Artificial neural networks based short-term forecasting of daily electrical energy consumption: an approach using consumption data subsets

Domen Kodrič, Gašper Podobnik, Martin Žukovec, Jan Gašpar

University of Ljubljana, Faculty of Electrical Engineering
E-mail: dk1824@student.uni-lj.si

Abstract. *The purpose of this paper is to present a possible approach towards artificial neural networks based short-term predictions of daily electrical energy consumption (EEC) in Slovenia and its practical application. Since power consumption depends on complex non-linear relationships between several influential variables, artificial neural networks are commonly used in creating EEC forecast models. A characteristic of power consumption is a superposition of multiple time-based patterns, which can be recognised, and through classification of past EEC data, data conforming to patterns can be grouped, and specific submodels can be designed. We show that the chosen constructed data subsets contain less noise, and that by using submodels built upon these subsets, we produce predictions that are more accurate than those made using a single dataset approach.*

1 Introduction

Electrical energy consumption (EEC) is not constant. It changes gradually through years, differs between seasons of the year, fluctuates from day to day (workdays, weekends), and can steeply increase and decrease from hour to hour. Power system operators, electrical energy traders, as well as electrical energy producers all benefit from accurate future EEC prediction.

In this paper, we describe an artificial neural networks (ANN) based model for producing short-term (1-3 days) forecasts of daily electrical energy consumption in Slovenia. We use past data on electrical energy consumption, past seasonal data, and past meteorological data. A characteristic of the power demand through time is a superposition of multiple patterns that depend on classifications such as day of the week, day of the year, holiday etc. An overview of the existing literature on (ANN based) short-term EEC forecast considering these factors offers two basic approaches to the problem. Approach A (example: [4]) involves assigning values for each variable (for example 1-7 for days of the week), and creating a single ANN. Approach B (by Lee already in 1992 [2]) employs classifying the data based on the aforementioned factors, and designing multiple ANN models, each assigned to produce forecasts for a specific type of day. It appears that in similar problems on short-term EEC forecast, approach A prevails [1, 3, 4, 5]. In the paper, we employ approach B: we dissect the data into subsets, and show improvement in prediction accuracy over a model built with a single dataset. Using two resulting specific models (considering only January

workday-days/weekend-days), we produce predictions of daily consumption for 18 days of January 2018, with results comparable to publically available predictions.

2 The data

Data on the energy flow from the transmission network in hourly intervals from Jan 1st 2016 onwards are available on the website of the Slovenian electricity transmission system operator (ELES) [7]. The company has, upon request, generously provided us the data for years 2010-2017 by e-mail [8] (69600 hourly EEC values).

The data represent the electrical energy flow from the transmission network into distribution networks and to directly connected clients; this includes losses in the transmission network [8]. We therefore forecast the electrical energy flow from the transmission network, and not the final electrical energy consumption. The two quantities are, however, closely related, and the term electrical energy consumption (EEC) will be used when describing correlations between variables for intuitive reasons.

Although the research primarily focused on the effect of date variables, we have considered the impact of meteorological variables in order to produce more accurate forecasts. We acquired meteorological data from the web servers of the Slovenian Environmental Agency [9]. These encompassed the average, minimal and maximal values of air pressure, temperature and relative humidity, average values of diffuse and total solar irradiation, and total precipitation, in hourly intervals. We acquired the data from the automatic station Ljubljana Bežigrad, because we estimated that due to its location (in the centre of the country) and a large population living in its vicinity would provide the most relevant meteorological data.

Empty data cells were filled using linear interpolation.

3 Data analysis

We avoided using too many meteorological variables, and only considered the most influential ones. In order to get a rough estimate of the importance of meteorological variables, we trained with Matlab R2016b a simple regression tree, which offers a straightforward insight into predictor importance. We trained regression trees first using hourly and then daily values of all meteorological variables as inputs, while the output of the regression tree was the value of EEC.

