Documentation of Practical Part of Exam for MLA Lecture

Performance comparison – regional differences in the wind farm
MLA-Group 5
Winter semester 2019/2020

Group Member

- Yihan Hu MN: 2328366
- Zhanke Liang MN: 2653044
- Li Liu MN: 2972291
- Marvin Stoeckel MN: 2062583
- Rukang Xu MN: 2729855
- Boxuan Zhang MN: 2479736
- Michael Luttmer MN: 9070140

Content

1. Introduction

Wind energy production has relatively safe and positive environmental characteristics and has grown from fringe activities to billion-dollar industries in the past few decades. It is one of the main keys in reducing CO_2 emissions. Even though it is one of the biggest problems of our time, there are a lot of restrictions in Germany, which dramatically reduce the area for the construction of wind farms.
This increases the aim of electric utilities companies to optimize the efficiency of their existing wind parks and also use the gained knowledge of previous investigations to place new wind turbines at the optimal position. Besides the conventional analysis, it makes sense to use machine learning algorithms to investigate problems of existing wind parks, because of the sheer number of different influencing factors.
So, the main task is to identify regional differences in each of the two given wind farms. Therefore, important ambient parameters have to be identified and analyzed.

It is assumed that the wind turbines influence each other, for example through the so called wake effect. One of the suggestions is, that a too small distance between units will lead to a loss of efficiency for a specific wind direction, because after passing a wind turbine, the laminar flow the air is not given anymore. As result, the turbulences regarding to different directions have to vary.
The position itself should also have an impact on efficiency and power output. If the wind turbine is for example in a valley, the wind speed should be different to a mountain top. More general: How do the ambient conditions influence the wind turbines to be a high or low performing unit. The investigation of the given dataset should lead to the identification of those units. Afterwards the reasons for this varying performance has to be examined.
The gained experience out of this investigations can be used from electric utilities companies to improve the knowledge of the influence of the environment in respect to the power output.

2. Concept and Methodology

In the following, the *CRISP-DM* process will be used relatively strongly. Since the work order is limited to the examination of the given data sets, models for knowledge generation are used and deployment is waived. The evaluation takes place with the help of classical investigations without machine learning as well as different machine learning methods among each other.
As already described in chapter one "Introduction", differences in the performance of the turbines should be shown and if possible the reasons for this should be identified. This will help to make power generation more efficient, which covers the area of "*Business Understanding*".
The data set is sometimes divided into different csv files for each turbine. These csv files must be merged to access all parameters simultaneously.
This happens during preprocessing. Certain errors in the data records are also corrected here. As such, duplicate timestamps, completely empty rows and columns, "Not a Number" values and outliers were identified and handled. In order to ensure that the measured values are valid, a comparison with the trust files is also carried out.

In the modelling process, different approaches are used to perform a validation. On the one hand, normal investigations are performed without the use of machine learning. Secondly, two clustering algorithms are applied. These two cluster algorithms were used because we want to find the turbines with the smallest differences of all values. So, we have to find the quadratic variance of the clusters. In addition, you can see the similarities of the turbines outside your own cluster very well.
Hierarchical clustering was used because the resulting representation is very well suited for this task. You can see how far apart individual turbines are from each other and the way the distance is calculated benefits the analysis. Due to the group's decision to use two clustering methods, kMeans was selected as the second method. It was chosen because of its ability to be applicable to many problems and because it works well with flat geometries and a medium number of clusters.
Lastly, an attempt is made to make qualitative statements about the relationship between the environmental conditions with regard to power output. For this purpose, self-organizing maps are used. These can give a visual overview of the relationship between two variables in a descriptive way.

In the fifth task area of CRISP-DM, the individual concepts are compared with each other in order to carry out an evaluation. For this purpose, a statement about the plausibility is made on the basis of the classical methodologies.

3. Technical Implementation

The programming language is Python 3.8.1.
Spyder is used for implementation. Other packages used are pandas 1.0.0.1 to use the datasets in a suitable way, scikit-learn 0.22.1-1 for the machine learning algorithms and matplotlib 3.1.3-1 and windrose 1.6.7 for the visualizations.
In some places numpy 1.18.1-1 is also used.

