable ***Y*** on one or more predictor variables $X = (x_1,...,x_n)$, where the data is given by $\{y^i, x^i\}_{i=1}^{N} \subset \Re^{N,n+1}$. The data is assumed be related by:

$$y^i = f(x^i) + \varepsilon \qquad x^i \in D \subset R^n \quad (\mathrm{i} = 1,..., \mathrm{N})$$

where ***f*** is an unknown continuous function, ε is a zero-mean error distribution that represents the dependence of the response variable ***Y*** on other unmeasured features, and ***D*** is the domain of the problem. The aim is to use the data to construct an estimate $\hat{f}(x)$ to the true function ***f(x)*** over the domain ***D***.

MARS is a spline regression model that uses a specific class of basis functions as predictors instead the original data:

$$\hat{f}(x_1, x_2,..., x_n) = \sum_{i=1}^{M} a_i B_i(x)$$

where:

a_i are the suitable chosen coefficients of the basis function B_i

M is the number of basis functions in the model.

The basis functions are such that:

$$\boldsymbol{B_i(x)} = \begin{cases} 1, & \boldsymbol{i} = 1 \\ \prod_{j=1}^{J_i} \left[\boldsymbol{s}_{ji} \cdot (\boldsymbol{x}_{v(j,i)} - \boldsymbol{t}_{ji})\right]_+ & \boldsymbol{i} = 2,3,... \end{cases}$$

where:

$\boldsymbol{J_i}$ is the degree of the interaction of basis $\boldsymbol{B_i}$

$\boldsymbol{s}_{ji} = \pm 1$ is the sign indicator

v(j,i) give the index of the predictor variable which is being split on.

$\boldsymbol{t_{ji}}$ give the position of the splits (known as knot points)

Although spline approaches are well known, they are not easily applicable when the number of predictors is high: for example 35 predictors with two levels each produces more than 34 billion of regions. MARS considers only a group of variables at the same time using a decision tree and introduces only as many nodes as needed. The MARS algorithm proceeds as follows. A forward stepwise search for basis functions takes place with the only basis function present initially: the constant one. At each step the split which minimises some "lack-of-fit" criterion from all the possible splits on each basis function is chosen. Splits are only permissible at the marginal predictor values. If the split was on basis $\boldsymbol{B_i}$ with predictor $\boldsymbol{x_*}$ at $\boldsymbol{t_*}$ this corresponds to the two new basis functions:

$$B_i\left[\pm(x_* - t_*)\right]_+$$

This continues until the model reaches some predetermined number of basis functions, which should be greater than twice the number of independent variables.

In the backwards stepwise, functions basis one at a time are removed until the lack-of-fit criterion is at a minimum. The basis that most improves the fit is removed at each step.

The MARS model is continuous in $\boldsymbol{D}$ and can be made to have continuous first derivatives by replacing the truncated linear basis function $\boldsymbol{B_i}$ by truncated cubic basis functions. Transform makes it possible to selectively blank out some regions of a variable in order to focus on the most promising zones. Then its excellent at finding interactions between variables and complex data structures.

Several modifications have been proposed on the basic algorithm. Denison introduces probabilities in the knot selection. API-Uniovi group has changed the forward/backward steps presenting different data in every step.

Final model was created with only three input variables: day of the week, day of the year and temperature and the output was the maximum daily load. Although more than 100 models were created, the last one has the following values:

Basis Function	Std. Dev	Variables
1	91.05	1
2	15.01	2
3	22.64	1,2
4	16.50	2,3

It is a piecewise cubic fit on 4 basis functions with possibility of two-dimensional interactions (1-2, 2-3).

Results with test data from Dec'98 shows percentages of 100% success for 15 and 20% of accuracy and over 89% for 9% accuracy, what we consider adequate. Obviously there are not results for Jan'99.

Training data uses complete 1997 set and partially 1998 (except January).