The air temperature consistently proved itself as the variable with the most influence on EEC. With significantly lower importance, total radiation and relative humidity followed; however, both of these are already related to the air temperature themselves, and we therefore did not include them. Instead, we added a variable *dT*, the change in temperature from the previous hour/day, to observe the effect it has on the change in EEC relative to the previous hour/day, as in human perception, the *change* of temperature also plays a role.

We then analysed the whole EEC dataset visually over time periods of different magnitude. We discerned a superposition of multiple patterns that the power consumption follows through time. Plotted daily consumption values against the days of a year (DoY) offer an insight into the difference between seasons. Figure 1 shows the difference in consumption between seasons of the year.
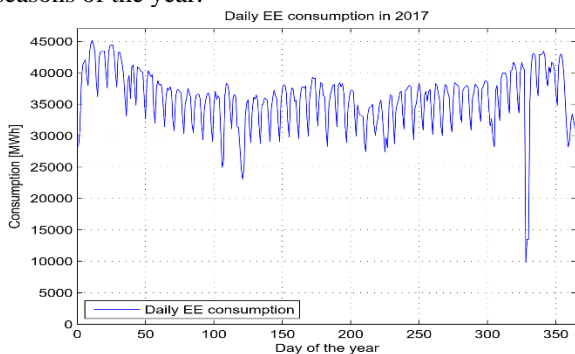

Figure 1. Daily EEC in 2017.

We can observe greater consumption values in the winter months. The second, smaller peak is seen around DoY 150-190 (the month of June and beginning of July). Noticeable deviations can be noted in the days around day 105 (Easter holidays), day 120 (May Day), and the substantial decrease in consumption around day 325 (a consequence of a weather induced power outage). These observations are referenced in Chapters 3.1 and 4.

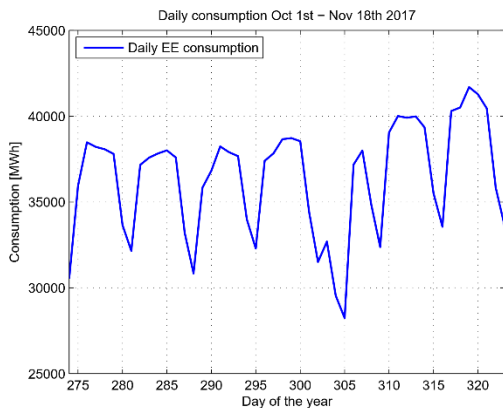The following Figure 2 represents the daily EEC through seven consecutive weeks in Oct and Nov 2017.

Figure 2. Daily EEC through 7 weeks of 2017.

We notice the periodic weekly pattern of EEC. Highs represent middle-of-the-week workdays (Tuesday -

Thursday), while lows represent Sundays. A deviation between day 301 and 304 can be attributed to two bank holidays and school holidays in that period.

Another pattern can be observed in the recurring shape of the hourly EEC curve with two distinct peaks in the morning and evening hours. Figure 3 shows the hourly values of EEC on Thursday, Oct 12th 2017.
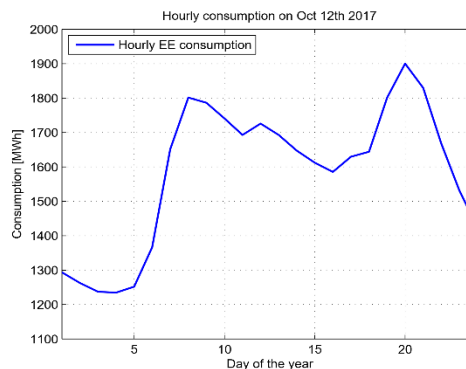

Figure 3. Hourly EEC on Oct 12th 2017.

3.1 Data classification

Considering the recurring patterns recognised in the previous chapter, we deduced that date variables (such as DoY, day of the week (DoW), and bank holidays) would need to be taken into consideration when creating a prediction model. In the following paragraphs, we describe an analysis of the effect that the day of the week has on the EEC pattern.