In addition to the files containing the measured values, there is always a so-called "Trust" file. It classifies the recorded measured values as valid or not valid. Here an adjustment must take place, which identifies non-valid measuring points and carries out a corresponding action.
Various problems have been identified with regard to the quality of the data. In addition to "Not a Number" values (NaNs), there are also problems due to duplicate time stamps. Sometimes the lines (timestamps) with exactly the same values occur twice. Sometimes there are measurement values with valid and invalid values for the same timestamp. In addition, there are timestamps that have completely different measured values for the same parameters and are simultaneously marked as valid by the "Trust" file. In addition, there are lines and columns in the raw data that consist exclusively of NaNs and therefore contain no information.
These problems are solved in the "Data Preparation" section, visualized in figure 2.1. The different files of the turbines are merged and rows and columns without information are deleted.
Whether measured values are valid is determined by matching each row and column of the "Trust" file and deleted with a matrix multiplication. To do this, every zero in the trust file is first made a NaN and then the multiplication is performed. As a result, the valid measured values remain and the NaNs in the trust file make an invalid measured value as a result of the product also a NaN.

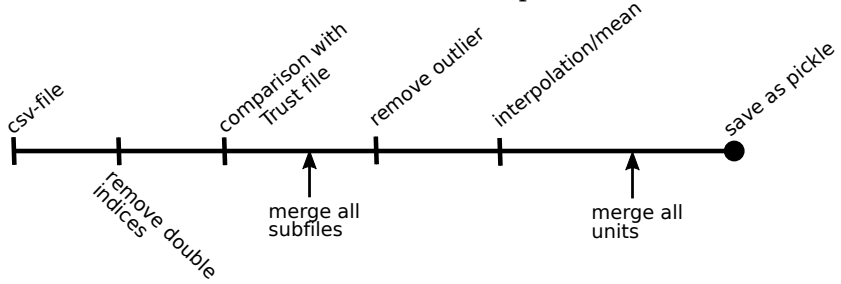

Figure 3.1: The preprocessing procedure

The duplicate time stamps are removed as follows: If the measured values of the same times are completely identical, all but one line is deleted. At times where there are valid and invalid measured values, the invalid ones are deleted. If there are completely different, valid measured values, it cannot be decided which data set is the correct one. The first measured value is selected and all others are deleted.

In the file “ENBW Data Signals Description.json” are the minimum and maximum values for each parameter stored. The values exceeding these limits have been classified as outliers and converted into NaNs. To prevent malfunctions, a limit of two percent has been introduced, which the algorithm was allowed to change without generating an error message.
The resulting set of measurements contained only NaNs and valid measurements. In respect to the position of these NaNs a different proceeding has been applied: NaNs between valid measured values are interpolated, based on the assumption that time courses have a relatively continuous course. NaNs, where are no given datasets before or after, the mean of the parameter has replaced the “Not a Number” value, because an interpolation is not possible and the mean value does not.

To be able to make better statements, two new parameters are introduced which are not available in the data set. The turbulence intensity and the density are calculated. quellen!

$$Amb\_Turbintensity = \frac{Amb\_WindSpeed\_Std}{Amb\_WindSpeed\_Avg}$$

$$Amb\_Density = \frac{p}{R_S * Amb\_Temp\_Avg}$$

with:

$R_S = 287{,}058 \frac{J}{kg} * K$

$p = 1000$ hPa

Since all three machine learning algorithms require scaled measured values, this step was also performed. As a final step in “Preprocessing”, the data sets of all wind turbines were merged into a single data set in order to adequately investigate the differences within a wind farm. So from many different files the number was reduced to one file per wind farm.

With hierarchical clustering, only the environment data (Amb_Windspeed_Avg, Amb_WindDir-_Abs_Avg, Amb_Turbintensity and Amb_Density) are used. To make the time-dependent measured values available for this type of machine learning, the mean value, standard deviation, skewness and kurtosis are calculated for each of these parameters. The model is then formed from these values. Here the "ward hierarchical clustering" is used. It has the goal to minimize the quadratic variance and this is also the goal of our investigations.
For both wind farms the optimum k was determined using the elbow curve. The elbow curve shows the sum of the deviation squares plotted over k and the parameter is selected by the elbow in the plot. This way an ideal k can be chosen which allows a compromise between variance and cluster size. This is done because the turbines should be divided into clusters as large as possible.
Self-organized maps: In an iterative process, a relatively high learning rate of 0,7 was chosen in order to adjust the weighting more closely. This is intended to achieve a better general explanatory power.
The following general rule of thumb was applied for the sum of neurons: quelle

$$Sum\ of\ Neurons = 5 * \sqrt{Sum\ of\ Measurements}$$

Additionally a new parameter was introduced for the self-organized maps which describes the seasons. The months December, January and February were defined as winter and assigned the value zero. Continuously every three months the value "Season" was increased by 0.333 to achieve a good visualization in the SOM.

