ISBN: 978-989-8533-80-7 © 2018

APPLIED MACHINE LEARNING: PREDICTING BEHAVIOUR OF INDUSTRIAL UNITS FROM CLIMATE DATA

Dieter Meiller[1] and Christian Schieder[2]
[1]*East Bavarian Technical University Amberg-Weiden, Kaiser-Wilhelm-Ring 23, 92224, Amberg, Germany*
[2]*BHS Corrugated Maschinen- und Anlagenbau GmbH, Paul-Engel-Straße 1, 92729 Weiherhammer, Germany*

ABSTRACT

The goal of this project was to develop a model and a working prototype for the evaluation of energy consumption of a factory in connection with climate data. The factory produces corrugated card board. This product is very susceptible to climate influences. At the end of the project, it should be clear what options there are for using data from the factory in conjunction with climate data and deriving a correlation between them. The following questions should be clarified: Which possibilities of prognosis and evaluation are there? Which data are needed? How big is the effort?

KEYWORDS

Machine Learning, Virtual Sensor, Climate Prediction, Industry 4.0

1. COURSE OF THE PROJECT

It is assumed that climate has a strong influence on production processes. Humidity and temperature do not only have an influence on the quality of products, but also on the energy consumption of industrial plants. In the project a model for the evaluation of plant data in connection with climate data was developed, which shows the possibilities to derive correlations between these data.

After starting the project, measurements of humidity and temperature values at the production site for corrugated board should be started soon. Data should be recorded for as long as possible and should be measured at intervals of one minute. For this, about ten data loggers were installed on site. We decided to construct complete new data loggers by our own. Reasons were: Commercial data loggers are either very expensive and must be configured extensively, or they are cheap devices that do not meet the requirements. The components for the data logger are: Arduino Pro Mini, a humidity sensor, a real-time clock with better accuracy than the clock on the board, and a NTC thermistor, which has a temperature accuracy of 0.05 degrees Celsius. The housing was made using a 3d printer. Then the data loggers were installed at different places at the factory. The company itself provides the data of the power consumption of single machines for the period of the measurement. After six weeks, the first data were collected, so we could start with our first analysis. The following describes the implementation and results of the data analysis.

2. DATA ENGINEERING

First, the available data sets are described briefly. On the one hand, it is climate data: humidity and temperature values from inside the factory-hall. Then, there are weather data from outside, also temperature and humidity values. Humidity values are measured as relative humidity (RH) in percent (labeled as humidity in the diagrams). Further, we have data sets with values for power consumption (watts) of individual parts of the plant. We analyzed two kinds of data: Long-term (over four months) and short-term (six weeks). In the first step, the data sets were preprocessed. This step was necessary to get data with consistent time frames and without recording errors. For data analysis, we used Python programming language in the Anaconda environment. This free collection of open source libraries offers many opportunities in the field of data

engineering and data science. The software is constantly being developed by an active community. Individual processing sessions can be performed in so-called "Jupyter-Notebooks" (Kluyver et al. 2016). All our processing and analysis steps were executed in such notebooks. For pre-processing, several notebooks were created using the Pandas library (McKinney 2011). This library can read data from different sources and offers many possibilities for processing, cleansing, analysis and visualization. In the first step, climate-data were merged with energy-data and brought to a uniform time scale: two-minute intervals in a uniform period. The plot shows the result of a short-time dataset for a single part of the factory (Fig. 1).