We used the hourly EEC data from Jan 1st 2010 to Dec 11th 2017. To determine the (dis)similarity of the pattern of the hourly EEC on different DoW, we calculated the standard deviation of the EEC for every hour of the same DoW, for all days of the week. To get an insight into normalized dispersion of the hourly values of EEC in different DoW, we calculated the coefficient of variation (CV) of the hourly EEC. ($CV = \sigma/\mu$; σ represents standard deviation and μ represents the mean value of the EEC in a specific hour.)

Figure 4 shows the CV of hourly values of EEC on all Thursdays in the dataset.
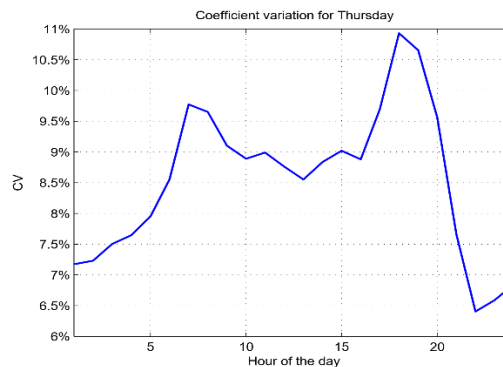

Figure 4. CV through an average Thursday.

A lower CV means a lower relative dispersion of hourly values, and demonstrates the degree of similarity of a day's consumption pattern. The peaks of CV in the

morning and evening show that the relative variance in EEC is greatest between the hours of 6-9 and 16-19 – periods that coincide with peaks in the average hourly EEC.

To evaluate the similarity of the consumption pattern between different days of the week, four charts are shown together in the following Figure 5. The blue, triangle-point curve represents the hourly coefficient of variation (CV) on Thursdays (same as Figure 4), the red, continuous curve represents the hourly CV on Saturdays, and the green, dashed curve represents the hourly CV of a combined set of Thursday and Saturday data (which gives one combined average around which data for both days are dispersed). Similarly, the black, continuous/dotted curve represents the hourly CV of a combined set of Thursday and Friday data.
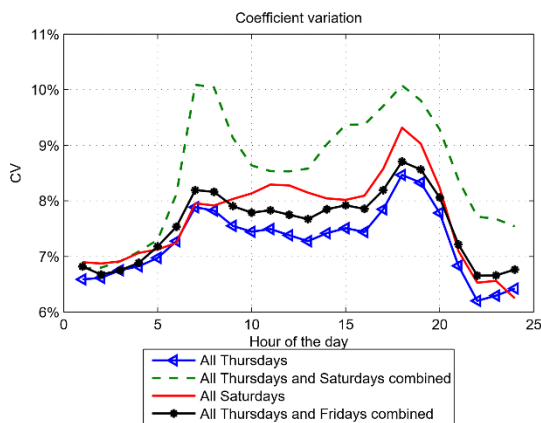

Figure 5. Comparison of CV's of different datasets.

Comparing the curves, the difference between CV's of the Thursday and the Thursday+Saturday datasets is the most obvious. While the CV of the Thursday dataset never exceeds 11 % (max 10.93 %, average 8.54 %), and the CV of the Saturday dataset peaks at 12.64 % (averages 9.39 %), the CV of the combined Thu+Sat is noticeably higher and peaks at 14.18 % (averages 11.04 %). CV of the Thu+Fri dataset, (max 11.41 %, average 9.13 %) presents a smaller increase from the Thu value.

The differences in CV of the datasets show that the consumption curve pattern is less predictable, when the datasets of two different days of the week are combined. Furthermore, the similarity between patterns of a workday and a weekend day (Thursday+Saturday) is substantially lower than the similarity of two workdays (Thursday+Friday). Since combining datasets of different days of the week into one dataset evidently produces a dataset with more noise, it follows that it may be meaningful to produce different prediction submodels, trained on certain subsets of similar data, at the cost of less data in a subset.

