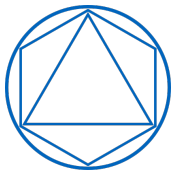
Technical University of Munich

Department of Mathematics



Project Report

Energy peak load prediction in a BMW plant

Reviewer: Prof. Dr. Massimo Fornasier
Chair for Applied and Numerical Analysis
Department of Mathematics
Technical University of Munich

Advisor: Dr. Ricardo Acevedo Cabra, Vadim Gorski

BMW advisors: Dr. Benny Kneissl
Jonathan-Edwin Asamoah
Dominik Schniertshauer

Composed by: Manuel Herrmann
Polina Labintceva
Daniela Simancas Mateus
Frank Vogeltanz
Burak Yazganoglu

Submitted on:

Contents

1 Introduction

This project is carried out for the Data Innovation Lab WS 2017/18 in cooperation with BMW Group. The aim of the project is to analyze and explain the energy consumption in one of BMW Group's plants, the Dingolfing plant, and understand its behavior. In addition, peak identification, consumption pattern recognition and overall energy efficiency have been the main themes around which the project is oriented. In this regard, three different data sets (namely ENERGO, IPS-T and IPS-L, which will be introduced in detail later in this report) belonging to several parts of the plant have been analyzed. After a lengthy phase of descriptive statistics and analysis of the data, predictive models were built on the three data sets. These range from basic analytic models to more sophisticated machine learning algorithms, depending on the need.
Introduction is followed by a more detailed description of the project and the setting. Later, detailed descriptions of the aforementioned data sets are presented. For each data set, respective descriptive analyses, as well as models and prediction methods are introduced. The evaluation of each method can be found at the end of respective sections. At the end of the report, a summary and outlook of the project can be found. Please refer to the Appendix for further plots, graphs and tables.

2 Motivation: Energy Transition and High Peak Load Windows

The demand for electricity in Germany varies over time. It shows peaks during the day and, in general, seasonal effects, as well. In the time before and after the standard working hours, especially private households, demand more electricity. Today, during the energy transition, the supply of electricity is also unpredictable [3]. Until now, there is no practical and low-cost way to store electricity available everywhere, which could compensate the gap between high electricity demand and low supply and vice-versa [6].
Customers with an steady high demand for electricity, who are able to shift their demand of electricity can benefit from an electricity costs reduction. For this purpose, they have to significantly decrease their demand in the so-called high peak load window (HPLW)[1]. This is the time of the highest average electricity demand of customers during a day. The HPLW are calculated and fixed for every season and they have to be taken into account on every working day (except Brückentage, i.e. long weekends and during the time from Christmas to New Year).
The goal of this project, to explore electricity data and uncover opportunities to leverage electricity for cost optimization, is motivated by the following regulation for BMW: its highest demand for electricity in a year within the HPLW has to be smaller than 90%[2] of the highest demand for electricity outside the HPLW of the year. The demand for electricity is always measured within an interval of 15 minutes. The following tables show the exact time slots of high peak load windows for BMW in 2016 and 2017.

Year	Month	HPLW-1 Start	HPLW-1 End	HPLW-2 Start	HPLW-2 End
2016	01	04:15 pm	07:14 pm		
	02	04:15 pm	07:14 pm		
	12	04:15 pm	07:14 pm		
2017	01	07:30 am	08:59 am	05:00 pm	07:29 pm
	02	07:30 am	08:59 am	05:00 pm	07:29 pm
	09	04:15 pm	07:14 pm		
	10	04:15 pm	07:14 pm		
	11	04:15 pm	07:14 pm		
	12	07:30 am	08:59 am	05:00 pm	07:29 pm

Table 1: Start and End of the high peak load windows in 2016 and 2017

[1] This field underlies the legal restriction "Sonderformen der Netznutzung gem. §19 StromNEV".
[2] Erheblichkeitsschwelle/ critical threshold: 10%

3 Organization of the Project and the Dingolfing Plant

Organization of the Project The project was conducted by a team of 5 TUM students, 3 data scientists and an energy manager from BMW, 2 representatives of the TUM Data Innovation Lab and experts from the Leibniz-Rechenzentrum (LRZ) over a time period of 4 months. Due to the significant number of stakeholders involved, a structured working approach formed the prerequisite for a successful progress of the project.

Upon receiving the first data set (ENERGO), technical issues needed to be resolved in the first place to enable access to the data. Since all data sets coming from BMW contain confidential information, safe storage was crucial. LRZ offers safe and powerful infrastructure and at the same time holds partnership with TUM. LRZ's infrastructure was put into disposition for this project by the Data Innovation Lab, and was hence used to store ENERGO and all following data sets provided by BMW. LRZ also offers web interfaces with RStudio server and computing clusters which were used for programming tasks. Furthermore, the LRZ supported the project with their technical support and their expertise in machine learning.

After a brief exploration stage with the first data set, 4 milestones were defined in agreement with BMW:

1. Statistical exploration of energy data and outlier correction
2. Detection of relevant features to explain and forecast the energy profile
3. Using machine learning methods for data exploration and developing prospective improvements
4. Documentation of project

To get an understanding of the technical process, a plant visit in Dingolfing was organized by BMW. After receiving two further data sets, IPS-L and IPS-T, the student team was regrouped in two subteams so that each group could work in parallel on one data set. To ensure steady progress and discuss open questions, independent weekly group meetings were held by the students as well as by students together with the cooperation partners of BMW. Representatives of the Data Innovation Lab were present during several student meetings to support the project with their mathematical expertise.

The Dingolfing Plant The Dingolfing production plant is the largest production site of the BMW group in Europe: approximately 1,600 BMW vehicles roll off the assembly lines every working day. The plant manufactures a wide range of cars: models of the 3 to 7 Series, as well as components for BMW's electric vehicles and car bodies for Rolls-Royce Motor Cars. In total, the location has a workforce of over 17,500 people, plus 800 apprentices. Starting in 2018, the new 8 series and from 2021 on, the completely new, electric and autonomous BMW iNEXT will be manufactured in Dingolfing, which will highly increase the importance of this plant [1] [5].
The plant is spread over different production halls. In the paint shop all the painting is done by robots. Quality control and surveillance of the robots can only be done by humans. In the assembly hall workers and robots work directly together. The shifts can differ in each part of the plant and also work on Saturday is possible. As more than 85% of the produced cars are individual orders and are already sold, the amount of production is extremely flexible and working hours will be increased when the demand increases, as well. When a car is ordered, the customer receives a date, which is called freeze day. This date is exactly 7 days before the production of the individual car will start and until the freeze day the customer has the possibility to change his customization of the individual car. This highly flexible manufacturing is made possible by just-in-time manufacturing.
Energy management is extremely important, as the whole production depends on the supply of energy, not just as electricity, but also as compressed air or heated water. To have a clearer understanding: the energy demand from Dingolfing plant can be compared to the one of a city with 250,000 inhabitants. Power blackouts of as short as 500ms can already disrupt the production process and lead to severe consequences. For these situations, Dingolfing is equipped with two diesel generators which then supply

electricity for emergency components. The whole production will be stopped and just emergency lights and equipment will be powered for approximately 30 min such that the plant can safely be evacuated. Still, the German power grid is very reliable compared to other countries, so in the last 30 years there were only four total blackouts in the plant Dingolfing.

4 ENERGO

4.1 Data Exploration and Electricity Supply

The electricity flow at Dingolfing is structured in a hierarchical way (cf. appendix, figure 43); the three power stations at the top are connected to the medium-voltage power grid (power station 1,2 & 3). Followed by 17 medium voltage stations (MS) and 229 focus stations (FS). Every MS is assigned to one specific part of the BMW plant in Dingolfing.
Each of these 249 stations has a unique identifier, the STDID. ENERGO data contains the energy demand of the plant from January 2016 until November 2017. It contains one measurement for each STDID every 15 minutes with its aggregated energy demand within this time span[3]. In this project only energy in form of electricity is considered. In total ENERGO provides over 77m unique measurements and the data has a size of more than 6 GB. The energy consumption of the plant is measured multiple times, in every hierarchical level once. However, there are levels in which not the whole energy consumption is recorded. This is due to the lack of measurement sensors, especially in the level of the focus stations. The amount of measured electricity in the MS level is nearly complete.
Another interesting aspect in terms of energy management is the availability of internal electricity sources. If there is no possibility of shifting the electricity consumption during the HPLW, the plant can use internally produced electricity to reduce the overall energy demand from external sources. At the BMW plant Dingolfing, there are two different internal electricity sources, photovoltaic systems (used for MS-9 (Assembly) and MS-12 (Dynamic center)) and combined heat and power systems (CHPS). A new system produces electricity with two equal generators in MS-10 which is used in MS-5 (paint shop, cooling water, cooling, extreme cooling) and an older system supplies electricity for MS-2 (hardening shop).
The electricity demand of MS-5 can almost be covered by the internal electricity provided by MS-10. Nevertheless, there is also the possibility to use external electricity provided by the power station, which is illustrated in Figure 1. The external electricity supply is used to cover peak loads, but there are some down times of the CHPS, as well, which can be seen in figure 45, in the appendix. In 2016 there were basically two major down times. One in march, where problems with the Technical Inspection Authority (TÜV) occurred and in summer, when the temperatures were too high and the CHPS could not be used.
Now, analyzing again the overall distribution of electricity demand in Figure 2, it is clear that the electricity is not equally distributed among the MS. Every MS demands electricity, except MS-10, which provides electricity.

Using the information about internal and external electricity sources, the overall electricity which had to be bought by BMW can be estimated. The estimate for January 2016 until October 2017 is 720.93 GW, as illustrated in figure 3. Figure 4 shows the electricity demand per day. The smaller electricity demand during weekends can be seen as well as a decrease during the Christmas season. The EDA can also be used to visualize on which day of the week there is production. Summing the electricity demand for all MS (except MS-10) within 15 minutes and grouping them by weekday, shows the distribution as illustrated in figure 5.

The sum of consumed electricity of the MS within 15 minutes is on average the same during working days and the weekend, respectively. But each day, the interquartile range differs. Another aspect is, that in the box plot as outliers visualized measurements are not necessarily false values; e.g. one sees

[3]The value from ENERGO is given in kilowatts per hour (kWh), therefore to calculate the energy consumption within 15 minutes, the value from ENERGO has to be divided by 4.

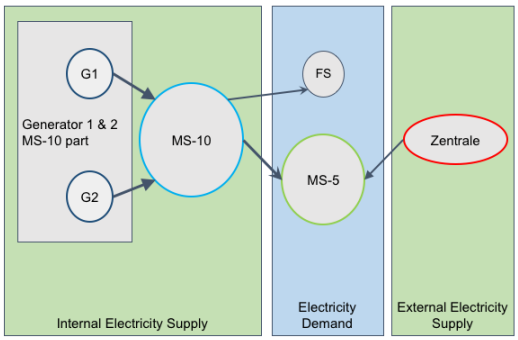


Figure 1: Electricity Supply of MS-5

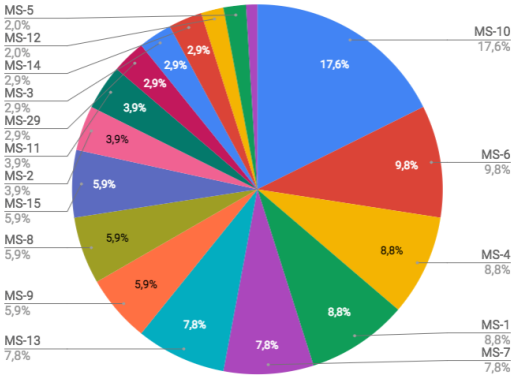


Figure 2: Distribution of Electricity Demand

[GW]	2016	01/2017 - 10/2017	01/2016 - 10/2017
HPLW	10.31	16.23	26.54
Not HPLW	397.28	297.11	694.39
Total	407.59	313.34	720.93

Figure 3: Estimate of total energy demand

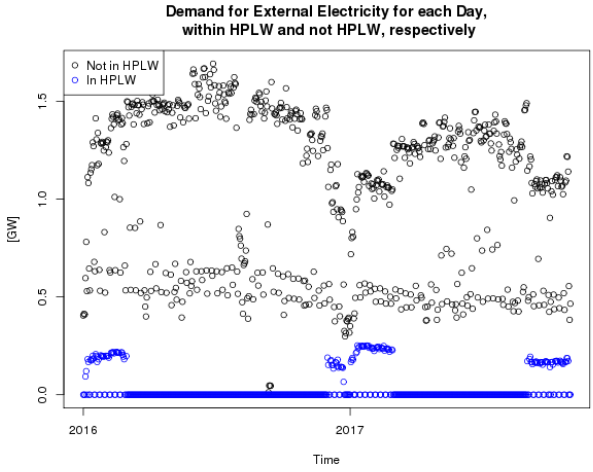


Figure 4: Electricity demand per day

the low data points during the Christmas season, where the production is stopped. As well, before and after Christmas the production was already reduced and at these days also the HPLW is active. It can also be seen that, the average electricity demand during the HPLW is not significantly smaller than outside of the HPLW. In terms of cost reduction, it is particularly important that the maximum electricity demand during the HPLW is at least 10% smaller than the maximum during outside of the HPLW, as explained in Chapter 2.

4.2 Outlier detection

For the project, the most important part of the energy profile is an anomaly-high (peak) consumption. However, not every peak in the data were true measurements. Therefore, it was needed to set apart true anomalies from wrong measurements.
During the analysis of ENERGO data, several types of wrong measurements were detected. The first group includes impossibly huge values which may be followed by negative ones for compensation. The second group consists of "too long" and "too short" measurements. If an error(e.g. wifi connection problem) has occurred in the equipment, the energy consumption could be counted for longer than 15 minutes interval. For energy balancing, after this long measurement a short one usually follows. For this reason, there would be consecutive wrong values. The third group contains missed data. This category contains all lost values equal to -1 as well as some "fake" zero measurements.
All these incorrect values caused by mistakes in measuring equipment. However, they all have a slightly different behavior, therefore all groups should be treated separately.