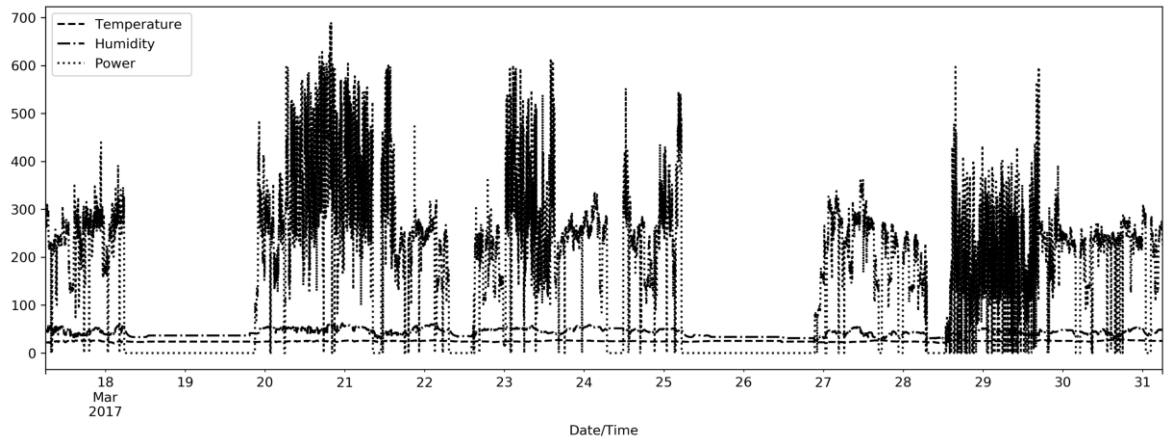

Figure 1. Plot Humidity / Power

3. DATA SCIENCE

The following analysis steps are devoted to the development of a forecast model. For this purpose, methods of machine learning were used. The technical basis was Python library Scikit-Learn (Pedregosa et al. 2011). Basically, a distinction is made between supervised and unsupervised learning. It is possible to use unsupervised learning to analyze climate data (Meiller 2017), but in this paper we focus on supervised learning. The difference is that during supervised learning there is a target size (y) of the data in addition to the individual features of the respective samples. In our case, these are the values for energy consumption. Because the values y are not given as discrete, but as continuous values, we can use regression analysis to predict these values. There is another term to be explained: „virtual sensors“ (Kabadayi et al. 2006). When it is possible to predict values, the model could replace the sensors for measurement. Because there are lots of expensive built-in sensors, this could be an option.

4. DATA PREPARATION

Data analysis and associated modeling was done by an iterative process of experiments. The aim was to have the best possible forecast of energy consumption based on the measured climate data. Different methods of machine learning with different parameters were tested. In addition, the input data has been preprocessed in various ways. Various combinations of possibilities were compared. This process took up most of the data analysis. At the beginning, the data were normalized, so that all data fit in an interval of [0,1]. This step is necessary because energy values are measured in scales other than climate data. After that, the data were split into two parts. The first part of the data (2/3rd) is the training set, the rest are test data. It was taken into account that the data is time-based, so that individual combinations of climate and power measurements have predecessors and successors in time. Therefore, we avoided to pick samples randomly to form training and test data from the original data set. The aspect of transient behavior would be lost as information. In order to be able to map time even better, higher-dimensional data were also generated from initial data (climate data): at each point in time, a vector was generated which represents a time interval around the selected point in

time (Kapoor and Bedi 2013). In Fig. 2, the process of feature generation is shown schematically, here with a time window of 80 minutes. The number of features (m) has thus increased from two features (humidity, temperature) to 2 * 80 features. The high-dimensional training data generated in this way then were used as input data for various learning methods.