Figure 6 is a visualisation of the similarity between all weekday datasets. The area of the circle represents the RMS value (in MWh) of the difference between respective hourly EEC values of two different DoW, for all 414 weeks (RMSd). Dissimilarity is greatest between Thu and Sun (RMSd = 319.0 MWh), and smallest between Wed and Thu (RMSd = 69.9 MWh).
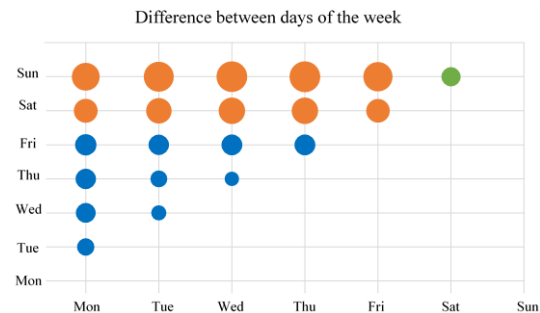

Figure 6. Comparison of (dis)similarity between days of the week.

Using a similar-days approach, Mandal [6] splits the days of the week into 4 classes (Monday, Saturday, Sunday, and weekdays comprising Tue-Fri). Regarding our dataset, we propose that if one were to account for Sat-Sun distinction (RMSd Sat-Sun = 123.3 MWh), it would follow to also account for at least the distinction between Tue and Fri (RMSd = 140.2 MWh), if not to take each DoW as an individual dataset (RMSd Tue-Mon = 102.4 MWh, RMSd Tue-Wed = 77.6 MWh, RMSd Tue-Thu = 95.6 MWh).

4 Devising a model for daily forecast

Mandal [6] uses a similar-days approach to produce 1-6 hours-ahead forecasts with a moving dataset, using data 65 days before the forecast day, and 65 days before and after the forecast day in the previous year. In our approach, we pre-determined the similar days, and our dataset (considering the analysis in Chapter 3) included the variable *day of the year* to account for seasonal similarity. Furthermore, we looked for similarity in all 8 years of data, and included *bank holidays* as an influential variable.

We first constructed a model using the entire dataset (daily EEC values from Jan 1st 2010 to Dec 11th 2017). Since the relation between the input variables and the output EEC value is non-linear, and the form of non-linearity is unknown, we have devised a model using artificial neural networks (ANN), an artificial intelligence method. We separated the data into a 'learning set' and a 'test set', and used the learning set to train in RapidMiner an ANN, which would produce the best dependency algorithm between the input (daily meteorological and date) variables and the output values of daily EEC. Input data of the 'test set' were then applied to the constructed ANN, and we evaluated the accuracy of the resulting predictions by comparing them to actual daily EEC values. Multiple combinations of the number of hidden layers and the number of neurons were tested, and minimization of the mean APE (absolute percentage error) was then used as the criterion in selecting the optimal values. The final ANN consisted of an input layer (6 neurons; one for each input variable), an output layer (one neuron for the output variable), and two hidden layers with 15 neurons each. Using the whole dataset, the mean absolute percentage error (APE) between predictions and actual daily EEC values was 3.6%. The maximum APE was 9.7 %.

4.1 Improvement of predictions using data subsets

We then trained multiple ANN's with different data subsets. We first trained an ANN using workday-day data only – we eliminated weekends and bank holidays. We also eliminated prediction-irrelevant datasets with atypically low EEC. The results showed an improvement in prediction accuracy compared to the general model – mean APE was 3.3 %, and maximum APE was 7.0 %. A model using only weekend data (without bank holidays) gave a mean APE of 3.7 %, maximum APE was 7.3 %.

Substantial improvements in prediction accuracy were reached when even stricter filters were applied, and forecasting models became more and more specific. A model using only workdays from March until May (days of the year 60 - 150) returned a mean APE of 2.5 %, maximum APE was 5.9 %. A model encompassing workdays from May to July (DoY 120 - 210) performed with mean APE of 1.9 %, maximum APE was 3.4 %.

5 A practical application of the model

We trained an ANN using only January workday-days, with no bank holidays and no prediction-irrelevant data. We produced a forecast (temperature predicted with [10]) of the EEC for the following day. Forecasts were made at midnight, with the previous day's actual EEC known. An analogous model was made for January weekends.