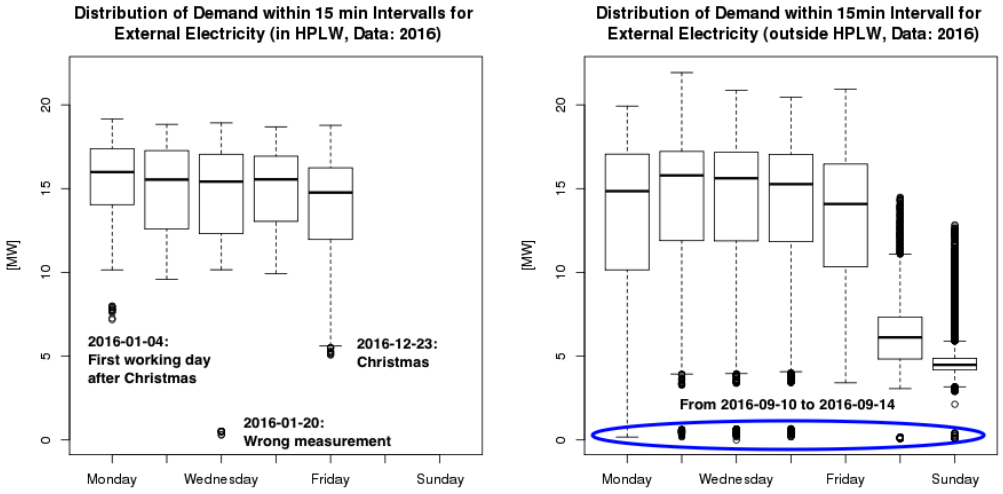


Figure 5: Distribution per Weekday.

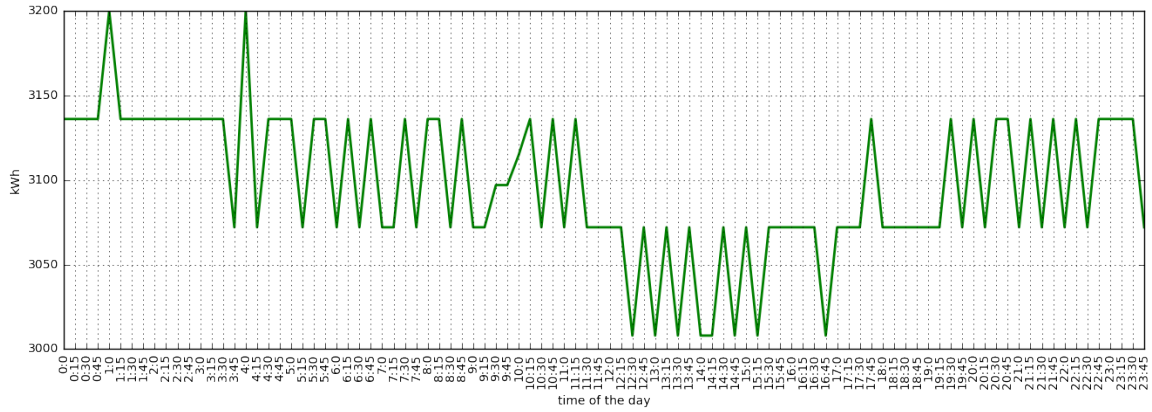
4.2.1 Missed Measurements

The energy demand of the whole plant decreases significantly during weekends and public holidays. The production on the plant is mostly stopped during these days. But some stations are still working, or have a regular maintenance activity. All non-production days are preceded by a smooth decrease in consumed energy, which is followed by a smooth increase during the day after.
In the ENERGO data, zero consumption does not only occur on weekends. There can be a single zero in the middle of the day, which is unrealistic from the production side. Also, there can be a set of consecutive zeros followed by a high value, which has a function of covering "fake" non-production time. All these wrong zero measurements were considered as a missed data together with the -1 values. There were around 1% of missed values and 6.5% of wrong zeros in the data set.
Negative values might affect basic statistics of the data and incorrect high measurements can give a wrong representation of peak loads distribution, therefore, it was needed to replace them. The first idea was to check the level below Medium Voltage Stations. As one MS connected to several FS, it can be possible to collect the data of the missed time interval on the FS level, then sum it up and receive the real consumption on the medium level. However, on the FS level, error in a measuring equipment occurred at the same time as on MS.
The second idea refers to the production cycle on the plant. If some working steps are always made on the same weekday, it is possible to use mean energy consumption of identical days (e. g. all Fridays of the previous month) to find the missed observations. A deeper analysis of energy profiles showed that the production changes from day to day. Therefore, it doesn't make sense to use the energy of the previous days as a base for replacement.
In order to eliminate -1's and wrong zeros without changing the distribution properties, it was decided to replace them with mean values. Single missed values were replaced with the average of the previous and following measurements. The consecutive missed values were replaced with a mean of the day. The figure below represents the example of how this replacement works.

4.2.2 Too long/too short measurements

As per request of the company, the outliers in the data, that were the result of a malfunction in the measurement equipment or wireless connection failure in parts or all of the plant were investigated. This particular kind of malfunctions caused the equipment to measure for more than 15 minutes. When the connection was restored, or the equipment realized that the previous measurement covered more than 15 minutes, the consecutive measurement is programmed to compensate for the elongated duration of the previous period, by covering a shorter period of time. Therefore, the expectation of the sum of two consecutive "too long" and "too short" measurements are the same. These errors reveal

(a) Wrong non-production time



(b) Arter replacement

Figure 6: Wrong measurements example

themselves by a pattern of a peak followed by a trough. In Figure 7 , we see an obvious example of this kind.

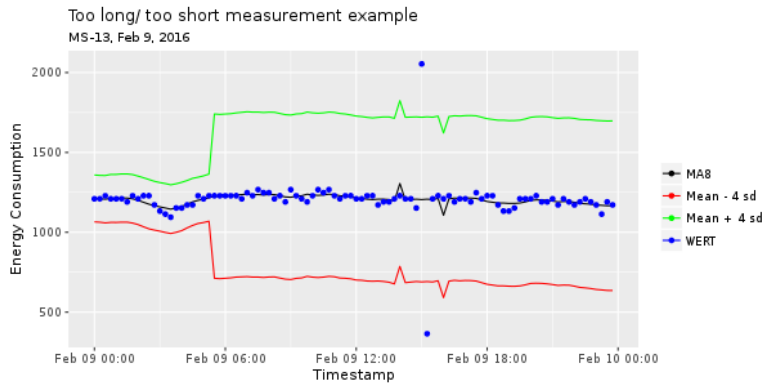


Figure 7: Too long/too short measurement example

It may sound easy to treat these measurements as any other outlier. However, their particular character of a peak followed by a trough makes it easier to isolate them. Therefore, in order to spot these errors, 4 columns for generated for every observation:

1. A column indicating if the observation was 3 standard deviations above the Moving Average
2. A column indicating if the observation following was 3 standard deviations above the Moving Average

3. A column indicating if the observation was 3 standard deviations below the Moving Average
4. A column indicating if the observation preceding was 3 standard deviations below the Moving Average

A pair of observations were marked as too long/ too short measurements, only if the too long measurement was 3 standard deviations above the Moving Average, the observation preceding the too long observation was not 3 standard deviations above or below the Moving Average, the too short measurement was 3 standard deviations below the Moving Average, and the observation following the too long observation was not 3 standard deviations above or below the Moving Average. As one of the main objectives of the project was to identify peak energy consumption, a more conservative attitude was adopted in detection of outliers and errors.
This conservative approach was also influencing in the parameter tuning. Moving Average parameter was set to 8, as a number smaller caused the average to be too volatile and a number larger was not adaptive enough to the changes in consumption, especially during shift changes. The parameter of the standard deviations were tested from 2 to 5, in decimal steps. The biggest change was in the outliers of MS-50, due to the spiky character of the entire time series. In order to make sure no real consumption data was not smoothed, the parameter around which the smallest change in the number of outliers was observed was selected, namely 4. In total, 39 pairs were marked, 17 of which belonged to MS-50.
A couple of ideas were introduced on how to treat these errors: Using Focus Stations to check if they represent the actual behavior without the failure, fitting a linear model to predict the observations preceding and following the pair and averaging the pair. Focus Stations displayed the same pattern as the Medium Voltage Stations, as the cause of the problem was local, rather than hierarchical. In order not to lose the information from the observations themselves, they were averaged.

4.2.3 Wrong Outliers

As already described in 4.2.1 and 4.2.2, the ENERGO data set contains some extreme outliers, which are not caused by energy peaks but by wrong measurements. To find and correct these “untrue ”peaks, it was mandatory to create one prepared data set.

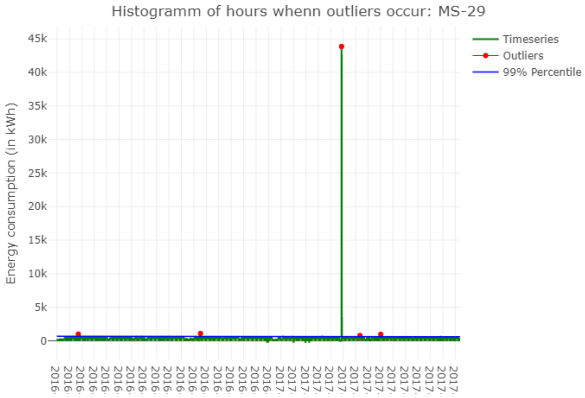


Figure 8: Marking too many Peaks as Outliers in MS-29 (Std Dev 3).

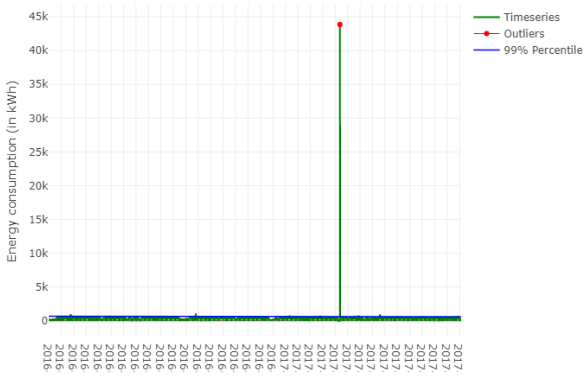


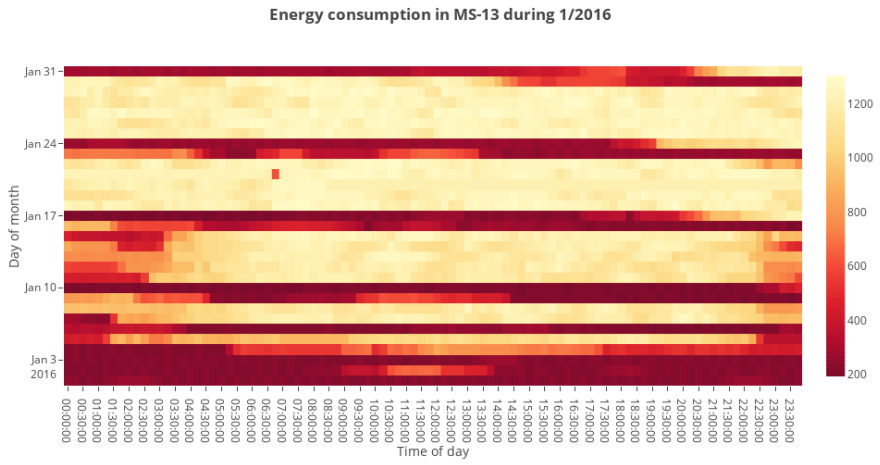
Figure 9: Marking only untrue Peaks as Outliers in MS-29 (Std Dev 9).

For the detection of outliers, the standard deviation method was used as a general idea. The first approach marked all energy values as outliers, when the difference between two consecutive values surpass the threshold of 3 standard deviations. Because the energy consumption varies strongly over a year, the standard deviation is not calculated over the whole data set but as a running standard deviation. Hereby, the running window consisted of 48 values (representing half a day) before and after the estimated value.

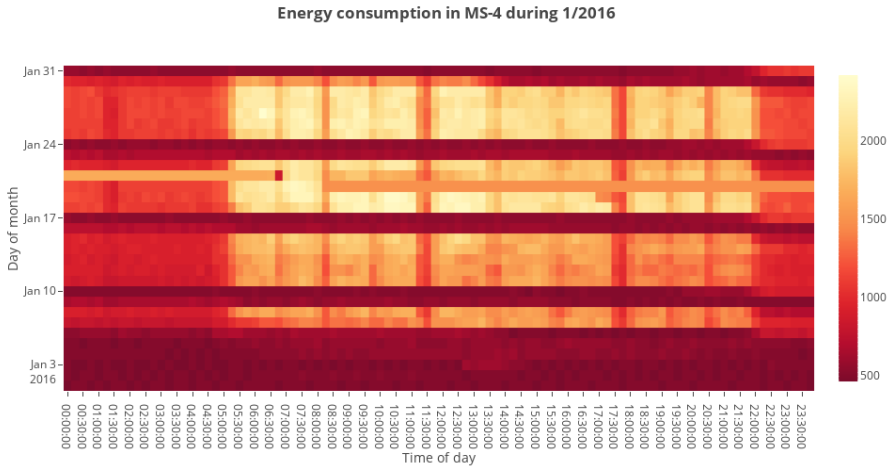
Additionally, to prevent too much smoothing of the data set by falsely correcting even properly measured energy peaks, only values surpassing the 99th-percentile of the energy values over two months are considered as possible outliers. So, only extremely high values can be detected. As can be seen in Figure 8, this approach still marked correct peaks as outliers, though.
By optimization, it was found that the optimal benchmark for marking values as outliers, is a threshold of 9 standard deviations (figure 9).
The so found outliers were then corrected by calculating the average of their previous and subsequent value. Besides, the data set also contained some negative values. Since negative energy consumption is unrealistic, these data points are set to zero.

4.3 Visualization

Proper use of visualization tools is fundamental for understanding of trends and behavior of the data. As a team who had never worked with energy or BMW plant data, it was crucial to have a general view of what happens during a day/month/year in a BMW production plant.