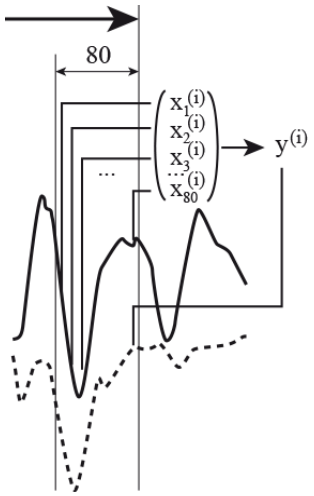

Figure 2. Feature Generation

5. MODEL TRAINING AND EVALUATION

The coefficient of determination(R^2) (see Formula 1) was used as quality measure for success control. It is interesting to know, that there are several definitions of R^2. The definition we used from the scikit-learn library is the same as the R^2_1 definition mentioned in (Kvålseth 1985). It is used to indicate how successful the model is. It compares the values for power consumption $\hat{y}$ estimated by the model from the test climate data with actual values for power from test data y. The values can be a maximum of 1, which would correspond to a match of 100%. Values close to 0 occur when taking constant values as estimated values $\hat{y}$. The estimate can be arbitrarily bad, values smaller than 0 indicate this. Using a multi-layered neural network, values of individual data loggers with current values of specific plant sections were evaluated. The best R^2 of 0.49 resulted from climate data measured on a machine located apart from other machines and its corresponding power data. We conclude that a measurement at a greater distance to the main-plant is more favorable, since there are no extreme fluctuations in temperature and humidity. Afterwards, further learning algorithms and parameters were tested to improve the results. One special preprocessing step was important to improve our results and to produce better visualizations: the smoothing of measured values using a Savitzky-Golay filter (Schafer 2011). The data were interpolated within a window of 351 minutes or less through a third degree curve (Fig. 3).

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n_{samples}-1}(y^{(i)}-\hat{y}^{(i)})^2}{\sum_{i=0}^{n_{samples}-1}(y^{(i)}-\bar{y})^2}$$

where

$$\bar{y} = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1} y^{(i)}$$

Formula 1. Coefficient of Determination

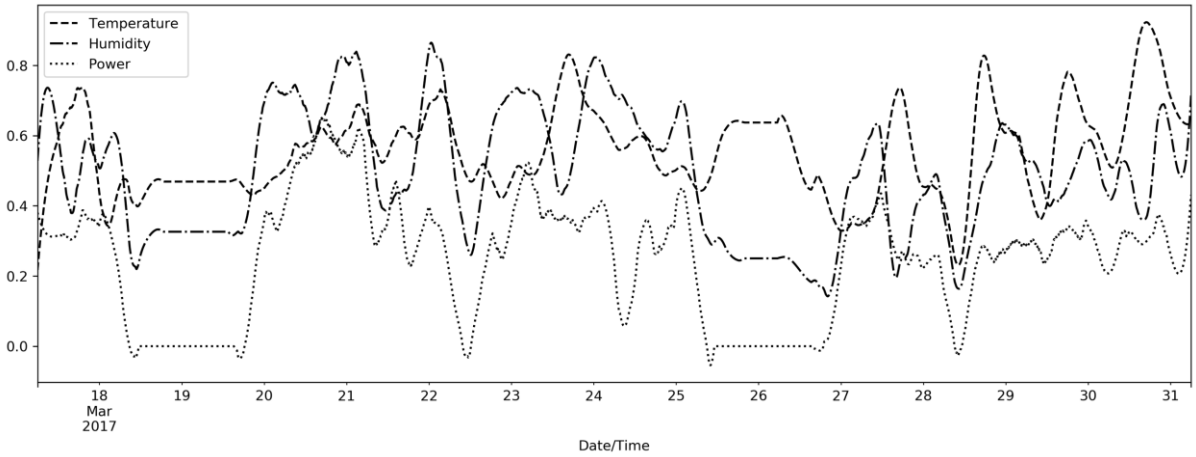

Figure 3. Plot Humidity / Power scaled and interpolated

We paid attention to split training and test data before preprocessing, hence before scaling and smoothing, otherwise information from test data could be leak in training data. In contrast to earlier studies (Radhika and Shashi 2009) (Rao et al. 2012), using the data mentioned to train a support vector machine gave mediocre results (R^2 value of 0.45). A trained neural network regressor gave an R^2 value of 0.58.

Better results of R^2 were produced with a random forest regressor: 0.68. Thus, it seems that using the latter regressor is a better choice than the former for making forecasts. Though, using a gradient boosting regressor produces a R^2 value of 0.71. Finally, after adjusting the alpha value and the number of hidden nodes, the neural network achieved the best value for R^2 of 0.77 (see Fiure 4).