4. Presentation and Discussion of Results

4.1 Classical investigations

Due to the smaller number of turbines, wind farm 1 can be better analyzed with classical methods. For the description of the wind direction, as well as the proportions of the different wind strengths, the visualization by using a "wind rose", shown in figure 4.1, is suitable. The orientation of the individual bars indicates the wind direction. The size of the different colored subdivisions is an sign for the speed components. The brighter a section is, the higher the wind speed
It can be seen that the wind in all turbines comes mainly from north-west to south-west. There is almost no wind flow from northern or southern directions. Towards the east there is again a small rise. The most striking difference compared to the other plants is turbine 2. Here, proportionally, there is hardly any wind speed of more than 9 m per second. Furthermore, one can see a relatively high similarity between turbine five and six, and between one and three. For the investigation of wind farm two, the conventional investigation without machine learning and the representation by the wind rose already reaches its limit. For this reason, we will omit the presentation here, since no information can be extracted from this plot.

If you only look at the power, the picture is different for wind farm 1 (figure 4.2 on the left): Here there are four average turbines, as well as one turbine each with significantly more and significantly less average power output.

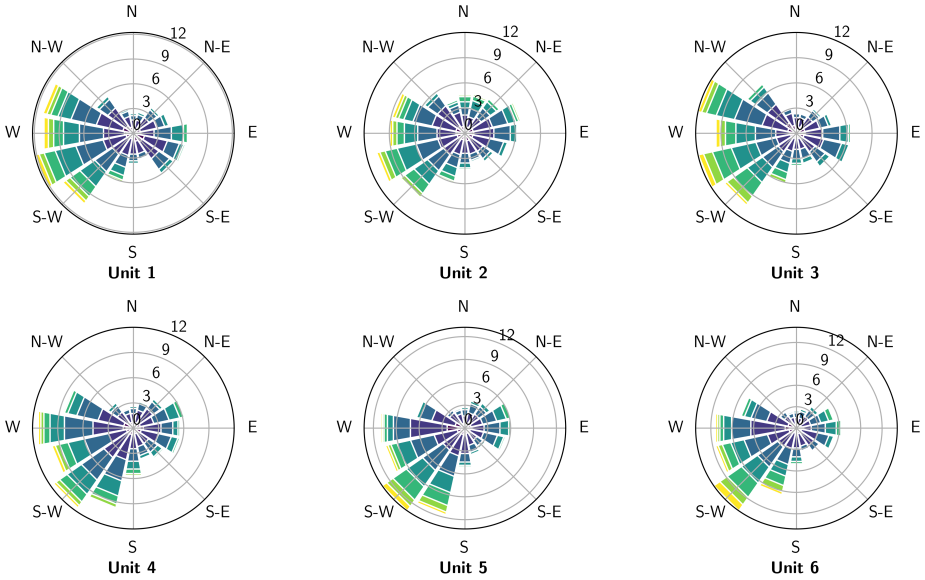

Figure 4.1: The Windrose plot of wind park one

On the right side of the picture the average power output of the wind farm two are listed. Like wind farm one, they are ordered from left to right according to the power output. You can see here that there is an outlier. Turbine twelve has on average 5 kilowatts less power output than the second worst unit. The remaining 17 machines are roughly in a continuous field. However, there are more or less conspicuous gaps between some turbines, which divides the units into four groups.

