Averaging and data enrichment: two approaches to electricity load forecasting

Wojciech Kowalczyk

Vrije Universiteit Amsterdam
Department of Artificial Intelligence
De Boelelaan 1081A
1081 HV Amsterdam
The Netherlands
wojtek@cs.vu.nl

Abstract. The paper presents two solutions of the electricity load forecasting problem. The first solution is based on a sophisticated scheme of averaging historical electricity load data and involves estimating values of three parameters that optimize a domain specific objective function (MAPE error). The second approach exploits a relation between daily average temperature and daily electricity peak consumption. This relation has been determined by a robust polynomial regression applied to the provided data. Daily temperatures for the forecast period (January 1999) have been reconstructed from daily temperatures in Bratislava, Budapest, and Vienna. They could be easily found on the Internet.

1 Introduction

During the EUNITE 2001 symposium a forecasting competition took place. The organizers of the competition provided some historical data on electricity load and average daily temperatures in Slovakia during years 1997 and 1998. The task of the participants was to make a forecast of the maximal daily electricity loads for all 31 days of January 1999. It should be noticed that daily temperatures for this period were not provided. The quality of solutions was measured by two error functions:

$$MAPE = 100 \cdot \sum_{i=1}^{n} \left| \frac{L_{R_i} - L_{P_i}}{n L_{R_i}} \right|$$

and

$$MAXIMAL = MAX(\left| L_{R_i} - L_{P_i} \right|),$$

where L_{R_i} and L_{P_i} denote *real* and *predicted* maximal daily load, respectively.

More information about the competition and data sets can be found on the competition site: http://neuron-ai.tuke.sk/competition.

This paper is our contribution to the competition. It presents two solutions. The first one is based on the assumption that "on average" the daily maximal electricity load (in short: DML) in January 1999 will be the same as in January 1998 and 1997. The motivation underlying this approach, together with all the technical details behind the term "on average" is presented in section 3.

The second solution is more elaborated and is based on the observation that there is a strong correlation between average daily temperature and DML. On the other hand, it is a common knowledge that the average daily temperature cannot be predicted with a reasonable accuracy (at least from the data that were provided). Even a much simpler task of estimating temperature averages on a monthly basis is quite difficult and requires observations for at least 10-15 years, [Chatfield, Ripley]. In other words: for making good predictions of DML we need daily temperatures for January 1999, and these temperatures cannot be predicted from the provided data.

To overcome this problem we decided to use additional data that could be found on the Internet and use it for modeling average daily temperatures in Slovakia. Consequently, we could generate better (as we hope) predictions of DML for January 1999.

We believe that in real life predictions of DML are made just a few days ahead, and they are (or could be) based on reliable (short-term) weather forecasts. Therefore, although our second solution might seem to be invalid (our forecast is based on some data from January 1999), we are convinced that nevertheless it could be successfully implemented in practice.

2 Exploratory Data Analysis

To get an idea of what is in the data we have generated a number of bar charts, scatter plots, histograms, etc., and visually inspected them. In this way we could develop better intuitions about the problem and make some useful observations. For example, let us look at Figure 1. It shows values of DML together with daily temperatures in January 1998 (for better readability we rescaled the temperature by multiplying it by 10 and adding 500). We can easily see that:

- there are three types of days: Sundays (with lowest DML), working days (with highest DML) and remaining days (Saturdays and holidays) with DML between extremes.
- the lower the temperature the higher DML. This is clearly visible in the last 2 weeks
- the first 6 days of 1998 (Thursday-Tuesday) have DML as we would expect for Saturdays and Sundays. This can be explained by the fact that there were too many free days close to each other: New Year (Thursday), Friday (between New Year and Saturday), Tuesday (holiday) and Monday between Sunday and a holiday).