Table 1 shows our daily prediction along with the actual daily EEC. The prediction on the website of the Slovenian electricity transmission system operator (TSO) serves as a reference. Absolute percentage error (APE) describes the accuracy of predictions; a green background marks the lower of the errors. For three of the days in the time span, predictions were not made.

6 Conclusion

Through the analysis of past EEC data and through practical examples, we showed that when using artificial neural networks, the creation of multiple submodels with appropriate data subsets results in an improvement of prediction accuracy. We devised such a submodel for a specific timeframe (the month of January) and produced practical predictions of the daily EEC.

When predicting hourly EEC, the selection of subsets could be made similarly, and additional classification could be introduced by periods of the day (e.g. peaks of CV during hours of high demand), or by days with similar meteorological conditions. Further specializing the submodels reduces the quantity of learning data in the submodel's dataset – an optimum has to be found between increasing the number of submodels and the resulting decrease in learning data per submodel.

Acknowledgement

Table 1: Real predictions in January

Date (Jan)	Actual EEC [MWh]	Our forecast [MWh]	APE of our forecast	Forecast at the TSO's website [MWh] [7]	APE of the TSO's forecast[7]
8	37948	37695	0,67%	40434	6,55%
9	39543	38513	2,60%	39754	0,53%
10	39813	38725	2,73%	39646	0,42%
11	39304	39415	0,28%	39568	0,67%
12	39793	40226	1,09%	39889	0,24%
13	35653	/	/	35280	1,05%
14	33732	/	/	33663	0,20%
15	40701	40151	1,35%	39680	2,51%
16	41472	41359	0,27%	41106	0,88%
17	41612	40597	2,44%	42071	1,10%
18	41258	40931	0,79%	41257	0,00%
19	41701	40307	3,34%	41694	0,02%
20	37414	36960	1,21%	36358	2,82%
21	33933	35108	3,46%	34179	0,72%
22	40265	40432	0,41%	39819	1,11%
23	40274	42368	5,20%	40941	1,66%
24	40760	42491	4,25%	41635	2,15%
25	41422	41786	0,88%	41271	0,36%
26	40771	40355	1,02%	40792	0,05%
27	36838	/	/	36665	0,47%
28	34123	33882	0,71%	33596	1,54%

References

[1] I. Chernykh, D. Chechushkov, and T. Panikovskaya. "The prediction of electric energy consumption using an artificial neural network." *Energy Production and Management in the 21st Century: The Quest for Sustainable Energy* 190 (2014): 1109.

[2] K. Y Lee, Y. T. Cha, and J. H. Park. "Short-term load forecasting using an artificial neural network." *IEEE Transactions on Power Systems* 7.1 (1992): 124-132.

[3] M. Buhari, and S. S. Adamu. "Short-term load forecasting using artificial neural network." *Proceedings of the International MultiConference of Engineers and Computer Scientists*. Vol. 1. 2012.

[4] L. Hernández et al. "Artificial neural network for short-term load forecasting in distribution systems." *Energies* 7.3 (2014): 1576-1598.

[5] K. Gajowniczek, R. Nafkha, and T. Ząbkowski. "Electricity peak demand classification with artificial neural networks." *Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on*. IEEE, 2017.

[6] P. Mandal et al. "A neural network based several-hour-ahead electric load forecasting using similar days approach." *International Journal of Electrical Power & Energy Systems* 28.6 (2006): 367-373.

[7] ELES d.o.o. »Prevzem in proizvodnja,« eles.si [online] Available: https://www.eles.si/prevzem-in-proizvodnja [Accessed 30. 1. 2018].

[8] Info ELES, »Poizvedba po podatkih za raziskovalne namene.« Personal email (Dec 21th 2017).

[9] ARSO METEO, »Uradna vremenska napoved za Slovenijo – Vreme podrobneje,« arso.gov.si [Online] Available: http://meteo.arso.gov.si/met/sl/app/webmet/ [Accessed Dec 21th 2018].

[10] The Weather Channel, »Vremenska napoved in pogoji« weather.com, [Online] Available: https://weather.com/sl-SI/vreme/danes/l/SIXX0002:1:SI [Accessed Jan 27th 2018].