(a) MS-13



(b) MS-4

Figure 10: Energy consumption in two different stations during January of 2016. Shift changes, weekends and holidays are easily spotted with a heat map. Low and high energy pikes are also evident in this graph.

Such visual analysis is important to understand and detect patterns of certain areas in the plant. For example, from **Figure 10** it can be inferred that MS-13 station, Lackiererei, (see graph with stations description 43), has a more steady work load throughout the day compared to MS-4, Karosseriebau,

where there is a large work reduction during night time.

4.4 Correlation Between Focus Stations

The electricity flow in the plant Dingolfing is hierarchical. Each Medium Stations (MS) is split into several Focus Stations (FS), and almost always one individual FS can be assigned to one MS. There is a great number of FS which means, that there are many real measurement points in the plant. Now the question arises, how the FS are correlated to each other. If the dependence structure is known, one could e.g. use the data to identify measurement errors or if the sensors are very reliable one could decrease the amount of needed measurement points and could use the dependence structure to virtually calculate the electricity demand of the correlated FS.
When all electricity consumption is measured correctly, the electricity demand of the MS should equal the sum of the electricity demand of the FS, as there is a hierarchical structure. Therefore the relationship between the FS and MS can be assumed to be linear and the sample Pearson correlation coefficient can be used as a measure of dependence.

Demand Nov 2017 [kW]	MS-13 STDID	449	447	482	451	448	452	450
463.424	449	1	1	1	0,98	1	0,98	0,98
459.368	447	1	1	1	0,98	1	0,98	0,98
441.361	482	1	1	1	0,98	1	0,98	0,98
374.148	451	0,98	0,98	0,98	1	0,98	1	1
457.056	448	1	1	1	0,98	1	0,98	0,98
358.836	452	0,98	0,98	0,98	1	0,98	1	1
270.356	450	0,98	0,98	0,98	1	0,98	1	1

MS-13.

Demand Nov 2017 [kW]	MS-3 STDID	119	140	4915	4926	138	4932	136	4928	134	120	141	4917	139	137	131	4919	4913
66.520	119	1	-0,34	0,27	0,6	0,01	0,68	0,55	0,54	0,23	0,71	0,02	0,6	0,37	0,49	0,1	0,63	0,52
19.887	140	-0,34	1	0,2	-0,15	0,01	-0,22	-0,21	-0,17	0,17	-0,24	0,05	-0,25	-0,05	0	0	-0,09	-0,05
81.755	4915	0,27	0,2	1	0,4	0,01	0,4	0,37	0,4	0,24	0,25	0,09	0,36	0,25	0,33	0,08	0,54	0,49
72.900	4926	0,6	-0,15	0,4	1	0,01	0,91	0,84	0,8	0,32	0,56	0,02	0,73	0,44	0,47	0,11	0,86	0,63
260	138	0,01	0,01	0,01	0,01	1	-0,01	0,01	0,01	0,03	-0,01	0,01	0	0,02	0,02	0,03	0,01	0,02
78.290	4932	0,68	-0,22	0,4	0,91	-0,01	1	0,84	0,81	0,31	0,63	0,02	0,78	0,43	0,5	0,1	0,88	0,65
56.248	136	0,55	-0,21	0,37	0,84	0,01	0,84	1	0,74	0,28	0,52	0,02	0,69	0,38	0,41	0,09	0,81	0,62
73.035	4928	0,54	-0,17	0,4	0,8	0,01	0,81	0,74	1	0,29	0,53	0,03	0,7	0,42	0,45	0,09	0,81	0,63
81.613	134	0,23	0,17	0,24	0,32	0,03	0,31	0,28	0,29	1	0,21	0,09	0,26	0,21	0,26	0,04	0,34	0,28
71.285	120	0,71	-0,24	0,25	0,56	-0,01	0,63	0,52	0,53	0,21	1	0,03	0,54	0,3	0,41	0,09	0,57	0,47
-	141	0,02	0,05	0,09	0,02	0,01	0,02	0,02	0,03	0,09	0,03	1	0,02	0,09	0,08	0,01	0,04	0,06
46.649	4917	0,6	-0,25	0,36	0,73	0	0,78	0,69	0,7	0,26	0,54	0,02	1	0,36	0,41	0,08	0,77	0,61
75.940	139	0,37	-0,05	0,25	0,44	0,02	0,43	0,38	0,42	0,21	0,3	0,09	0,36	1	0,37	0,08	0,44	0,34
76.560	137	0,49	0	0,33	0,47	0,02	0,5	0,41	0,45	0,26	0,41	0,08	0,41	0,37	1	0,09	0,51	0,43
9.293	131	0,1	0	0,08	0,11	0,03	0,1	0,09	0,09	0,04	0,09	0,01	0,08	0,08	0,09	1	0,11	0,09
89.625	4919	0,63	-0,09	0,54	0,86	0,01	0,88	0,81	0,81	0,34	0,57	0,04	0,77	0,44	0,51	0,11	1	0,73
99.130	4913	0,52	-0,05	0,49	0,63	0,02	0,65	0,62	0,63	0,28	0,47	0,06	0,61	0,34	0,43	0,09	0,73	1

MS-3.

Figure 11: Correlation between FS of MS-13 and MS-3 in November 2016.

Analyzing the correlation of the FS for each MS in December 2017, the result can basically be grouped in 3 categories: a) very high correlation of its FS, b) most FS have a low correlation (around 0-0.3) and only a few FS have a higher correlation (0.7-0.8) and c) few correlated FS (0.4-0.7) and many highly correlated FS (0.9 - 1). Interestingly, mostly MS for non-production parts of the plant (e.g. MS-3 BIZ and kitchen) fall in category b). MS-13 shows a very high correlation among its FS, so it is in category a). For both MS the correlation of its FS is compared in figure 11.
The plots show, that indeed in some MS, the FS are highly correlated and in other MS at least a subset of FS show a high correlation. On the other hand, not every MS shows this pattern and one would have to know in advance which FS are correlated and thus measure sensors could be removed. Another aspect one should consider is, that all MS, except MS-13, show low correlated FS, the longer the observed time intervals are. So analyzing e.g. MS-3 for January 2016 to December 2017, shows that only a few FS are correlated between 0.8-0.9.

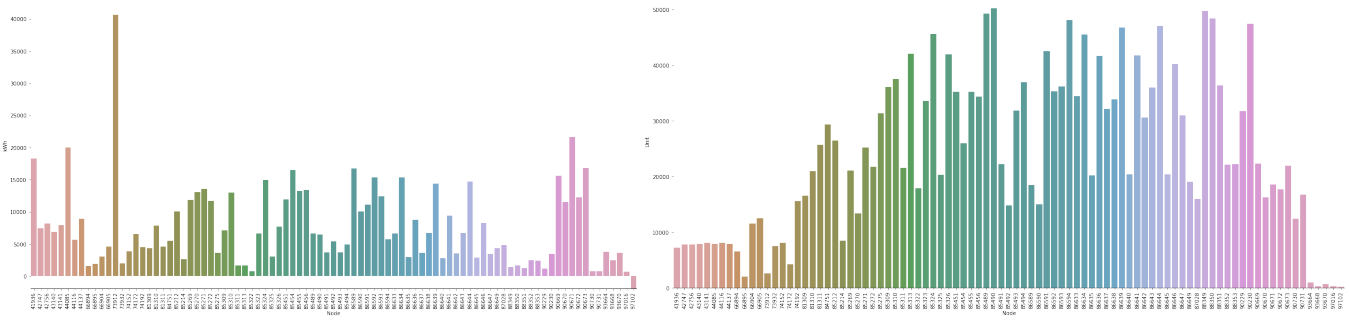
4.5 Findings and Conclusion

Analysis of ENERGO data set was the first step in the understanding of how this complex BMW plant works and how energy is distributed over it. The main interest for the analysis was the MS level. Using this data the energy profile of each production station can be investigated. By suitable visualization of the data, the main patterns in production schedule were identified and most energy consuming stations were found.
In order to avoid wrong statistical interpretation of the data, the real peaks were separated from the wrong values, caused by mistakes in measuring equipment. Wrong values were replaced appropriately. For the whole picture of energy consumption, internal energy resources of the plant were analyzed.

Also, the correlation between MS and FS levels was estimated. All these findings are useful for further analysis of more specific production data (IPS-T, IPS-L), for the building of prediction models, and for the processes/production optimization.

5 IPS-T Data

The IPS-T data set contains information about production together with energy consumption on MS-stations responsible for car body assembly. The production stations are called *nodes* and each node is in charge of a specific operation for a single car model. There are 89 different nodes in the IPS-T data set. For instance, node 330THL3_.SPS, is the station where rear left door for car model BMW 5 Series (G30) is assembled.
An advantage of IPS-T data set with respect to ENERGO, is that for each node there is information about the energy consumed and the count of car bodies at a certain point in time; while in ENERGO, there is no production data. The main disadvantage of IPS-T data set is that tracking of a specific car around a plant cannot be done, since whenever a car is counted, only its model is stored instead of a unique identifier. Besides, IPS-T covers much smaller time frames: IPS-T started at the end of August 2017, ENERGO started in January of 2016.
Figure below represents the energy consumed by each node together with the number of cars passed during one month. These plots show that not all nodes have a forward dependence between the number of proceeded units and consumed energy.



Energy consumption of each node Number of units passed through the node

Figure 12: Production during one month

Summarizing, IPS-T data set can be used to analyze energy profiles when performing certain tasks on different car models, as well as to identify which nodes are independent of units produced.

5.1 Pre-processing and Exploratory Data Analysis

Raw IPS-T data came in a very long table, where a new row was added every time there was an increase of (roughly) 0.1KW in accumulated energy consumption. At the same time, another row was added every time (roughly) one car was dispatched from a working station, i.e. node.
The first step taken in pre-processing was shortening the table by accumulating all car units and energy measurements in ranges of 15 minutes. Just like in ENERGO data, this lead to 96 energy measurements per day as well as car unit counts. This modification allowed to reduce the size of the data-set, and lead to clearer plots without loss of detail.
Simple scatter plots on every node after the first step showed presence of outliers that were several orders of magnitude away from most of the data. Hence, the second step in pre-processing was outlier detection and removal. Each node's data were treated as samples of separate populations with their own corresponding statistics. The method chosen for outlier detection employed the median and MAD

(median absolute deviation) of each node since they are more robust statistics than mean and standard deviation [2].
In the end, all measurements more than 10 MADs away from the corresponding median were smoothed out by taking the average of the two previous and two posterior measurements.

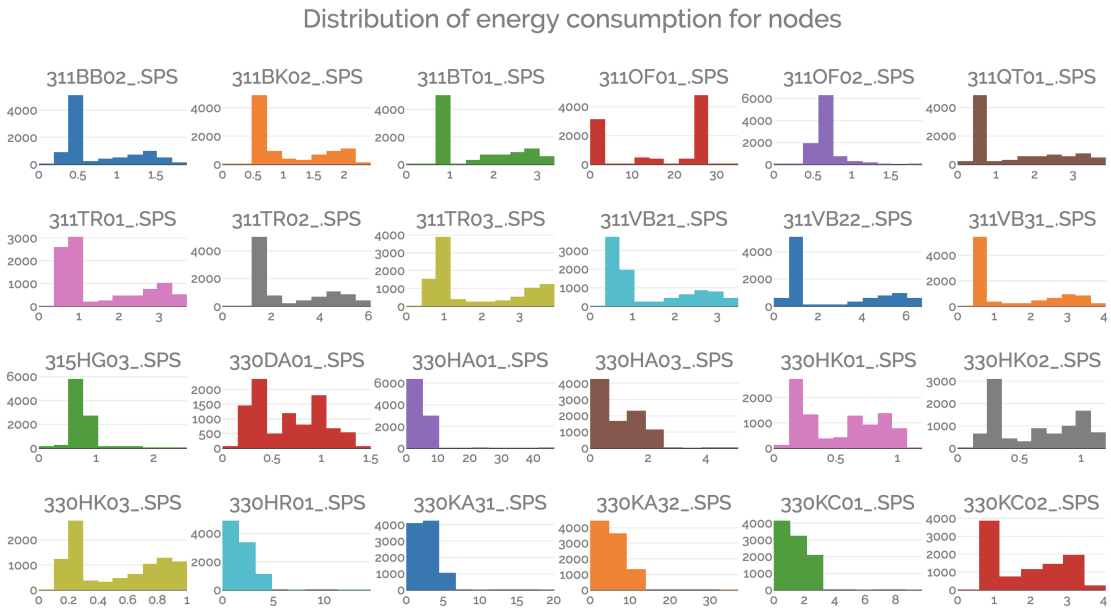


Figure 13: Histograms displaying energy profiles of 24 randomly picked nodes.

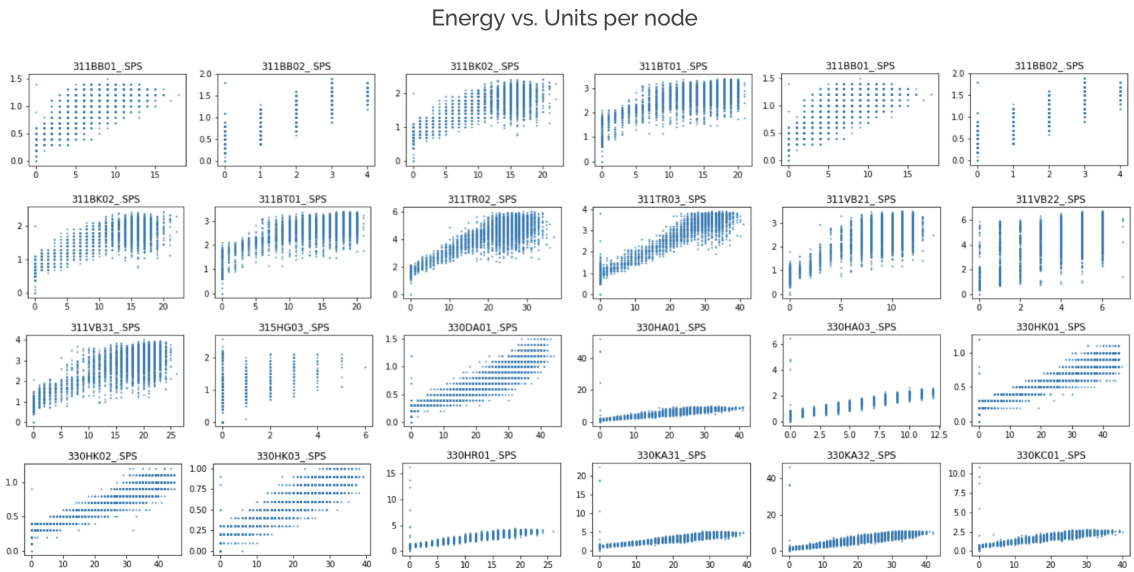


Figure 14: Scatter plots showing the relationship between car units and energy consumed for 24 randomly selected nodes.