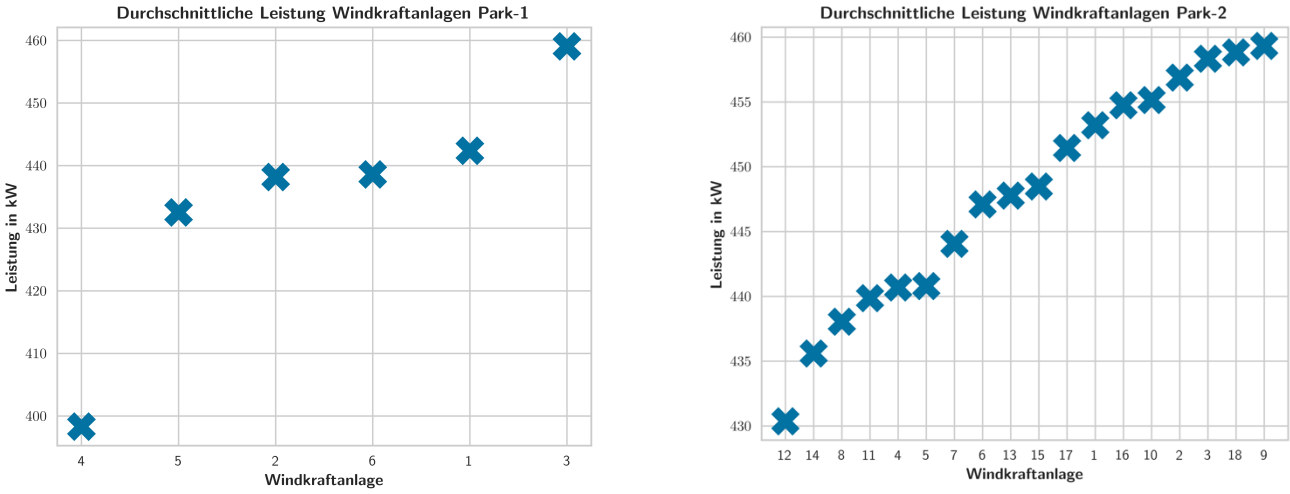

Figure 4.2: The mean power output of the two wind parks

4.2 Hierarchical investigations

The results of hierarchical clustering are shown in Figure 4.3. Here the Euclidean distance to the other turbines(x-axis) is shown on the y-axis. Input parameters are the environmental conditions. In comparison to the diagram 4.2 there are only few differences. Turbine four and three have been assigned to the same group in clustering, but turbine one is also in the same group. This unit, however, is shown in the middle field in terms of power output. Clearly different to all other turbines is unit two. It has

hardly any similarities with the rest. However, if you compare the picture with picture 4.1, there are great similarities. Turbine five and six are also assigned to each other here. The same is true for units one and three, plus unit four. Number two is also clearly different. As an observation it becomes obvious, turbines hardly differ in power output, when they have similar environmental conditions.
In wind farm two an analysis becomes much more difficult, but the comparison of environmental parameters to the average power output fits better. Here the wind turbines (4,5,7) are next to each other in both figures. The red cluster (9, 10, 18) is in the high-power output group. The same applies to the remaining clusters, except for wind turbines three, eight and sixteen. Here three and sixteen have high power output, however turbine eight is rather bad.
Nevertheless, the assumption that the ambient conditions can be used to draw direct conclusions about the power output can be reinforced, particularly in the case of wind farm two. In this representation through hierarchical clustering, the individual groups are also very well visualized, which was one of the main tasks.