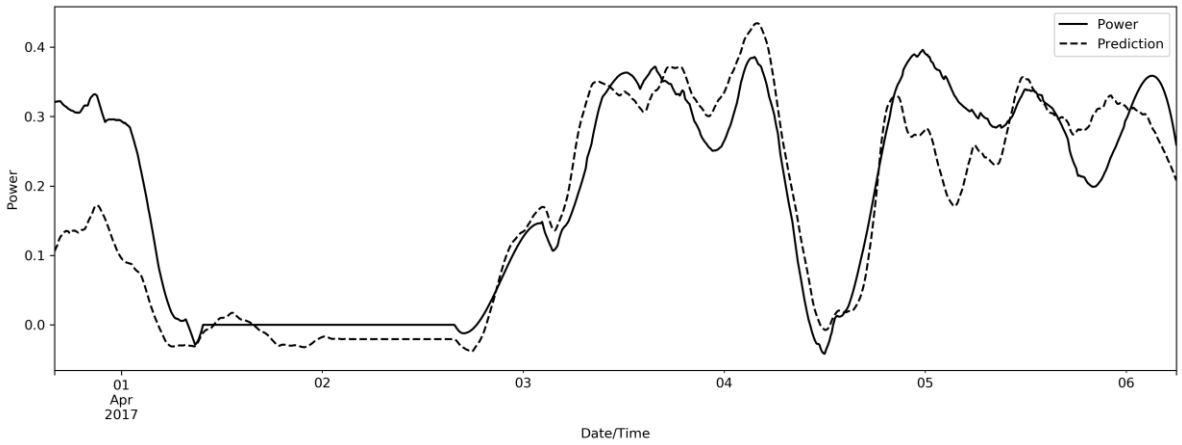

Figure 4. Prediction of power consumption (neural network)

In the next step, the long-term climate data were analyzed. The R^2 value for data from the already mentioned location was 0.4 (see Figure 5).

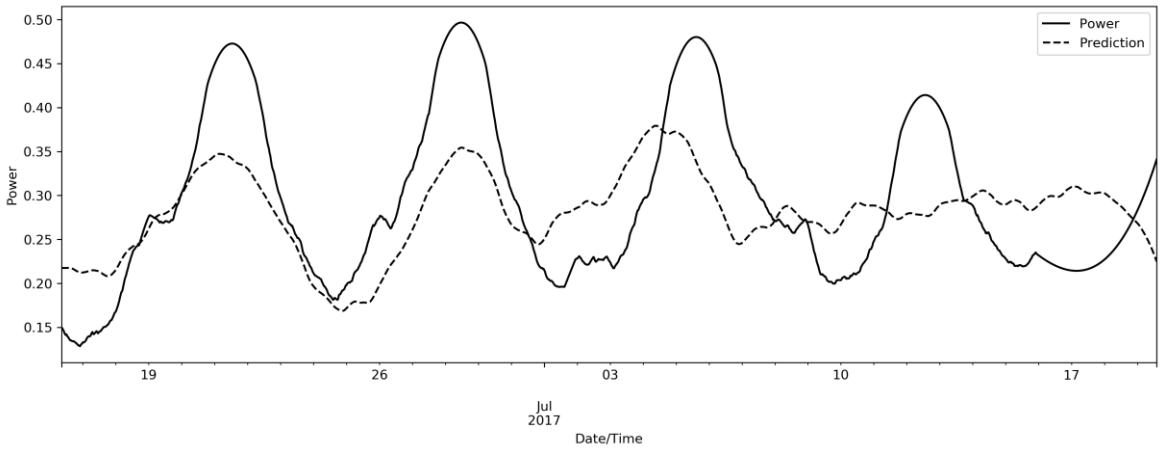

Figure 5. Long term prediction of power consumption

Thus, a short-term estimate with data of three weeks is better than a longer-term one with data of several months. Nevertheless, a trend can be predicted over a longer period of time, since the R^2 value is above 0 and also the curve of the estimated power values shows a similar progression to the actual values in the first days.

In addition to the climate data measured in the hall, weather data from outside the factory were included in the analysis. The R^2 value for predicting power consumption with weather data (two months) using a neural network was even good: 0.76. The curve-progression shows that the power consumption decreases when the humidity of the weather (outside) increases. This could be the proof that high humidity allows a static discharge, which reduces mechanical friction and so power consumption (Mardiguian 2009). Thus, a correlation between weather and power consumption of the system can be determined (Figure 6).