Finally, it was possible to observe the energy profile of each node, by analyzing their histograms. A first look at these graphs suggested that nodes are not completely independent from each other and could be clustered by their energy profiles. Additionally, a glimpse at scatter plots of car units vs. energy showed that in many cases, as expected, there is a relationship between the amount of car units passing by a node and the energy consumed by it.
Following this initial analysis, the team decided to apply regression and clustering models on IPS-T

data for further information discovery.

5.2 Linear Models and Feature Selection

Two different implementations of linear regression in R and Python were used. Furthermore, different features combinations were applied on these models to obtain the results.

IPS-T is the time series data, so it is possible to extract a lot of information from date and time. Every day of the week or hour of the day can be coded by the corresponding average of energy which was consumed on this day of the week or hour. Therefore, categorical features such as day of the week were coded by an average value of the target variable. In this case, it is important to ensure that the average value is calculated only within the train data set (or within the currently observed fold in the cross-validation), otherwise, it may bring information about the future to the model.

To estimate the generalization ability of the solution cross-validation was used. Time series data is characterized by the correlation between observations that are near in time. Cross-validation techniques such as `KFold` and `ShuffleSplit` would result in meaningless correlation between training and testing instances on time series data. To avoid it, `TimeSeriesSplit` from `sklearn` can be used to cross-validate time series data samples that are observed at fixed time intervals. This class is a variation of k-fold which returns first k folds as train set and the $(k+1)$ th fold as test set. Unlike standard cross-validation methods, successive training sets are super-sets of those that come before them (Figure 15).

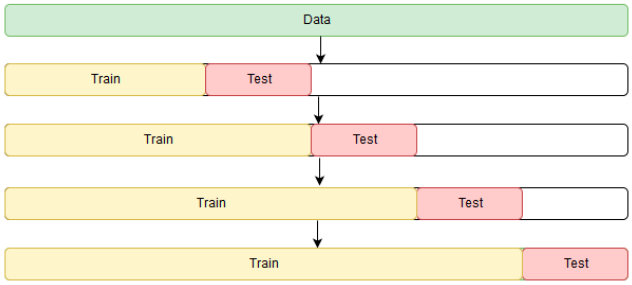


Figure 15: Time Series Cross-Validation

The initial idea is to linearly relate Energy Consumption to the Number of Units processed per 15 minutes. In addition to the number of units passed through the node, the lags of the target variable can be used. In presented model lags from 96 to 193 were used. This means that energy consumption of the previous day for each 15 minutes interval used as a variable. Models were built using `stats` package in R and `sklearn` in Python. Finally, the models are evaluated and compared using Mean Absolute Error and Mean Absolute Percentage Error.

Based on the different feature combinations, three different sets of the parameters were used for linear regression model:

1. Lags of target variable + is weekend + mean of day + mean of hour
2. Number of units + is weekend + mean of day + mean of hour
3. Number of units + lags of target variable + is weekend + mean of day + mean of hour

The time series plots on the figure 16 shows the prediction for one node. The left one represents the forecast without using units as a feature, the right one – with units. Prediction based on the LR model depicted in red, compared to the real values in blue having the number of observation as the x-axis (one observation is one 15 minutes interval), and the energy consumption as y-axis.

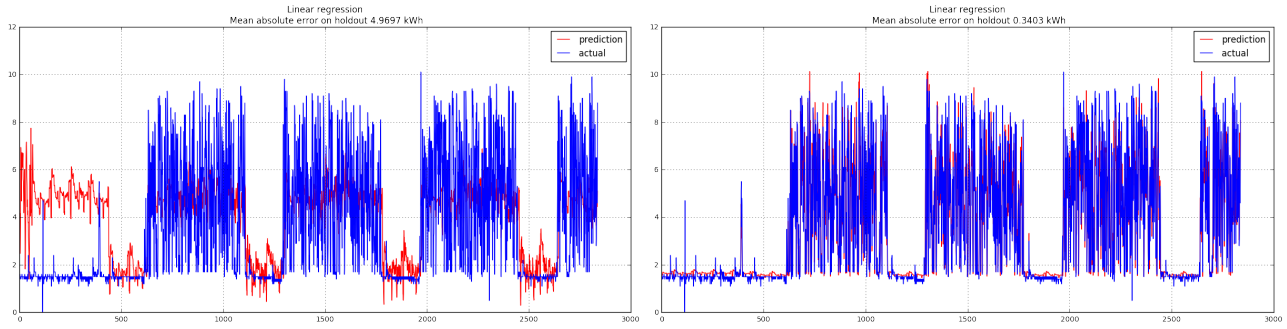


Figure 16: Linear regression model predictions

It seems that the combination of lags and units, as in the third set, should give the best result, but for 75 nodes over 89, it works worse by 1,6% in average than set number two. Table 2 shows the MAE of predictions on holdout part of the data.

Node	Mean Absolute Error		
	Features Set 1	Features Set 2	Features Set 3
330RH23_.SPS	7.9188	0.3102	0.3161
330VB03_.SPS	7.3540	0.2800	0.2868
330TVL2_.SPS	5.5037	0.2310	0.2325
330HA01_.SPS	4.9697	0.3403	0.3658

Table 2: Mean absolute errors of predictions

Afterwards, more production schedule related feature were tested. In order to have a better fit, features Shift and Day of Week were introduced. Most models were improved or not affected by this addition. Out of 89 Nodes, 75% of the nodes had a MAPE lower than 14.1%. Every node except one had a MAE lower than 1 kWh.

5.3 Clustering Nodes

The linear models in most nodes performed very high. Therefore, in order to extract the most information from the data, the 6 nodes with low correlation in the linear models were further investigated (See Figure 17.) One of the low-performing nodes was a dryer in the body shop. From our visit to the plant, and the talks with the company, it has been pointed out that dryers consume energy related to the duration that they are working, and not the number of units they process.

For simplicity, the shifts with no units were taken under consideration, namely the night shifts. For the 98 complete days that there are, every night shift between 22:00 and 06:00 were extracted. Therefore, for every day, there were 32 observations of energy consumption. In particular, the ramp-up behaviors were under scrutiny, the greatest energy consumption drivers through the night.

From this plot 18, it can be observed that there are several clusters of behaviors through the night. These clusters can be obtained via an algorithm that would group similar behaviors through the night together. The data could be handled as 98 observations to be grouped, with 32 dimensions. However, usual clustering methods do not take into account the fact that the data at hand is a time series data, which requires a special understanding of the dimensions. Therefore, it is important to look into Time Series Clustering. Time Series Clustering focuses on the similarity - or dissimilarity - between curves and groups curves together, rather than points in a 32-dimensional space.

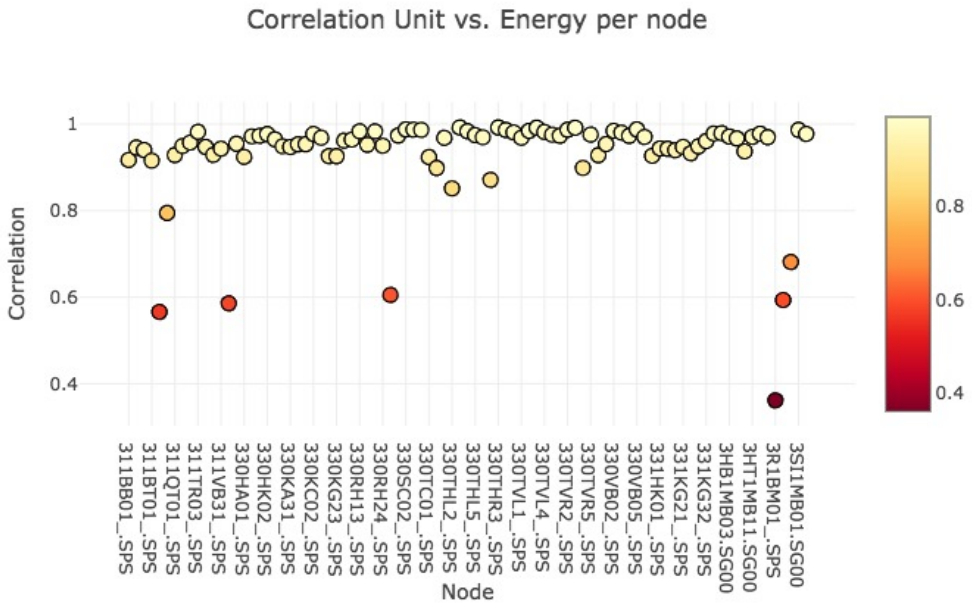


Figure 17: Correlation between Energy Consumption and Units for every node

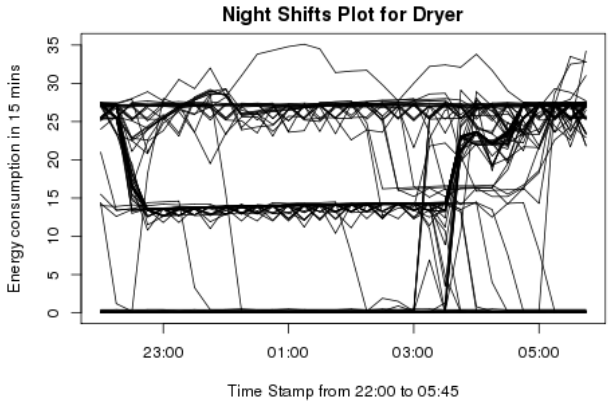


Figure 18: Energy Consumption during 98 nights for the dryer node

Consequently, it was important to determine a suitable dissimilarity measure to feed in the clustering algorithm. Dissimilarity is a general term for a measure of distance between two observations under a certain criterion. Then, using the observations, a dissimilarity matrix is generated. Afterwards, this matrix is fed into a clustering algorithm, resulting in the observations grouped in different clusters. For the Night Shifts Data, 3 different dissimilarity methods were used to generate a dissimilarity matrix.

1. Correlation: Computes dissimilarities based on the estimated Pearsons correlation of two given time series.
2. Frechet Distance: Computes the infimum of maximum distances between two curves
3. Dynamic Time Warping: Computes optimal match between time series regardless of their acceleration. Used prominently in speech recognition.

Each method is used to create dissimilarity matrices. These dissimilarity matrices are then used to create dendograms. The clustering method is chosen as Hierarchical Clustering, as the number of clusters, or the number of observations per cluster is not intuitive from the data. Below, the plot of the dendogram generated by the dissimilarity measure Fréchet Distance can be seen.

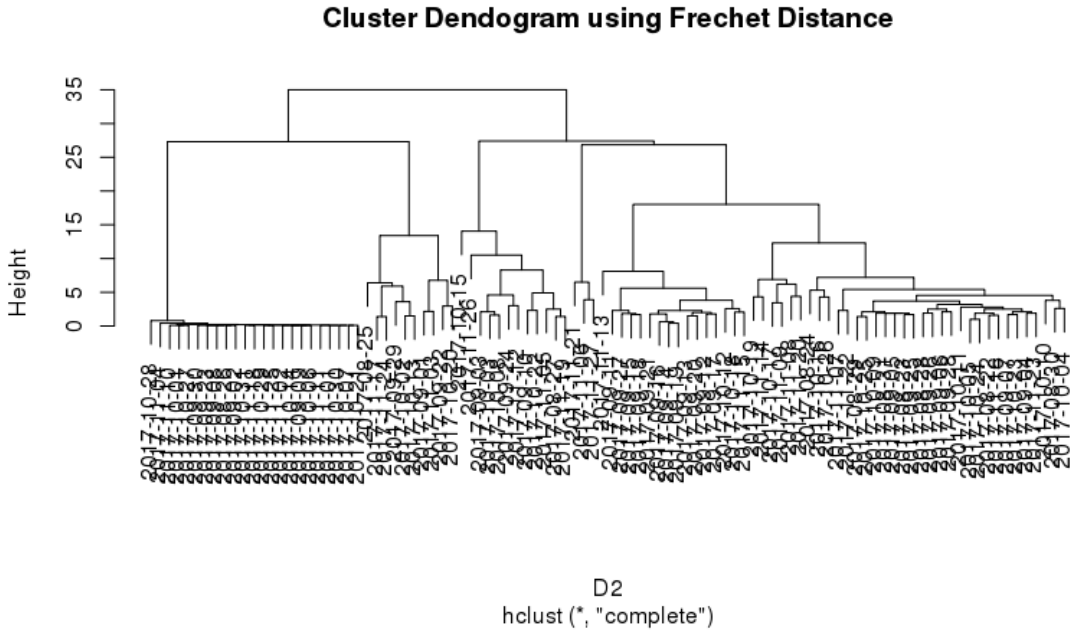


Figure 19: Dendogram using Frechet Distance

For each method, 5 or 6 clusters were the most sensible result. Then, these clusters were plotted separately, to closely examine the consistency of behaviors captured by the clusters. Below plots show the 6 clusters using the Fréchet distance method. They show the different behaviors in the energy consumption, irregardless of where the drops and the ramp-ups start. Taking a close look at the weekdays that are present in each cluster, following conclusions are reached:

Cluster 1: Regular workdays, (Tue - Thu) except any holidays. During these nights, energy consumption starts from a high value, as there was production before this night shift, followed by a drop during no production hours, followed by an increase in consumption due to the beginning shift

Cluster 2: Fridays, or any day that had a production shift before a night and no production on the next day. This group includes a Monday followed by a holiday (3.10.2018), which shows how robust the clustering is to the production

Cluster 3: Days during which the dryer was almost completely shut down. These days are regularly Saturdays, but also include bigger holidays.