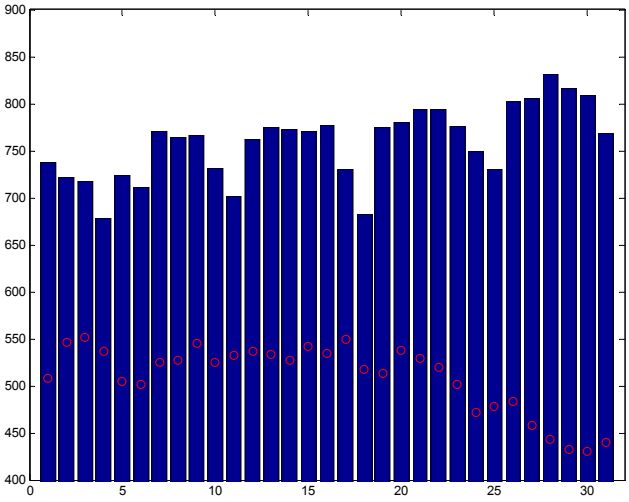

Fig. 1. Daily Maximum Load (DML) in January 1998 (bars) and average daily temperature (markers). For better readability the temperature have been rescaled (see the text above).

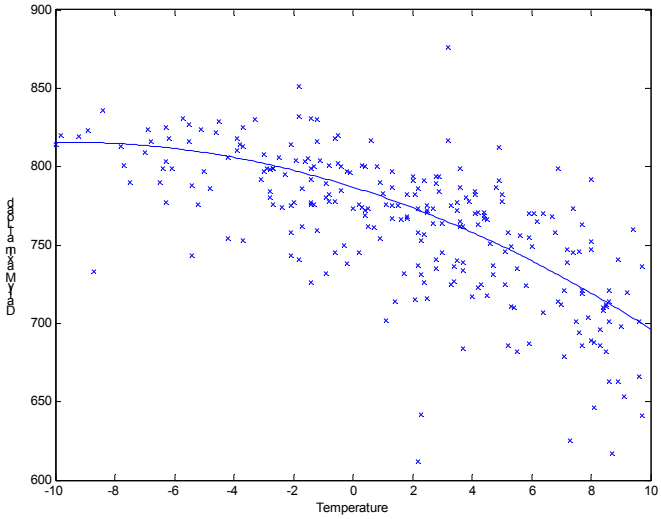

Fig. 2. Scatter plot of DML versus daily temperature. Only observations with temperature between –10°C and +10°C are shown. Non-working days are not displayed. The solid line shows the best (robust) quadratic fit.

The scatter plot shown in Figure 2 demonstrates a strong correlation between daily temperature and DML. The solid line represents the best quadratic polynomial that has been fitted with robust regression (50 "worst" observations have been ignored).

In our explorations we paid very little attention to non-winter months. We are convinced that these months are intrinsically different than winter months because factors like the time of sunrise and sunset, summer vacations, rain, etc. strongly influence DML.

3 Averaging approach

As we noticed earlier, daily temperature is highly correlated with DML. Unfortunately, the temperature data was not available for the forecast period (January 1999). Therefore, we could either try to predict the temperature in January 1999 separately, and then use it (forecasted temperature) for making the final prediction of DML, or, predict DML directly from historical values of DML.

In the literature on Time Series Analysis the problem of predicting temperature gained some substantial attention. Unfortunately, with respect to our problem, the results are very discouraging: forecasts are usually built for monthly averages and are based on data that spans periods of at least 10-20 years, [Chatfield], [Ripley]. In this situation we gave up the idea of generating daily temperature forecasts for January 1999. Also predicting the average monthly temperature for this period would be very risky (we had observations from the last 4 years only). Therefore, we decided to ignore the temperature data completely and to assume that DML in January 1999 will look, on average, similarly to DML in January 1997 and 1998.

3.1 The model

Our model is very simple. We just assume that it can be characterized by values of 3 parameters: average DML for Sundays, average DML for working days, and average DML for the remaining days (Saturdays and holidays). Let us call these parameters A_SUN, A_WORK, A_SAT, respectively. Clearly, to find these 3 "averages" we will use data from January 1997 and 1998.

The word "average" has here a special meaning: we want to minimize the MAPE error function, so for us "average" means "the value that minimizes the MAPE error function".