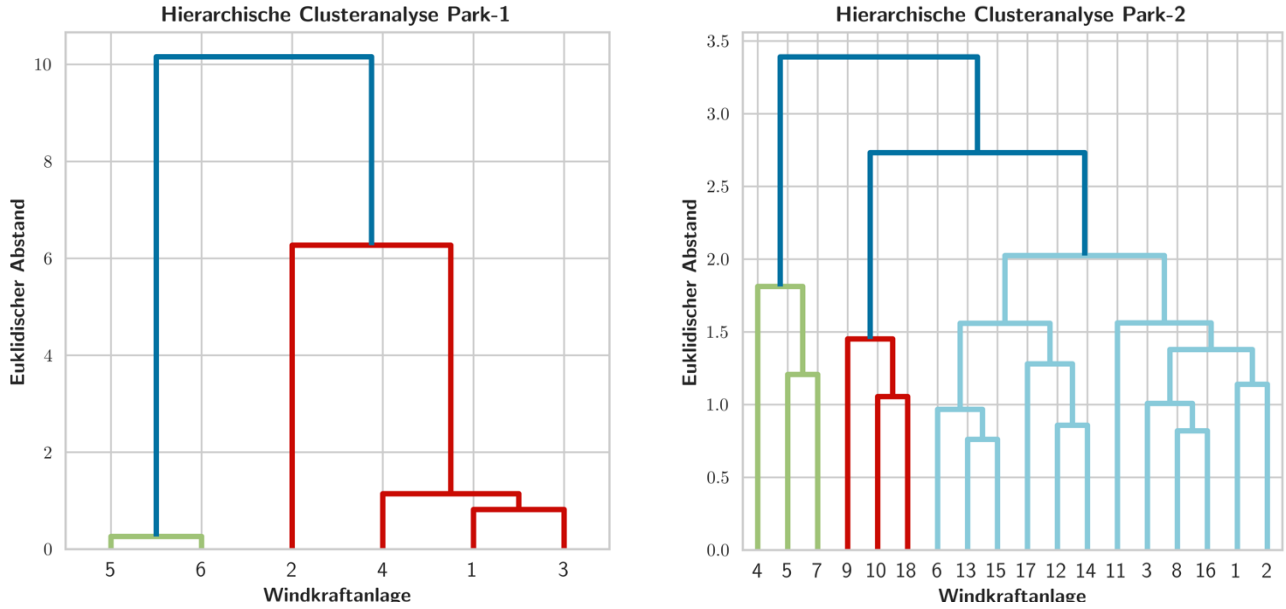

Figure 4.3: The results of the hierarchical analysis

4.3 Investigations with kMeans

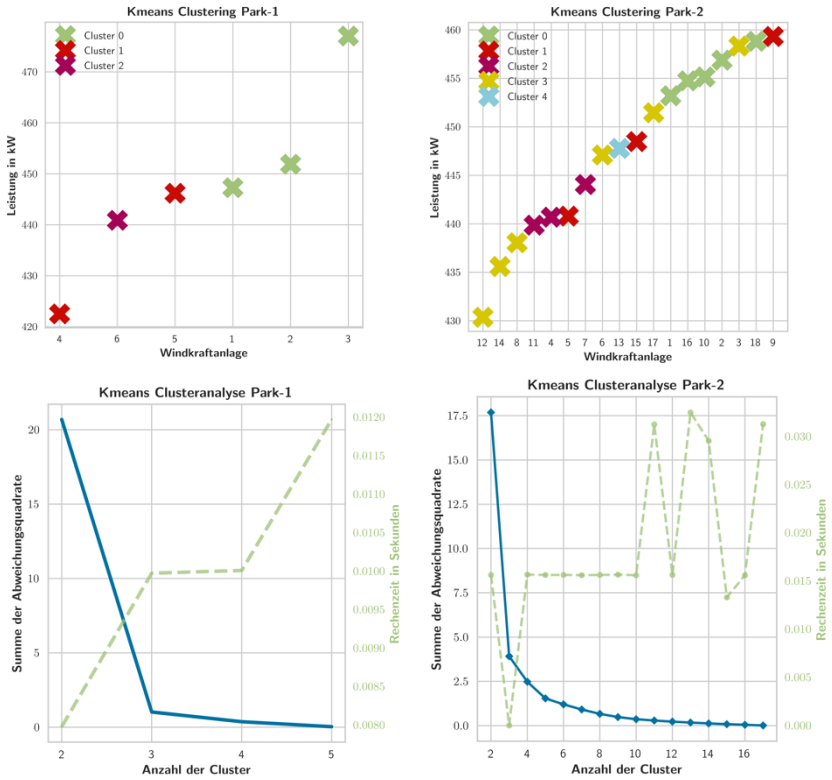

Figure 4.4: The results of the kMeans clustering with the elbow curves

4.4 Qualitative statements with self organizing maps

Even if self-organizing maps are used to reduce dimensions, one can also make qualitative statements, or at least make assumptions. After having identified high and low performers in the individual wind farms, one can perhaps make a statement here as to which environmental parameter has the greatest influence on the power output.