Cluster 4: Sundays, or any night that did not have a production shift before the night, and a production shift in the following morning. This cluster is symmetrical to Cluster 2.

Cluster 5: This cluster groups the times when the dryer was on throughout the night. The highest average energy consumption is produced by nodes in this cluster.

Cluster 6: This cluster groups 3 nights that display a somehow different behavior than the others. It would have been more suitable if these observations were also in the first cluster. However, this could also be representing a different behavior that we do not have enough samples due to the short duration of the data.

5.4 Results

Linear models were able to predict energy consumption with low error rate, which is nice for production-solutions because these models are easy to implement and they have interpretable results. Despite the good performance of the linear regression model, it is still possible to improve it by playing with the features. It also will be interesting to try Random Forest and XGBoost and compare the results.

Clustering was used to classify energy profiles of 6 nodes during the night shifts. As a result of this

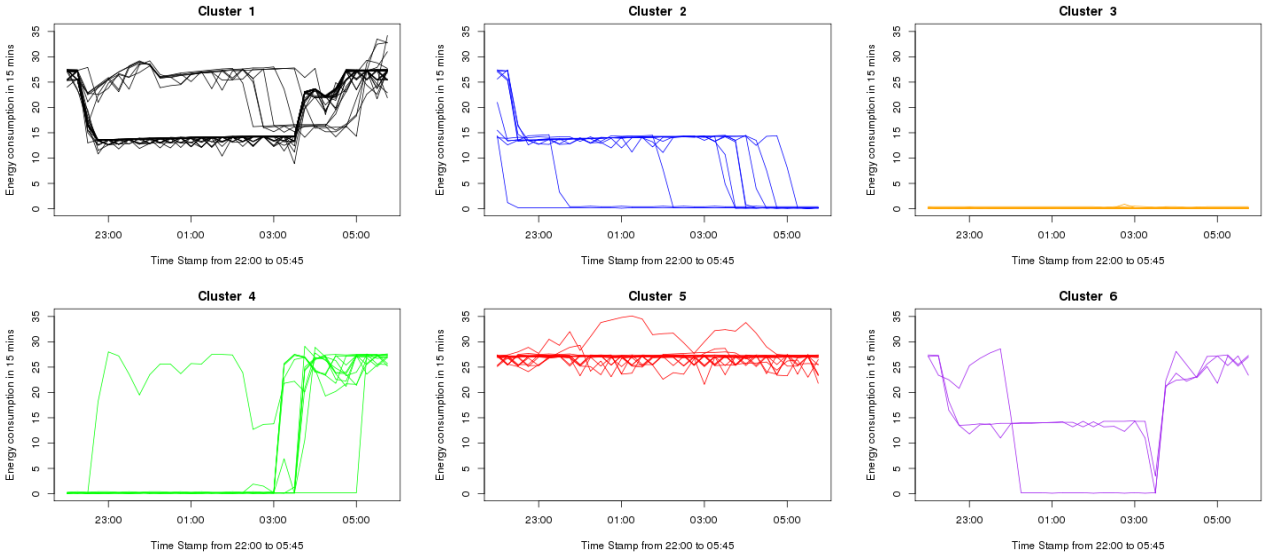


Figure 20: Clusters 1-6 using Frechet Distance

exercise, it was possible to group energy profiles. A main indicator proved to be the day of the week (or presence of a holiday), as ramps ups and downs depend on this factor. It remains to be discussed if, since ramp ups and downs *have* to happen, it is possible to minimize the time between the ramp up and the first car body being processed.

6 IPS-L Data

6.1 Data Exploration and Basic Descriptive Statistics

The data set IPS-L provides the logistic perspective of the cars in the paint shop (MS-13). This means that every time an individual car passes a production station in the paint shop the time, the car, its color, variant and the station itself are tracked. For November and December 2017 the data is complete and for April and October 2017, respectively, only a few cars were tracked.
After removing duplicated measurements there are in total 323,369 unique measurement points from 46,547 individual cars in the IPS-L data. In the data, 102 different colors were used and 982 different body specifications (variants) were produced. The variant of a car is a more specific description than only the car model; It e.g. describes if the car has 3 or 5 doors, etc. 943 variants were tracked only one time in the data and for these cars the painted color is unknown.
The distribution of colors reveals an interesting fact. Even thought a huge number of different colors were used during the observed time period, more than 90% of all cars were painted in one of just 14 colors. The same holds for variant. Over 90% of the cars is one of just 18 different car variants. This shows the high flexibility the plant Dingolfing provides.

In the paint shop there are 36 different working stations. They are identified by a unique abbreviation and its description. In total there are 4 different production lanes, on which cars are painted and dried afterwards. These lanes are connected at the beginning and in the end. All cars start at Fueller Decklack and the production flow ends in the warehouse. Some of the production stations can be categorized into similar groups. For example there are eight working stations which denote the drying process (Auslauf Trockner: Z20554x0). To get a better impression how the stations are connected, graphs, showing the production path, are created in the next step. Later the influence of color or variant on the production flow will be analyzed.

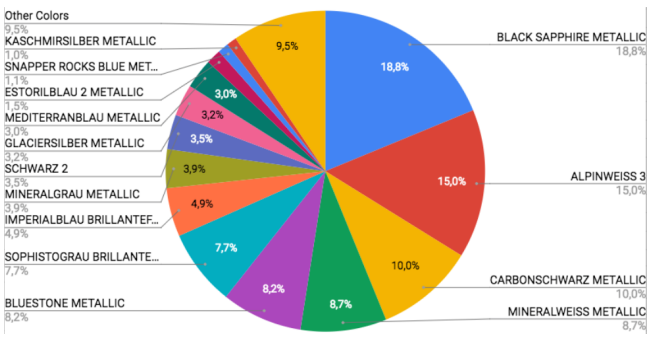


Figure 21: Distribution of Colour

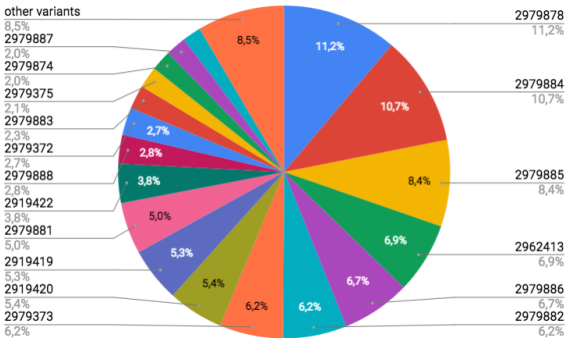
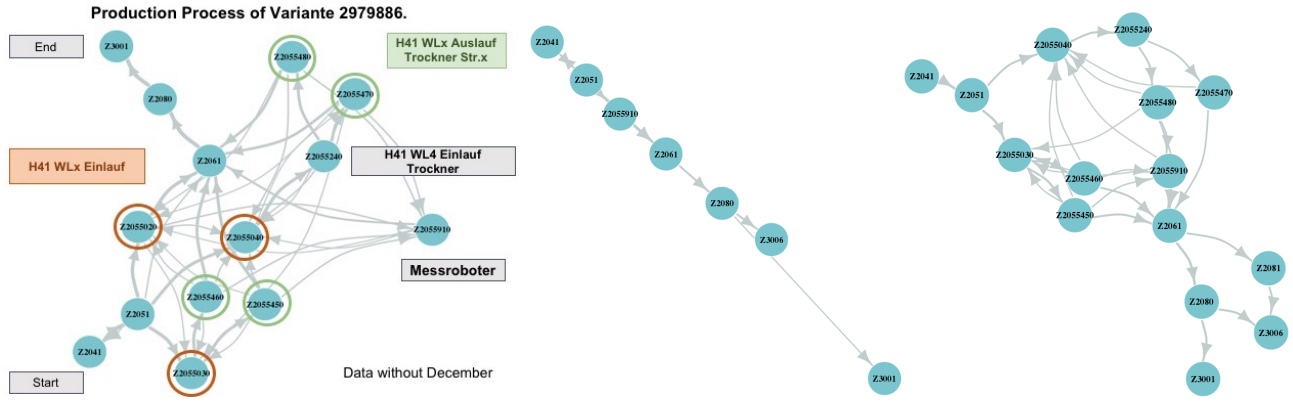


Figure 22: Distribution of Variant

6.2 Visualization

The EDA of the IPS-L Data motivates to visualize the sequence of production stations each individual car takes. The visualization is done using the R package `iGraph` and `NetworkD3`. `iGraph` plots graphs and has graph algorithms implemented. `NetworkD3` creates interactive graphs, where the view of the model can be changed by dragging and dropping vertices of the network. This is especially helpful when many cars are shown in one single graph and their production flow differs.
As mentioned in the last chapter, some working stations can be divided in groups. Now, by comparing the production flow of the different colors and variants, one remarks, that basically the sequence of these groups follows the same pattern. All cars start with Fueller Decklack (Z2041) and the last station will be the warehouse (Z2080, Z2081). This is illustrated in the following example, where the production flow of all 2,425 cars, with the variant 2979886 which were build before December 2017, is analyzed.



PF of Variant 2979886 | PF of Color Spacegrey Metallic | PF of Color Atlaszeder Metallic

Figure 23: Comparison of Different Production Flows

Nevertheless, the production flow of a certain color or variant can highly vary. One reason for that is that e.g. some colors need more coats of lacquer than other colors do. Here, the sequence of the two colors Spacegrey Metallic (50 cars) and Atlantic Cedar Metallic (154 cars) are compared. For the first color only a few stations are used and the Atlantic Cedar Metallic needs many more production stations. One sees as well, that some of the production stations are used several times by one car as e.g. they need more coats of lacquer (e.g. Z2055030).
Another aspect can be visualized. By aggregating the production steps per hour, it can be seen that at the paint shop there is 24 hour production on 5 days per week. As well one can see that the amount

of production is decreased during the night. This corresponds to the information provided during the visit to the plant. At the paint shop, there are three shifts, from 5 am to 1 pm, 1 pm to 9 pm and 9 pm to 5 am. The night shift can clearly be identified using the plots in figure 24. As well one can see the production stop in the week of Nov 1st, as it was public holiday. There is also a sudden decrease in the production on Dec 7, which will be analyzed in chapter 6.4. Additionally, the heat map shows,

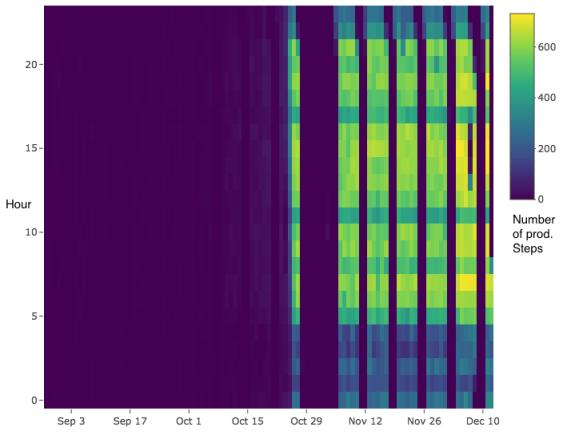


Figure 24: Number of Production Steps within 15 Min

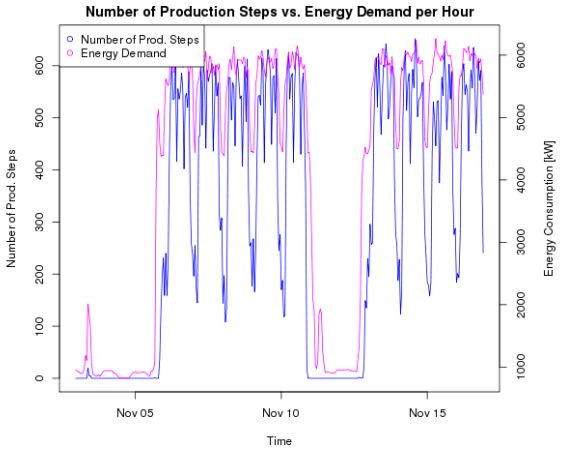


Figure 25: Comparison between Number of Production Steps and used Electricity

that the number of tracked production stations is not constant for every hour of the day which can bee seein in figure 26. It decreases during the night, but there are same regular drops e.g. at noon as well. This motivates to investigate, if some production stations are used more frequently at a specific time point of the day or if it is just random.

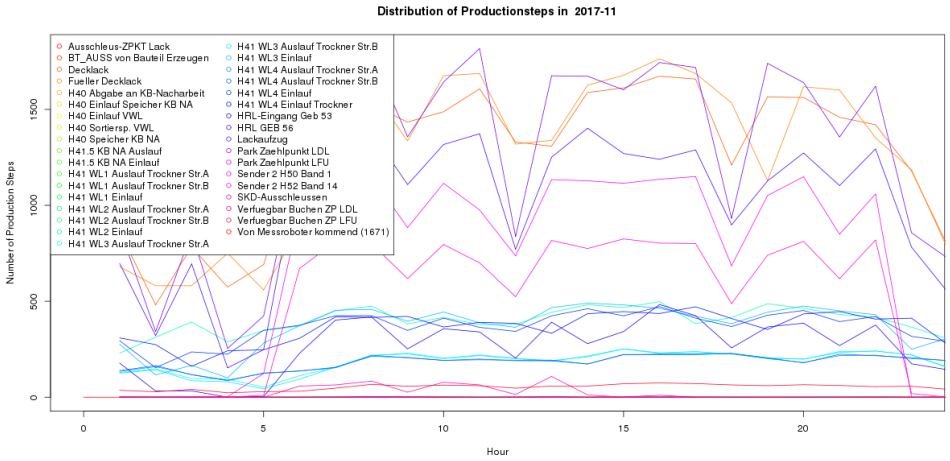


Figure 26: Distribution of Number of Production Steps by Working Station

6.3 Mapping to ENERGO Data

After receiving the IPS-L data set, there were now two data sets available for data exploration:

ENERGO: containing information about the energy consumption in the BMW factory in Dingolfing
IPS-L: containing information about production steps in factory building MS-13 (paint shop)

To meet the goals of detecting correlations between production program and energy consumption as well as to develop models for the prediction of energy consumption, both data sets needed to be merged. Since the time measurements in both data sets differ, data preparation was required.