To be more specific, let us consider some observations $x_1, x_2, ..., x_n$. It is well known that the "normal" average, i.e, $m=(x_1+...x_n)/n$ optimizes the Mean Squared Error:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \ .$$

However, we are interested in optimizing the *MAPE* error function, and no explicit formulas are available for that. We can, however, find this minimum numerically.

3.2 Minimizing the MAPE error

Finding an *m* that minimizes the *MAPE* error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - m}{x_i}\right|$$

is very simple. We know in most cases both error functions, i.e, MAPE and MSE, behave in a similar way. This means that an *m* that minimizes MAPE is close to the arithmetic average, $\underline{m}$. Therefore, to find an optimal m it is enough to search for it in a small neighborhood of $\underline{m}$, e.g., in [0.8$\underline{m}$, 1.2$\underline{m}$]. This can be done by calculating the value of the MAPE error for, say 10000 points that are uniformly distributed over this interval, and picking the best one.

3.3 Bootstrap

It is well known that when estimating parameters of probability distributions from a small number of observations (as in our case), it is useful to resample the data (with replacement), say 1000 times, estimate parameters for these 1000 samples and then average them. This method, called bootstrap, is widely used in statistics, [Efron].

We used this method for estimating our parameters. In addition to estimating the MAPE-average, we also estimated the "normal" average, the minimum and the maximum.

3.4 Results

The whole estimation process (minimization of the MAPE error combined with the bootstrap method) yield the following table (MMEAN denotes here the value that optimizes MAPE):

	MIN	MEAN	MMEAN	MAX	MSE	MAPE
Working	763.248	793.2551	795.0047	827.725	0.0198	0.0195
Saturday	714.284	741.5027	737.1926	785.262	0.0221	0.0209
Sunday	681.644	708.823	707.6144	729.432	0.022	0.0208

We can see that there are some differences between MSE-means and MAPE means; the expected improvement in error reduction is almost neglectable. However, with respect to MAXIMAL error, the MAPE predictions seem to be better. For example, the value that minimizes the MAXIMAL error on "working days" is 795.4865 is closer to the corresponding MMEAN than to MEAN.

The final forecast is therefore based on values of MMEANS and is provided as the WK1.TXT file.

4 Data Enrichment Approach

As we mentioned earlier, the knowledge of daily temperatures is essential for making good forecasts. As these temperatures were not provided by the organizers of the competition we decided to get them the Internet. For example, the Temperature Data Archive of the University of Dayton, http://www.engr.udayton.edu/weather/, contains daily temperatures for most capital cities for the last couple of years. We have used data from three cities that are closest to Slovakia: Bratislava, Budapest and Vienna.

4.1 Modeling the average temperature in Slovakia

After fetching the data we converted it from Fahrenheit to Celsius scale and replaced missing values (very few) by averaging the closest neighboring values. Then we constructed several linear models for the 4 years (1995-1998) using the temperature data that was provided by the organizers. It turned our that the data from Budapest is the best predictor: the correlation was 0.9809. A slightly better model (with correlation 0.9812) has been obtained by using data from all three capitals:

$$t\_Slovakia = 0.8405 \cdot t\_Budapest + 0.2503 \cdot t\_Bratislava - 0.0996 \cdot t\_Vienna - 1.2278.$$

By applying the model to temperature data for January 1999, we found the following estimates of the average daily temperature is Slovakia during that period:

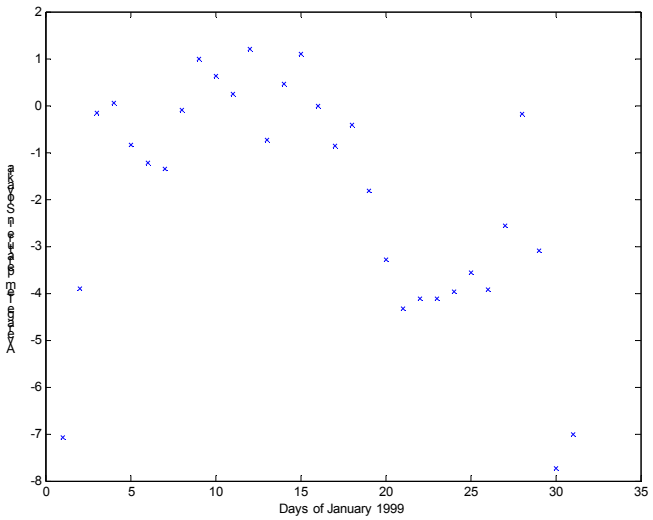

Fig. 3. Estimated average daily temperature in Slovakia in January 1999.