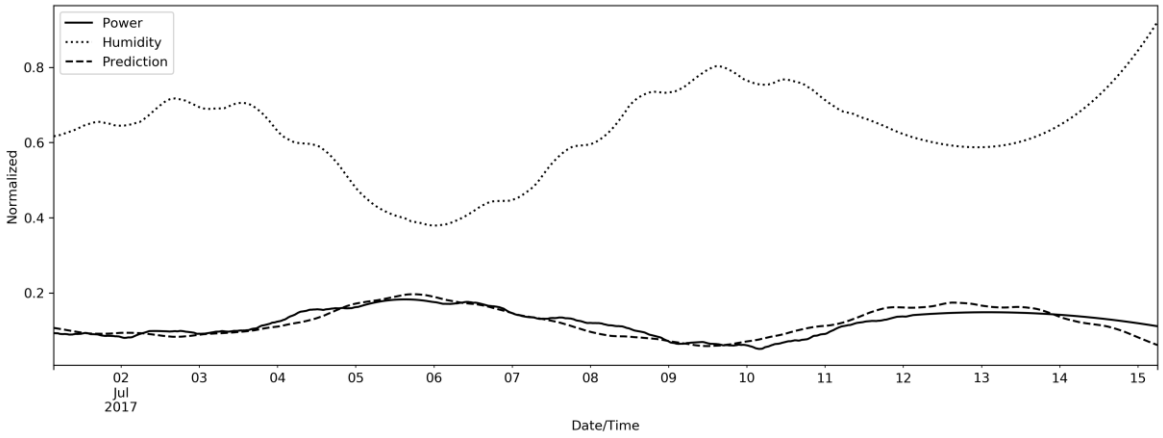

Figure 6. Prediction of power consumption using weather data

Finally, the relationship between the weather outside and the climate in the hall was examined. The input data were weather and climate data over a period of four weeks. As before the data were split into training and test data. It turned out that it is possible from training data to generate a correct estimate of nine days of indoor climate, with an R^2 value of 0.62. The progression of the curve for interior humidity corresponds to the estimated values (Figure 7).

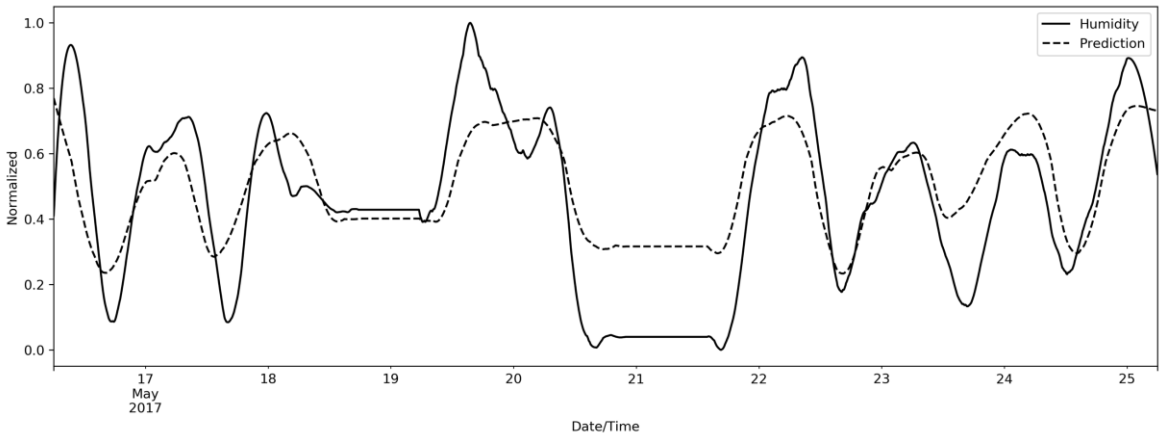

Figure 7. Correlation between weather and climate

Examination of a longer period of three months did not give a good result (R^2 value of -1.78) (Figure 8).