As described in section 4.1, in ENERGO energy usage is measured as aggregated energy consumption over 15 minutes time intervals. In contrast, the IPS-L data set contains the timestamps when a car crosses a light barrier of certain production step, e.g. top-coat lacquer. To merge both data sets, each timestamp in IPS-L and its associated production step is assigned to the corresponding 15 minutes interval of ENERGO. This joint data set was then used to construct and verify hypotheses in IPS-L (see figure 27)
As the student group was told during the plant visit in Dingolfing, painting colors need to be stored at continuant temperatures to guarantee high-standard and constant results. Therefore, the influence of weather on the energy consumption is also investigated in section 6.5.1. Accordingly, the joint data set was expanded by the weather data for November 2017. The data, taken from meteomatics.com, are precisely measured for the GPS coordinates of the BMW factory in Dingolfing (48.644217, 12.477284). The weather data contain 3 columns (figure 28):
average wind speed over last 15 minutes (in km/h), average temperature within last 15 minutes (in °C) and radiation within the last 15 minutes (in kW/m^2).

	X	IDENTNR	FARBE	FARBBEZEICHNUNG	ZAEHLPUNKT	ZAEHLPUNKTTEXT	ZEIT	VERSIONID	TimeDelta	QuarterlyTime
1	1	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z2041	Fueller Decklack	2017-04-12 07:33:20	2962405	1.0	2017-04-12 07:45:00
2	2	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z2051	Decklack	2017-04-12 08:31:07	2962405	2.3	2017-04-12 08:45:00
3	3	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z2055020	H41 WL2 Einlauf	2017-04-12 10:50:49	2962405	4.2	2017-04-12 11:00:00
4	4	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z2061	Lackaufzug	2017-04-12 15:04:54	2962405	0.2	2017-04-12 15:15:00
5	5	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	SAPZ2080	Z2080 an SAP	2017-04-12 15:13:56	2962405	0.0	2017-04-12 15:15:00
6	6	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z2081	HRL-Eingang Geb 53	2017-04-12 15:13:56	2962405	15.9	2017-04-12 15:15:00
7	7	682024...	A89	IMPERIALBLAU BRILLANTEFFEKT METALLI	Z3006	Sender 2 H52 Band 14	2017-04-13 07:09:25	2962405	0.0	2017-04-13 07:15:00
8	13	694024...			Z2041	Fueller Decklack	2017-08-29 20:50:21	6048954	3.7	2017-08-29 21:00:00
9	14	694024...			Z2051	Decklack	2017-08-30 00:33:46	6048954	56.4	2017-08-30 00:45:00
10	58	694024...			Z2061	Lackaufzug	2017-09-01 08:55:19	6048954	0.1	2017-09-01 09:00:00
11	59	694024...			SAPZ2080	Z2080 an SAP	2017-09-01 09:02:04	6048954	0.0	2017-09-01 09:15:00
12	60	694024...			Z2080	HRL GEB 56	2017-09-01 09:02:04	6048954	117.7	2017-09-01 09:15:00
13	1...	694024...			Z3006	Sender 2 H52 Band 14	2017-09-06 06:41:18	6048954	0.0	2017-09-06 06:45:00
14	8	694223...			Z2041	Fueller Decklack	2017-08-28 15:46:33	6048956	0.7	2017-08-28 16:00:00
15	9	694223...			Z2051	Decklack	2017-08-28 16:25:38	6048956	15.1	2017-08-28 16:30:00
16	10	694223...			Z2055040	H41 WL4 Einlauf	2017-08-29 07:32:34	6048956	0.8	2017-08-29 07:45:00
17	11	694223...			Z2055240	H41 WL4 Einlauf Trockner	2017-08-29 08:23:07	6048956	0.8	2017-08-29 08:30:00
18	12	694223...			Z2055480	H41 WL4 Auslauf Trockner Str.8	2017-08-29 09:09:15	6048956	28.9	2017-08-29 09:15:00

Showing 1 to 18 of 369,157 entries

Figure 27: IPS-L and ENERGO merged.

	TimeIntervall_15min	Energy	Wind	Temp	Rad
30	2017-11-01 07:15:00	888.8	5.6	4.3	43797.2
31	2017-11-01 07:30:00	890.8	5.8	4.4	159077.3
32	2017-11-01 07:45:00	897.2	5.9	4.6	135335.2
33	2017-11-01 08:00:00	892.8	6.1	4.8	110206.6
34	2017-11-01 08:15:00	892.8	6.2	4.9	242125.2
35	2017-11-01 08:30:00	893.2	6.2	5.0	423935.4
36	2017-11-01 08:45:00	890.8	6.3	5.1	357018.9
37	2017-11-01 09:00:00	890.8	6.4	5.2	350808.7
38	2017-11-01 09:15:00	892.8	7.2	5.3	443701.9
39	2017-11-01 09:30:00	893.2	8.0	5.4	661456.6
40	2017-11-01 09:45:00	898.8	8.8	5.4	554558.8
41	2017-11-01 10:00:00	895.2	9.6	5.5	681981.2
42	2017-11-01 10:15:00	892.8	9.3	5.6	634566.6
43	2017-11-01 10:30:00	888.8	9.0	5.7	695512.8
44	2017-11-01 10:45:00	892.8	8.8	5.8	570554.4

Showing 29 to 45 of 2,880 entries

Figure 28: Weather Data merged with ENERGO.

6.4 Time Differences at Main Nodes

As seen in figure 23, cars of a certain color or variant do not have a unique production flow. An obvious reason is, that it can be chosen between stations of the same category which fulfill the same production task. To analyze how long the production between two major stations takes, now only the time span between two production nodes of the station Z2041, Z2051, Z2061, Z2080 and Z2081 is considered. Additionally, only cars which are in two subsequent nodes on the same day are taken into account, to exclude possible shift breaks.
Analyzing how many production steps per time interval were made, one clearly sees an anomaly December 7 2017, from 2 p.m. until 4 p.m., where almost no production steps were tracked. There are two possibilities what could have happened. No data was tracked due to either a production stop or a data recording problem. An obvious check would be to look at the ENERGO data in order to identify the amount of demanded energy during this time slot. But the ENERGO data for December 2017 is not available. To analyze what might have happened, the production of the previous days will be compared to the production of Dec 7 2017. In detail, the distribution of the time spans between two nodes will be compared. In order to create equal comparison groups, for Dec 5, 6 and 7, only the measurement points from the time windows noon - 1 pm and 6 pm - 7 pm are used. For the computation of the time spans, only time intervals are considered, where two subsequent production stations strictly follow the given sequence Z2041, Z2051, Z2061, Z2080/Z2081. E.g. when only the measurement for Z2041 and Z2061 were recorded, no time interval is taken into account, as the time point for Z2051 was not tracked. Like this the real production span is not falsified. The time span of

the observations has to be chosen carefully. When the interval is too big, too many cars before and after the missing data window are measured and the distribution of productions spans will be wrong. When the interval is too little, it might happen that no car is tracked.
In the following two plots, the distribution of time spans between each major station and its subsequent station is compared for Dec 05, 06 and 07. One can see, that the distribution of time spans for Decklack

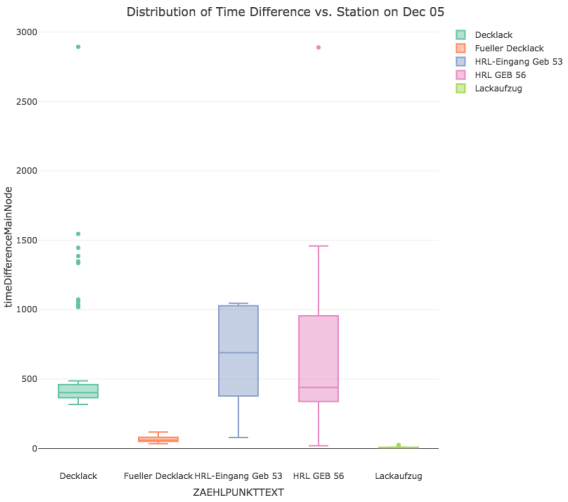


Figure 29: Dec 5

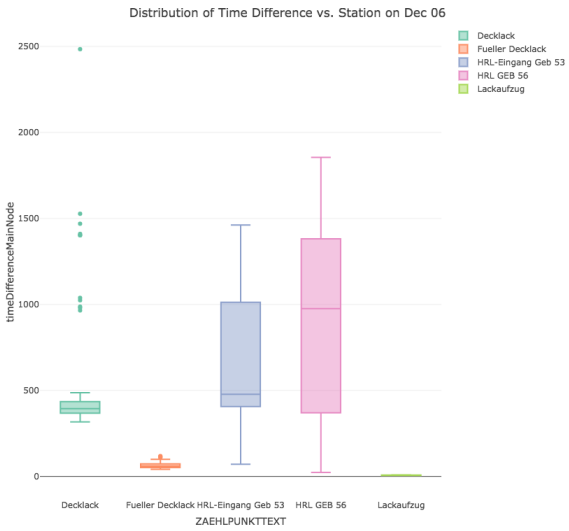


Figure 30: Dec 6

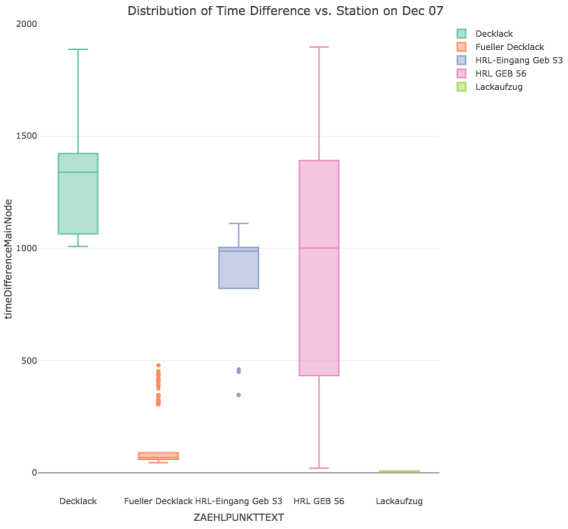
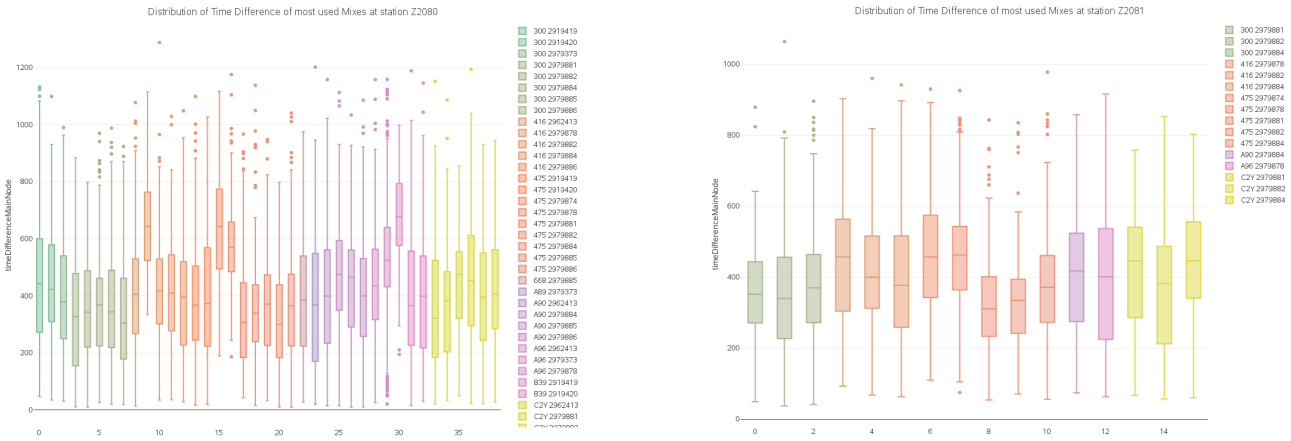


Figure 31: Dec 7

Figure 32: Distribution of Time Difference at major Stations.

is significantly higher on Dec 07 than usually. As well for "Fueller Decklack" there are more cars with a longer time spans than usually and also for "HRL Eingang Geb. 53", the distribution of time spans is significantly longer. Therefore it took longer for significantly many cars to move from one production station to the next one. However, on this day there was no change in how the production is done. Therefore, it can be assumed that the cars were waiting inside the stations and that there was a production stop from Dec 07 2 p.m. to 4 p.m.
An EDA of the time span between two major stations in dependence of the used color or variant, shows than an influence of these features is visible. Mixture models are now build, to analyze the interaction of color and variant in a more detailed way. Therefore for each unique car both attributes are merged and used as a new feature. There are over 100,000 possible combinations of Color and Variant in our observed data. Interestingly, there were only 720 different combination used and over 50% of the observed cars are one of 39 combinations of color and variant. Now the same time span

analysis as before is done. The result in figure 35 shows, that the same color used for different variants



Time Span at Station Z2080 Time Span at Station Z2081

Figure 35: Comparison of Time Difference to Next Station for Most Used Mixed Models.

has a different distribution of time spans between the major stations.

6.5 Models to predict Energy Consumption

After analyzing and visualizing the IPS-L data in the sections above, the next step in the project was to apply the ENERGO and IPS-L data sets for predicting the future energy consumption. After some research and discussions about promising methods for such tasks, the following machine learning and statistical methods were considered for further examination:

1. Train a Neural Network
2. Regression Analysis
3. Developing a Random Forest

As described in 4.1, only for November 2017 complete monthly data were available. Accordingly, the data set is significantly too small to train and validate a neural network architecture. Therefore, this approach was dismissed for the project. The following paragraphs explain the results of the regression analysis and the random forest method in detail.