After the modelling it is available a model to predict every day of the week,just introducing its position in the year and its temperature. That will be the starting point for the patterns generated with the Kohonen map.

2.6. Effect of Holidays

The holidays cause a notorious effect of load descent. That is specially truth in December/January, where there are several consecutive holidays and even working days are limited by Christmas Holidays.

It is possible to observe differences between the decrease of load according to the day of the week and even its effect on other working consecutive days. January has two holidays in the first week, one of them immediately consecutive to the last day with data, Jan 1st. But in 1999 that holiday is in a Friday and it adds a bigger problem. A holiday so close to the weekend produces special effect on itself and on the next Saturday affected by the inactivity of the previous day.

There are not enough holidays on Friday to produce a certain model, so we have decided to predict values firstly without the effect of holiday and then to introduce the effect of holiday based in the relation between closer Sunday and the day of the web.

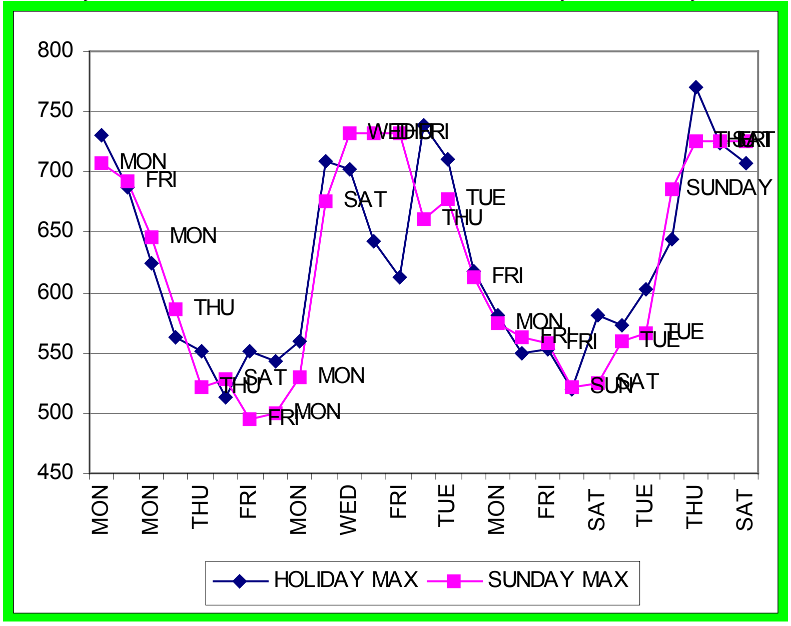

Fig. 8: Connection between load in a holiday and closer Sunday

There are not enough holidays on Wednesday to produce a certain model for the 6/1/1999, so we have decided to make a calculus of the distance normalised, with the five values of previous non-working Fridays. For doing this a SOM based on weeks with a holiday on Tuesday, Wednesday or Thursday was trained, and the majority pattern was adopted.

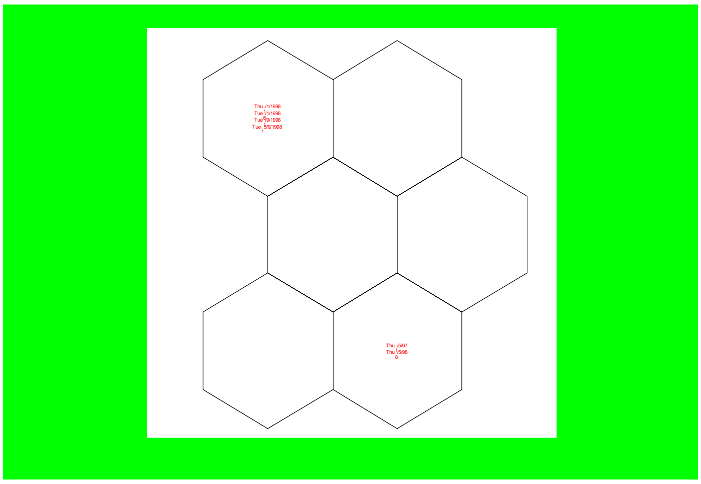

Fig. 9: Self Organising Map of the weekly consumption patterns for weeks with a holiday on Tuesday, Wednesday or Thursday