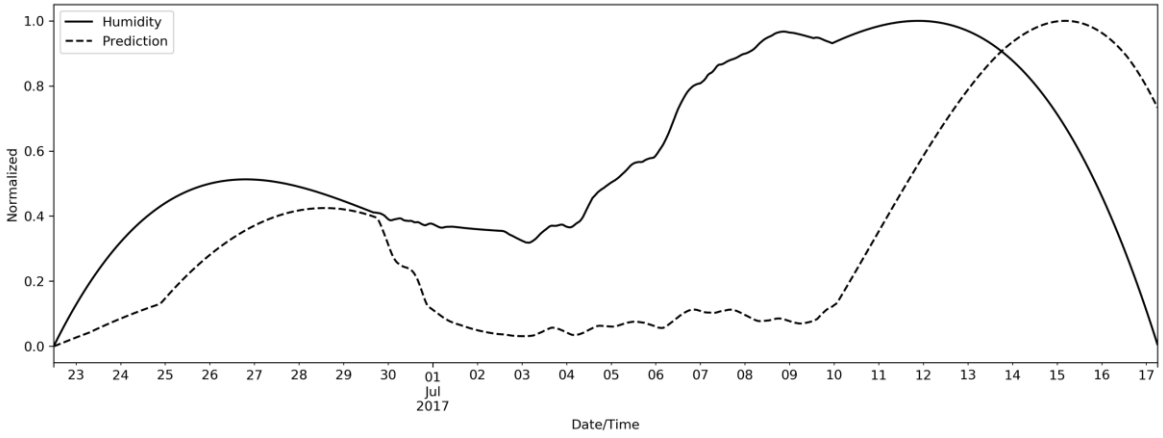

Figure 8. Correlation between weather and climate (long term)

6. CONCLUSION

The project dealt with the evaluation of plant data in combination with climate data in order to prove any correlations. The feasibility of this approach was proved. A correlation between the climatic data in the hall and the power consumption of the system parts could be proven. Weather conditions certainly have an influence on the power consumption of the plant components. For stable and correct forecasts, however, continuous measurements of power consumption should be available. The influence of the weather on the indoor climate could be proven. Short-term forecasts of the indoor climate can be made from existing weather data.

ACKNOWLEDGEMENT

We thank Norbert Städele, Nadine Fröhlich, Andreas Gmeiner, Susanne Kriesche and Florian Schöler-Niewiera for their valuable suggestions.

REFERENCES

Kabadayi, S. et al, 2006. Virtual sensors: Abstracting data from physical sensors. In *Proceedings of the 2006 International Symposium on on World of Wireless, Mobile and Multimedia Networks*, pp. 587-592.

Kapoor, P., and Bedi, S., 2013. Weather forecasting using sliding window algorithm. *ISRN Signal Processing, 2013*.

Kluyver, T. et al, 2016. Jupyter Notebooks-a publishing format for reproducible computational workflows. *In ELPUB*, pp. 87-90.

Kvålseth, T.O., 1985: Cautionary Note About R2. *In The American Statistician.*, Vol. 39, No. 4, pp 279-285.

Mardiguian, M., 2009. *Electro Static Discharge: Understand, Simulate, and Fix ESD Problems*. John Wiley & Suns, Hoboken, New Jersey.

McKinney, W., 2011. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, Seattle, USA, pp. 1-9.

Meiller, D., 2017: Using Interactive Visual Analytics to analyze influences of climate on industrial production. *In International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing 2017 and Big Data Analytics, Data Mining and Computational Intelligence 2017*. Lisbon, Portugal, pp. 357-358.

Pedregosa, F. et al, 2011. Scikit-learn: Machine learning in Python. *In Journal of machine learning research*, *12*(Oct), pp. 2825-2830.

Radhika, Y, and Shashi, M., 2009. Atmospheric Temperature Prediction using Support Vector Machines. *In International Journal of Computer Theory and Engineering*, Vol. 1, No. 1, pp. 55-58.

Rao, T., Rajasekhar, N., and Rajinikanth, T. V., 2012. An efficient approach for weather forecasting using support vector machines. *In International Conference on Computer Technology and Science (ICCTS) IPCSIT*, Vol. 47, pp. 208-212.

Schafer, R. W., 2011. What is a Savitzky-Golay filter? *IEEE Signal processing magazine*, *28*(4), pp. 111-117.