6.5.1 Linear Model

At the beginning of the prediction process it was planned to start with a mathematical model providing a quick and flexible implementation, to rapidly test if relationships between energy consumption in the paint shop and different production features, e.g. car type, car color, exist. Hence, a linear regression model was chosen as benchmark model best matching the described requirements.
Considering the huge amount of parallel activities in a factory, it was quite unlikely from the beginning to find a perfect linear relationship between certain features and energy consumption. Nevertheless, this method is still very meaningful:
Even if the relationship between features and energy consumption is not linear but e.g. logarithmic or exponential, the existence of the correlation would still be visible in the slope. Consequently, a found linear correlation between features and energy consumption can be used as a backtest for the accuracy of a more complex and nontransparent models (e.g. neural networks).
Before the prediction started some data preparation was needed in advance. As it has been shown in section 6.1, the 17 most produced car models and 14 most painted colors represent over 90% of all produced car types and car colors respectively. To reduce the complexity of the data set, only these main colors and car models were concerned for further prediction. In addition, since factory building MS-13 (paint shop) contains a huge amount of production steps, which are independent of

the painting process, like storing of cars, only data from the painting station, "Einlauf ", were used for the prediction.
As a first step the intuitive assumptions of linear correlations between the features color and car model respectively and the energy consumption were analyzed. The results in the table below show that both color and car model are statistical significant features for the energy use of MS-13 (paint shop). Hereby, both F-value and R^2 suggest, that the car model seems to be slightly more relevant than color.
During the plant visit in Dingolfing the energy manager of BMW explained that in the process of painting the change of colors would cause an extra energy need. Using this information, the effect of color changes and car model changes were investigated. With an R^2 values between 10% and 12% and F-values above 200 (see figure 36), the information of the energy manger were supported by the data.
Discussions about the production process led to the assumption that the most relevant factor for the energy consumption could be the number of working steps, i.e. painted cars. This indicator reached by far the highest F-value and R^2 value. Moreover, in a model containing all mentioned regressors, the number of working steps was the only regressor with statistical significance. To validate this result,

Model	R^2	Adjusted R	F-value	p-value	Regressors significant on p-value (0.001 level)
$EnergyUse = \beta_0 + \beta_1 Col_1 + \dots + \beta_{14} Col_{14}$	0.1827	0.1767	30.67	< 2.2e-16	Color 1,2,3,4,5,6,7,11
$EnergyUse = \beta_0 + \beta_1 Car_1 + \dots + \beta_{17} Car_{17}$	0.2288	0.2215	31.07	< 2.2e-16	Car 1,2,3,4,5,6,7,10,14, 17
$EnergyUse = \beta_0 + \beta_1 WS$ (WS: Number of working steps per interval)	0.2051	0.2047	463.5	< 2.2e-16	WS
$EnergyUse = \beta_0 + \beta_1 ColC$ (ColC: Number of color changes per interval)	0.107	0.1065	215.3	< 2.2e-16	ColC
$EnergyUse = \beta_0 + \beta_1 CarC$ (CarC: Number of car type changes per interval)	0.1207	0.1202	246.5	< 2.2e-16	CarC
$EnergyUse = \beta_0 + \beta_1 Col_1 + \dots + \beta_{14} Col_{14} + \beta_{15} Car_1 + \dots + \beta_{31} Car_{17} + \beta_{32} WS + \beta_{33} ColC + \beta_{34} CarC$	0.246	0.2319	17.44	< 2.2e-16	WS

Figure 36: Results of Linear Regression.

a binary case for the feature number of working steps was developed. In a binary case, the regressor is either fully considered in each regression step or not considered at all. By doing so, it is easier to investigate its effect.
Therefore, the regression was repeated, where very few production steps (1 to 5 steps) are assigned with a 0. In contrast, a high number of production steps (19-22 steps) are assigned with 1. The following regression led to almost a doubling in the R^2 value and the adjusted R^2 value (40.25% and 40.1% respectively). This result strongly supports the assumption that there exists a correlation between the number of working steps and energy consumption.
During the visit at the plant in Dingolfing the student team learned that painting colors need a continuant temperature to ensure a high-standard and constant result. Therefore, the idea was developed to investigate whether the weather influences the plants energy consumption. As described in section 6.3, the weather data of November 2017 for the location of BMWs factory in Dingolfing were used for the regression. To examine, whether and which weather feature affects the energy consumption, several linear regressions were conducted. The results are listed in the table below. It is indicated that temperature and wind speed are statistically significant for the energy use. In contrast, with a p-value of 0.8781 and a F-value of 0.02353 (figure 37), the assumption that radiation effects the energy consumption can neither be rejected nor confirmed.
After examining the significance of certain regressors, the last objective of this projects regression

Model	R^2	Adjusted R	F-value	p-value
EnergyUse = β_0 +β_1 WindSpeed	0.02277	0.02243	67.06	3.905e-16
EnergyUse = β_0 +β_1 Temperature	0.06735	0.06703	207.8	< 2.2e-16
EnergyUse = β_0 +β_1 Radiation	8.175e-06	-0.0003393	0.02353	0.8781
EnergyUse = β_0 +β_1 WindSpeed +β_2 Temperature	0.08746	0.08683	137.9	< 2.2e-16
EnergyUse = β_0 +β_1 WindSpeed +β_2 Temperature +β_3 Radiation	0.1178	0.1168	128	< 2.2e-16

Figure 37: Results of Linear Regression on Weather Data.

analysis was the prediction of future energy consumption. Firstly, 80% of the data from the data set were randomly picked for the training set, while the remaining 20% of the original data set formed the validation set.
As shown above, with a R^2 value and an adjusted R^2 value of 24% and 23% respectively the final model in figure 36 was the statistically most significant one, containing the 3 features color, car type and number of production steps. Therefore, this model is used for a first prediciton.
The accuracy of the prediction is measured by the root-mean-square deviation (RMSD), which is 112.3 kWh. With an average energy consumption of 1426.4 kWh, the prediction error is about 7.8%.
Since there exist many mathematical models, which use the previous value of the dependent variable to calculate the new value, e.g. Wiener Process, this idea was adopted. Consequently, the average energy use of the previous day was introduced as a new regressor. For the now updated model the previous energy consumption was combined with the feature number of production steps, which was the most statistically significant feature so far.

$$EnergyUse = \beta_0 + \beta_1 AverageEnergy_{previousDay} + \beta_2 WS \tag{1}$$

Model 1 has a R^2 value and an adjusted R^2 value of 33.3% each. Additionally, the model also has a lower RMSD of 92.5 kWh, which corresponds to a prediction error of about 6.5%. These results indicate that the energy consumption of the previous day is statistically significant for the current energy consumption.

6.5.2 Classification

In MS-13 the information about the demand for electricity is quite aggregated, both, in terms of time and location. Contrary to that, the IPS-L Data provides a detailed overview of the cars properties, like e.g. color, and the time point they were painted. Therefore another idea is to analyze the relationship between the time point a car with a certain color and variant was painted and dried. It can be assumed that these production tasks are the most energy-intensive tasks in the paint shop.

		Predicted	
		HPLW	not HPLW
Observed	HPLW	1259	14
	not HPLW	7924	177

Figure 38: Fall

		Predicted	
		HPLW	not HPLW
Observed	HPLW	613	54
	not HPLW	2365	240

Figure 39: Winter

Figure 40: Result of the classification as confusion matrix.

The gained information from the EDA figure 26 motivates to investigate if a car of a certain color and variant is always painted at a certain time point of the day. For the classification a random forest was chosen, it generates different uncorrelated decision trees. Its advantage is that it trains very fast on large data sets and its decisions are understandable (in contrast to e.g. a neural net). For the implementation the R package `RandomForest` was used, as it works with categorical features, as well. To measure the accuracy of the classification, the data set was randomly split in a training set (80%) and a test set (rest of the data, 20%). The model was build on the training set. On the test set the classification through the model for each observation and its observed value could be compared. A first try to classify the time point of a day (divided in 15 min intervals) using the features color and variant only showed a weak relationship (cf. appendix figure 44). Clearly, in this example the difference between two 15 min intervals usually is not significant enough for a classifier to identify it. Theoretically, a better criterion for the grouping is, whether a measurement was during the HPLW or not. This means, if a special color or variant is on purpose not painted during the HPLW, will be analyzed. A reason for this could be, that an increase of electricity consumption during the production of a certain color or variant is already known. As the HPLW for fall is shorter than for winter, both classifications are made independently in order to determine if an effect is visible. The result in the confusion matrix 40 shows, that indeed a differentiation can be made and that some colors and variants are classified explicitly during the HPLW or not HPLW. However, in both cases the classifier yields a high amount of false positives. This can be caused by the extreme imbalance of the classes. The data points measured when the HPLW is active, are only 6% and 10%, respectively, of the whole data set. A resampling of the data, where the under-represented was duplicated and over-represented group was randomly decreased, was not successful. Further techniques to improve the classification in imbalanced groups are discussed in [4].
For the classification the 4 stations "H41 WLx Einlauf" are used as a reference point, as the electricity demand should be the highest at this working stations. As well different time lags were used, but they couldnt improve the result. The result of the classification cannot confirm the thesis, that different colors or variants are on purpose not painted during the HPLW. Even though the IPS-L data provides a detailed view of the logistic perspective in the paint shop, these information cannot be used to identify which features lead to a higher electricity demand.

6.6 Color Clustering

The regression in 6.5.1 has shown the statistical significance of number of color changes for the energy consumption. Therefore, the first objective was to examine the painting chronology to determine frequent color combinations. If these frequent painting sequences contain combinations with numerous color changes, this would indicate that there would exist potential for energy savings by optimizing the painting order.
Like in section 6.5.1 the data set was reduced to the 14 most used colors to scale down the dimension of possible color combinations. To find frequent color combinations different clustering methods were applied. The popular approaches of K-Means (clustering by numerical values) and K-Modes (clustering by categorical values) were implemented. To apply both methods a corresponding matrix needed to be developed. Hereby the matrix is of dimension 14×28 for K-Means and 14×29 for K-Mode. The 14 rows represent the 14 most used colors in the examined time interval. Columns 1 to 14 represent the previous colors to each row and columns 15 to 28 the subsequent colors to each row. So fields are the aggregated counts, how often a certain color combination appears. For example, if at an investigated time point color 5 is painted and the previous color was 3 and the subsequent color is 2, then the fields (5,3) and (5,14+2) of the matrix would be increased by 1. For K-Mode this 14×28 matrix has been expanded by adding a new column 1, which contained the categorical information about the color (i.e. 1-14). According to the structure of the 2 matrices, K-Mode and K-Means were meant to cluster frequent color combinations.
Both clustering methods could not find clusters containing more than one color, so each color forms its own cluster. The results of K-Mode and K-Means are supported by the figures 41 and 42. The

histograms show the distribution of previous and subsequent colors for the 14 colors. All histograms are clearly unimodal distributed, which leads to the conclusion that colors mostly follow themselves, which reduces the number of color changes. Therefore, it can be inferred that the painting sequence is already optimized with regards to color.

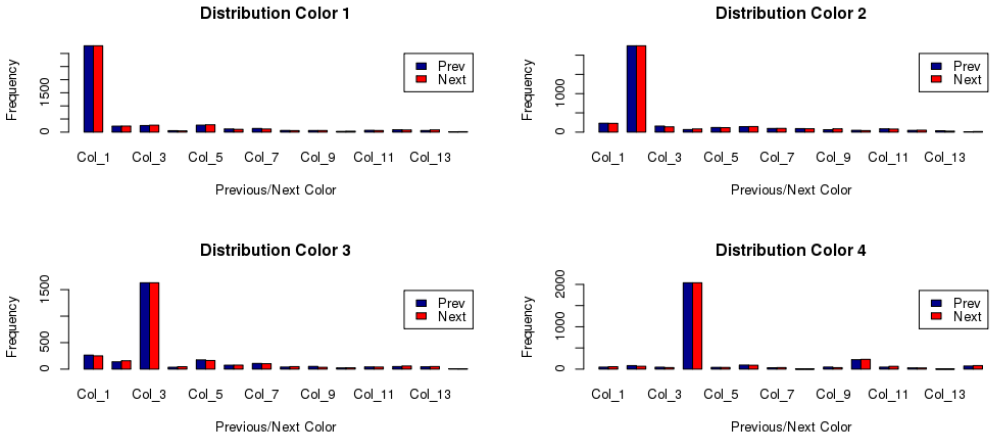


Figure 41: Histograms of Color Sequences for Colors 1-4.

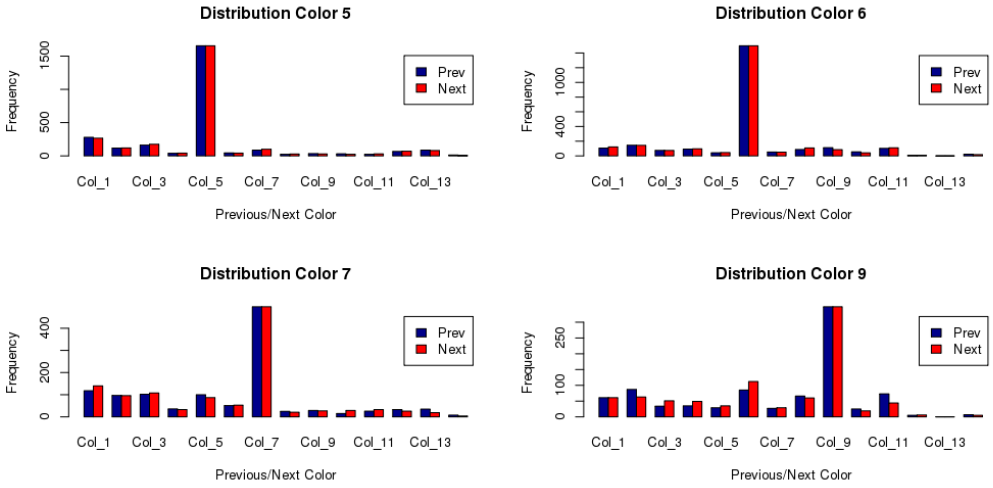


Figure 42: Histograms of Color Sequences for Colors 5-7 & 9.