3. Configuration of Final Model

With the results produced before, the main forecasting system is composed of the following steps:

1. Determination of the temperature evolution on January 1999 as the interpolation of a piecewise linear interpolation, being the values of the points the temperature of December 31st, average of January of 95, 96, 97 and 98 and average of February 1995, 96, 97 and 98.
2. Determination of the basic load of January 1st, 1999 with the APIMARS technique.
3. Extension of the prediction to the first week (up to Jan 7th) with the patterns identified with the Kohonen network.
4. Determination of the maximum loads of the rest of January's Mondays with the APIMARS technique.
5. Extension of the prediction to the rest of the days of the week (up to Jan 31st) with the patterns identified with the Kohonen network.
6. Introduction of a compensation for holidays 1st and 6th according to the levels of previous Sundays.
7. Compensation of the days before and after a holiday with the use of a Kohonen network.

Final results were evaluated with the prediction of a APIMARS model developed specifically for every day of the week and the transitions between Sundays and Mondays were studied to determine if some problems of discontinuity could be present in those days.

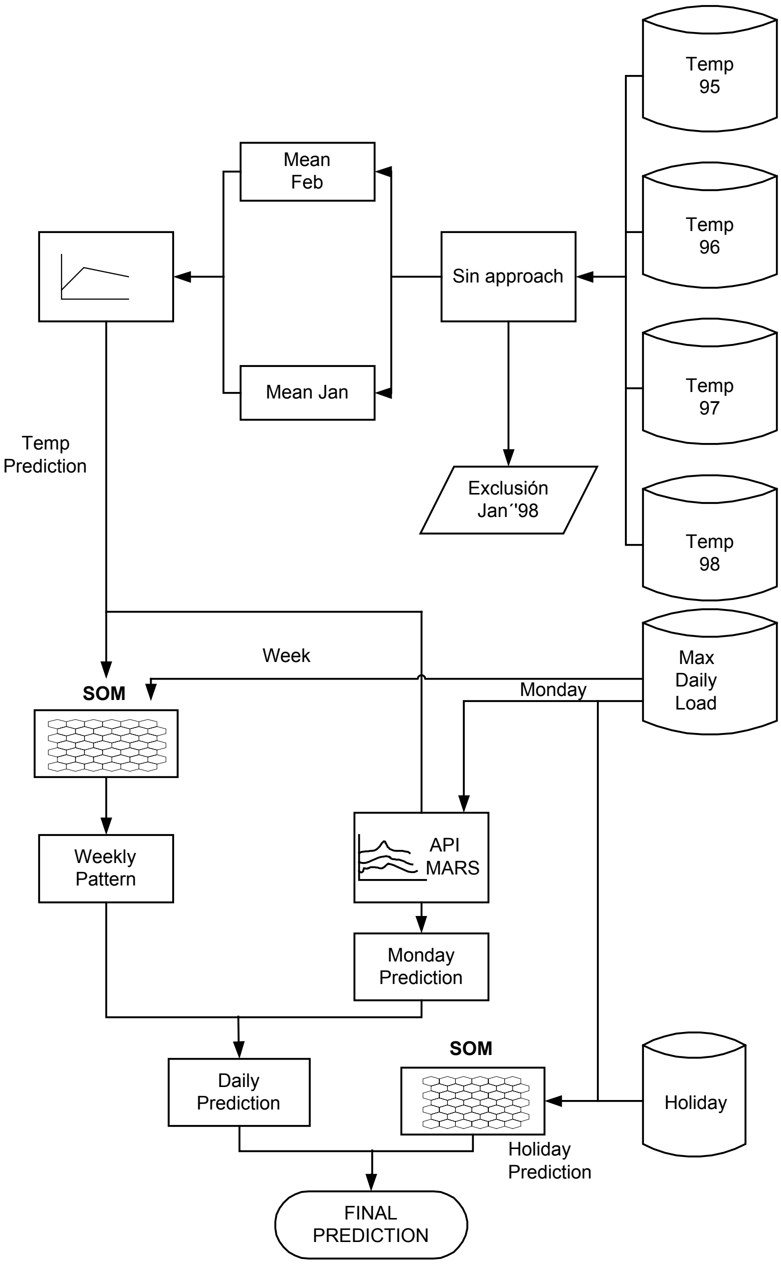

Fig. 10. Flowchart of the global process.

4. Data Processing

The data has been treated and pre-processed before its application in each model, although treatment is different due to the different techniques used. In general the main steps of pre-processing includes:

- Search of outliers and statistical characterisation of each variable: mean, maximum, minimum, standard deviation.
- Euclidean normalization for data introduced in Self Organizing Maps and distance-based methods.
- Normalization between 0,1-0,9 of data introduced in multiplayer perception and similar techniques, in order to ensure the avoid the "slope 0" zones.
- Division of data in two groups according to the 2/3-1/3 rule (training-validation). Any percentage indicated in this paper refers to validation only values, not to the complete or training sets. Although we use normally 10% of training data for cross validation in neural networks, they are not included here as the technique was not finally included.
- When a too small set of data is available, for example when only December was used for modeling, a "leave one out" strategy was used. Bootstrapping was not considered.