4.2 Modeling the temperature – DML relation

To model the relation between the temperature and DML we used all observations for working days such that the temperature was between –10°C and +10°C, see Figure 2. We noticed numerous outliers and non-linear characteristic of this relation. Therefore we assumed that the dependency could be modeled by a polynomial of second degree. To eliminate the influence of outliers we run the robust regression algorithm [3]. This algorithms iteratively removes from the data set points that are most remote from the regression curve, and finds a new curve for the smaller data set. We iterated this algorithm 50 times, ending up with the following formula:

$$DML(t) = -0.3105 \cdot t^2 - 5.9725 \cdot t + 786.9231.$$

By applying this formula to the temperatures calculated in the previous step we've got predictions for all working days of January 1999.

4.3 Non-working days

Because the number of observations for non-working days was too small to repeat the above procedure, we simply assumed that DML for such days could be calculated by multiplying the average DML for nearest working days by a constant. These constants have been estimated from our first solution: for Sundays this constant turned out to be 0.8901 and for Saturdays (and holidays) 0.9273.

By combining all three sorts of predictions we have generated our second solution that is included in the file WK2.txt.

Both solutions, together with daily temperature estimates can also be found in the appendix.

5 Conclusions

Two solutions have been presented. The first one, based on a simple averaging idea, should provide forecast with the MAPE error of about 2.5% and the MAXIMAL error of about 30-50. The second one, based on reconstructed daily temperatures, should result in a slightly better prediction (the MAPE error should be below 2%). The first few days of January, esp. New Year, will probably cause a big MAXIMAL error – these days are not typical. Instead of correcting predictions for these days manually, we left them as they are, for the sake of simplicity of the whole modeling process.

6. References

1. Chatfield, C.: The Analysis of Time Series. Fifth Edition, Chapman and Hall,1996.
2. Efron, B., and R. Tibshirani: An introduction to the bootstrap. Chapman and Hall, 1993.
3. Rousseeuw, P.J., and A.M. Leroy: Robust regression and outlier detection. John Wiley, 1987.

7 Appendix: Summary of Results

The following table contains our predictions for January 1999.

Day	Solution 1	Temp.	Solution 2
1	737.1926	-7.0826	754.478
2	737.1926	-3.9011	746.9212
3	707.6144	-0.1691	701.312
4	795.0047	0.0534	786.6034
5	795.0047	-0.8299	791.6658
6	737.1926	-1.224	736.0459
7	795.0047	-1.3446	794.392
8	795.0047	-0.0979	787.5049
9	737.1926	0.9907	723.9298
10	707.6144	0.6181	697.0297
11	795.0047	0.2456	785.4378
12	795.0047	1.2083	779.2529
13	795.0047	-0.7316	791.1264
14	795.0047	0.4651	784.0781
15	795.0047	1.0957	780.0061
16	737.1926	-0.0091	729.7493
17	707.6144	-0.8663	704.8187
18	795.0047	-0.4059	789.2959
19	795.0047	-1.8208	796.7684
20	795.0047	-3.2893	803.2082
21	795.0047	-4.3229	806.9379
22	795.0047	-4.1252	806.2759
23	737.1926	-4.1257	747.646
24	707.6144	-3.9742	717.1819
25	795.0047	-3.5677	804.2782
26	795.0047	-3.9299	805.5981
27	795.0047	-2.5645	800.197
28	795.0047	-0.1713	787.937
29	795.0047	-3.0958	802.4363
30	737.1926	-7.7252	755.2966
31	707.6144	-7.0191	724.1161