7 Conclusions

The main goal of the project is to predict the energy consumption on the BMW plant. Energy consumption of the plant must be reduced during the peak load windows. However, energy consumption is influenced by many different parameters and peak loads usually occur spontaneously.
The most meticulous part of the project was the data explanatory analysis. Three different datasets were examined for a better understanding of the production processes, their connection to each other, and their influence on the energy demand of the whole plant.

7.1 Results

The mapping of the ENERGO data set to IPS-L showed that the used aggregation of 15 minutes and the division the the FS are too imprecise for mathematical modeling. Therefore, it can be inferred that different data sets showing the energy consumption should be taken into account.
Secondly, several energy relevant features could be detected by regression analysis (e.g. number of working steps and color) and were used for predicting future energy consumption. The influence of external effects (e.g. weather) on the energy consumption was proved.
Color clustering showed that the significant feature "number of color changes" is already optimized in the production process by the avoidance of high frequency color changes.
The machine learning approach by a random forest indicate that the features color and variante are not sufficient to determine if a car is in a HPLW.
For IPS-T dataset, linear models were able to predict energy consumption with low error rate. Whenever there is a predicted peak during HPLW, the production schedule can be altered to avoid the peak. Clustering was used to classify energy profiles of 6 nodes during the night shifts. As a result of this exercise, it was possible to group energy profiles. For every cluster, there could be an energy consumption reduction within the cluster by optimizing the schedule of the ramp-ups using the minimum energy profile within every cluster.

7.2 Outlook

For IPS-T model, more data would be necessary to make sure that the model is accurate along the year. In addition, a production pipeline can be created to make predictions on a regular basis. In order to analyze and predict the peaks better, all peak loads should be logged. These markers can then be used to formulate a classification problem. The production line can be mapped to the energy consumption data better, in order to extract more detailed information.
As to what it respects to the future of Energy peak prediction in BMW, it is valid to suggest to replicate some of the techniques or algorithms used to predict energy profiles in other production plants. In spite of data structures being likely changing from one plant to another, the data processing pipeline could be re-utilized. Also, being Dingolfing one of the biggest BMW plants, there are no hints on why the same algorithms wouldn't perform well in production plants of lower complexity.
The IPS-L data provides the view of the paint shop MS-13. In the BMW Plant Dingolfing, there is a second paint shop, MS-5. In MS-5, both, heat and cooling is needed at the same time, which makes this MS to the MS with the highest electricity demand (c.f. figure 26). Chapter 4.1 already focused on its electricity demand and the availability of internally produced electricity. In IPS-L, there are five working stations which indicate an exchange of cars from MS-13 to MS-5 and vise versa. Here a further investigation should be concerned. According to information provided during the visit to the plant, the same technologies (e.g. same kind of paint robot) are used in the plant Leipzig, with a different setting of parameters. Here, a good comparison with different parameter settings and the therefore resulting electricity demand can be made over a longer time interval. If further monthly IPS-L data would be available, this information could be used to apply state-of-the-art machine learning methods, as e.g. neural networks (c.f. chapter 6.5). For these models could be used to test the presented findings and predict an energy consumption.

References

[1] BMW. *Welcome to BMW Group Plant Dingolfing.* `https://www.bmwgroup-werke.com/dingolfing/en.html`. [Online; accessed 31-January-2018].

[2] Wikipedia contributors. Median absolute deviation — wikipedia, the free encyclopedia, 2017. [Online; accessed 14-February-2018].

[3] A.-M. Gruber. *Zeitlich und regional aufgelöstes industrielles Lastflexibilisierungspotenzial als Beitrag zur Integration Erneuerbarer Energien.* PhD thesis, Technische Universität München, 06 2017.

[4] H. He and Y. Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications.* Wiley-IEEE Press, 2013.

[5] Reuters. *BMW to build new 8 series at German Dingolfing plant from 2018.* `https://www.reuters.com/article/us-bmw-dingolfing/bmw-to-build-new-8-series-at-german-dingolfing-plant-from-2018-idUSKCN1BX2ZB`, 2017. [Online; accessed 31-January-2018].

[6] Wolf-Peter Schill. Residual load, renewable surplus generation and storage requirements in germany. *Energy Policy*, 73:65 – 79, 2014.

8 Appendix

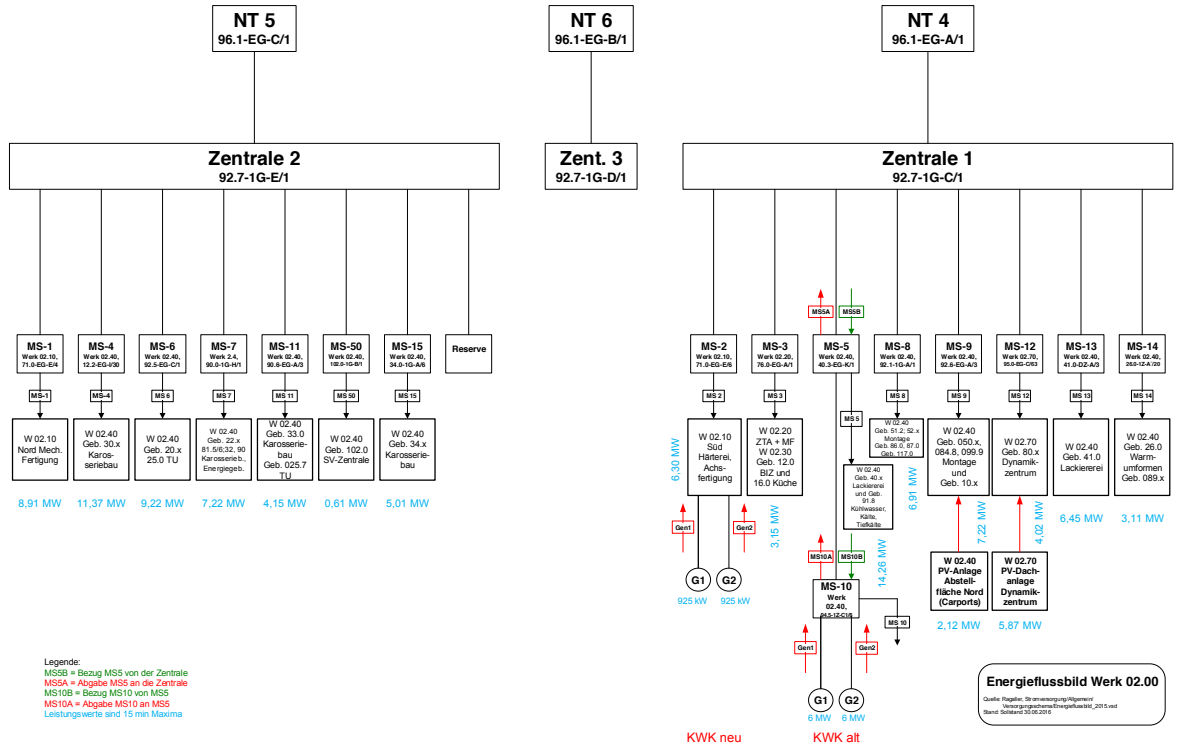


Figure 43: Energy flow in the plant Dingolfing.

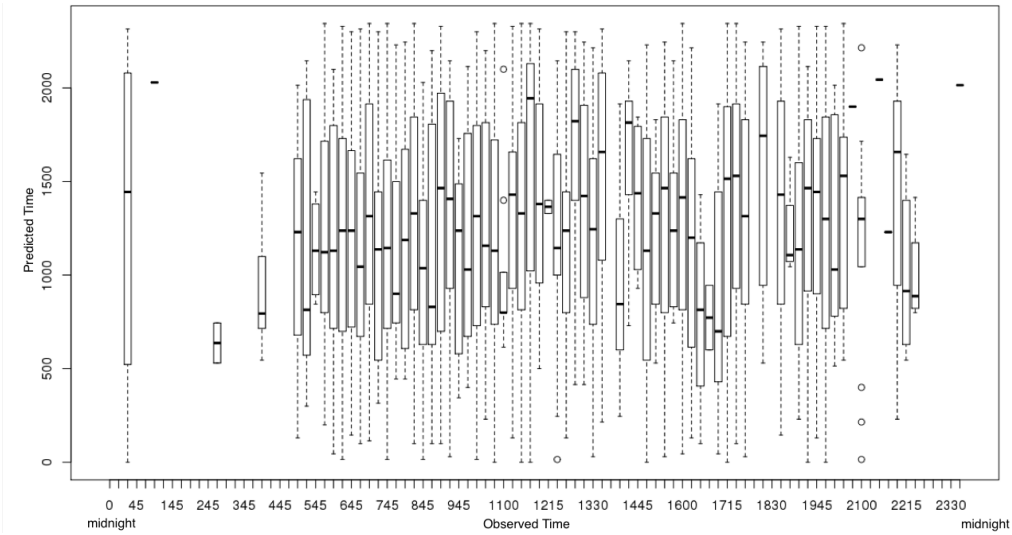


Figure 44: Predicted vs. Observed Values of the Classification.

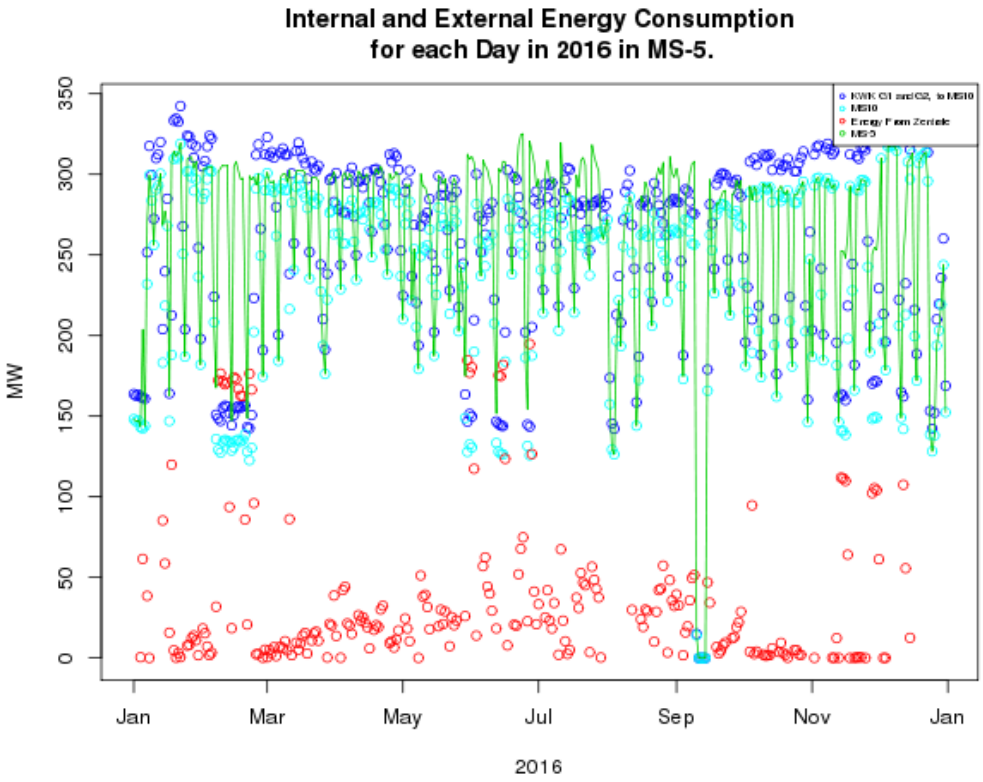


Figure 45: Energy Demand of MS-5.

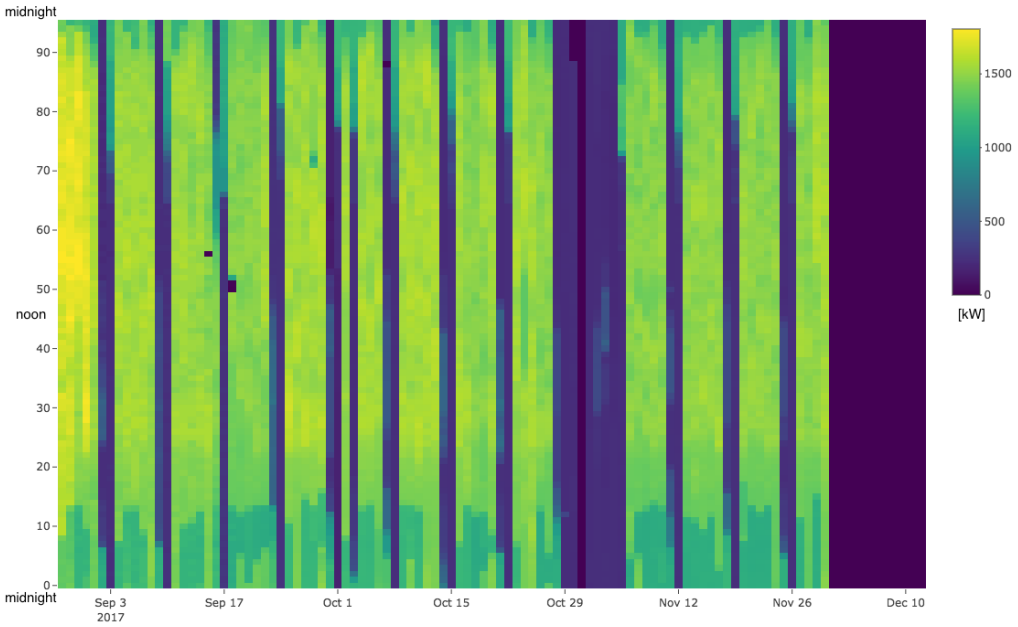


Figure 46: Energy demand of MS-13.