- Training and testing sets were selected randomly 5 times. Results of each trained test were compared. If differences were bigger than 10%, trainings were repeated. In other case final result is the average of the five partial tests.
- Selection of nodes, interactions and functions in the multi regressive Spines were done by experience and after successive tests in different conditions.
- Cyclic variables as day of the week were introduce after a zooidal transformation in order to ensure that Sunday and Monday were detected consecutive by the model. This was also applied to the month when several years were introduced.

5. Equipment

In order to develop the simulations the team has used Windows 2000 servers based on INTEL and UNIX machine from HP and DIGITAL. Software was own developed during the last 8 years by the team in different projects and applications, in C, Fortran, TCL and Java and it is grouped under the APIPAK (for statistical processing) and XDPM (for neural networks) names.

Other well-known software as Matlab, SNNS and Excel software was also used for testing in different stages of the project

6. Conclusion

Here it was presented an alternative approach to the problem of electric demand forecasting. Considering the limited number of existing patterns and the absence of other interesting parameters as humidity, thermal sensation or economic evolution, the approach was based in a combination of statistical analysis, clustering and neural networks.

That hybrid model makes use of several simplifications and it is additive: predictions are being improved considering the different effects on a previous first value. Different techniques are used for every step as it is considered that the variability and the absence of important data does not permit the correct behaviour of an unique model.

Although estimations seems to be realistic and work with very small errors for December'98, we think they will be very influenced by the real temperatures and the economic cycle.

7. References

1. Denison, D. G. T., Mallick, B. K. and Smith, A. F. M): Bayesian MARS. Technical report, Department of Mathematics, Imperial College, London. To appear in Statistics and Computing .(1997)
2. Friedman, J.H.: Multivariate Adaptive Regression Spines (MARS). Technical Report no.102, November 1988, Laboratory for Computational Statistics, Stanford

8. Appendix: Daily Values forecasted for January 1999

Day	Prediction
1	724
2	717
3	681
4	792
5	749
6	715
7	744
8	792
9	756
10	727
11	794
12	807
13	810
14	802
15	794
16	758
17	728
18	797
19	810
20	813

21	805
22	797
23	760
24	731
25	798
26	811
27	814
28	806
29	798
30	761
31	732