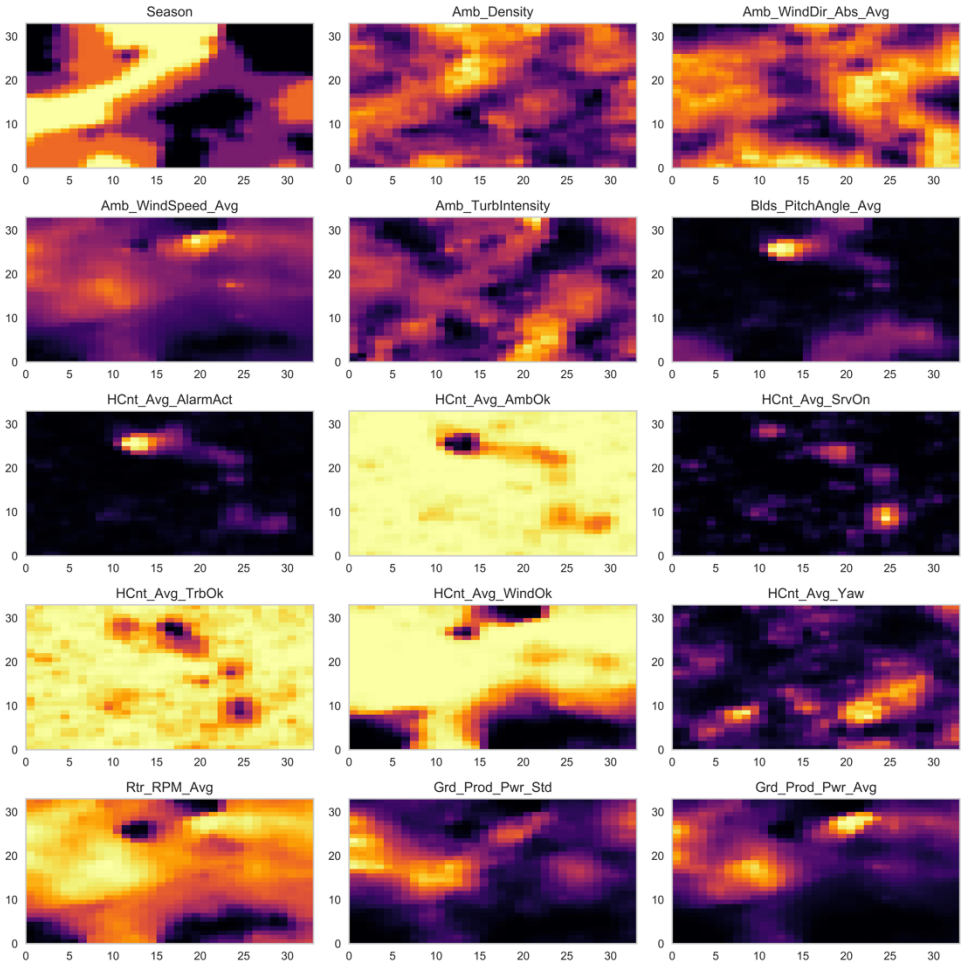

Figure 4.5: The results of the self organizing maps

In

5. Applicability Analysis

Especially for large wind farms like park two, a conventional analysis of the environmental parameters compared to the power output is very difficult. This is where machine learning becomes an obvious choice. Conventional statistical methods can only be used here to verify the results.
From the given wind farms it can be seen that the power output can be concluded with relatively good accuracy using hierarchical clustering based on the environmental parameters. However, the turbines are regulated from a certain power level upwards, so that high wind speeds no longer lead to a better output. The self-organizing maps made it clear that the operating company should pay very close attention to the wind speed. Ideally, positions should be chosen which do not necessarily have very high wind speeds, but a steady current. After all, this is by far the most important factor for increasing electricity production and in this case the energy can be used best.
It is also clear that the ambient temperature does not have as great an influence on the energy output as one might think. The different densities do not result in significantly higher values, but a correlation can occur due to the different seasons and the different wind percentages that occur in them.

kMEANS

For all selected procedures, however, no qualitative statement could be made. The size of the individual differences between the environmental parameters of the turbines can only be guessed at. Also the connection between the individual parameters and the performance is only qualitative. In this context the selected algorithms are limited.

6. Summary and Outlook

Briefly summarize the most important aspects of the documentation. Outline which steps you would recommend to the industry partner with regard to the task.

7. References

Use the citation style of the IEEE. The sources are numbered chronologically according to their occurrence. Further information can be found here:
https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf
https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf