Udacity Machine Learning Nanodegree

Capstone Proposal

Dependence of the soiling rate of solar concentrators on weather parameters.

1. Domain Background

Figure 1: Solar Through Collector, source: www.dlr.de

Concentrating Solar Power plants (CSP plants) capture and concentrate sunlight and transfer its energy to a heat transfer fluid. Currently there are installed about 5000 MW of CSP power worldwide (source: cspplaza.com). It is predicted that due to the rapid development of emerging markets including Morocco, South Africa and China the CSP capacity will have a huge increase in the next years.

Soiling (= “getting dirty”) of these solar concentrators (glass mirrors) is an important issue that significantly reduces plant efficiency and causes high cleaning costs for the plant operators. For a lot of potential plant sites there is no soiling data available. Also, in state-of-the-art cost-effectiveness studies during the planning phase of a CSP plant the soiling issue is not considered in sufficient detail: the soiling rate is assumed as a constant factor independent from plant location, without variation during the year and constant during the entire plant life. A recent study by (Wolfertstetter, 2017) does an in-depth analysis on the effects of soiling on CSP plants. Soiling is dependent on several environmental factors like wind speed and direction, condensation and aerosol particle concentration. All of these parameters vary during the time of a year and strongly depend on the location of the plant. In the work of (Wolfertstetter, 2017), a new measurement device is designed, that measures the cleanliness of solar concentrator (mirror) probe. During 3 years the cleanliness of the mirror has been logged together with the above-mentioned weather factors. The soiling rate is then found as the time derivative of the cleanliness.

2. Problem statement

In the work of (Wolfertstetter, 2017) a prediction model for the soiling rate is built. The model is based on linear regression. The independent parameters (features) are the weather parameters like temperature, humidity, windspeed, mirror orientation and others. The dependent variable (label) is the soiling rate. As a result of the regression analysis it was found that the Pearson coefficient was > 0.3 for none of the features. None of the features shows strong linear correlation with the dependent variable. Only 4 of 43 input features showed a weak correlation to the labels, and all the other features did not correlate significantly with the labels. It is estimated that the model can make reliable predictions for a maximum of 2 months only. The aim of this work is to find a better prediction model.

3. Dataset description

a. original dataset with 1-min resolution

This dataset provides 1-min time resolution for all 43 measurement channels (features). It contains 874k data points. The measurements were performed over approximately 3 years. The label (soiling rate) is not available in minute resolution, but in 1-day resolution. The cleanliness is measured in minute resolution and could be used as label. However, the measured cleanliness is heavily biased due to several optical and atmospheric effects (see Figure 2). Also, the cleanliness is affected by “cleaning events”, when the mirrors are cleaned. In this case a step appears in the data after the cleaning. And, most importantly, the cleanliness is an “absolute” value that changes with time depending on the applied cleaning schedule, whereas the soiling rate is a relative value representative of particle adhesion independently of the applied experimental protocol. Thus, the dataset cannot be randomized and only a time-series analysis might work on this dataset.

The soiling rate is the time derivative of the cleanliness. In Figure 2 the soiling rate can be seen as the slope of the fitted (red) line. In (Wolfertstetter, 2017) it is stated that the time derivative in 1-min resolution does not make sense because of the following reasons: 1) the cleanliness shows strong fluctuations during the 1-day measurement cycle (can be seen as green “blobs” in Figure 2). 2) The fluctuations are caused by biased due to optical effects in the measurement device.

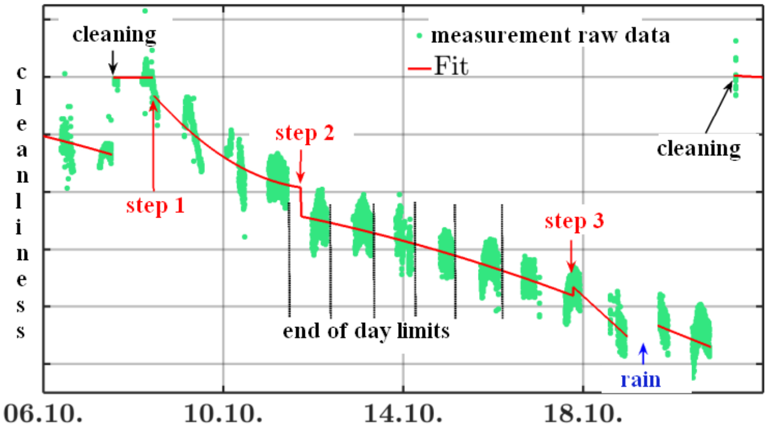


Figure 2: 1-min resolution data set cleanliness measurements (green dots) and the data fit (red curve) in order to create the 1-day resolution soiling-rate dataset. The figure includes 3 cleaning events that cause steps in the fitted curves. Rain effects the cleanliness, too.

b. main dataset with 1-day resolution, for training and validation

To solve the issue of strong fluctuations in the cleanliness values, a new dataset was created from the original: It contains 544 data points in 1-day time resolution. A 1-day soiling rate label was created in (Wolfertstetter, 2017). As stated above the soiling rate is created by fitting a curve through the cleanliness data. The mean slope of the curve is calculated for each day and taken as soiling rate. The curve-fitting process as seen in Figure 2 is fairly complicated as it takes into account a variety of optical and meteorological effects. It is a semi-automated process done by experienced scientists at the DLR.

c. data set with 1-day resolution for testing

Another dataset with 1-day time resolution, with 280 data points is available for testing. It has the same structure as the data set described in b. but is based on data recorded at a different location.

4. Solution statement

As the soiling rate consists of floating point values and required no clasification, the problem to solve is a regression problem. Solving the problem means provide a model that accurately predicts the soiling rate dependent on meteorological features.

After data preparation, advanced regression algorithms should be applied, like Linear Regression, SVM, Decision Trees, AdaBoost and XgBoost, and MLP Regression NN. If available, the feature importance functionality of the algorithm should be used to gather important information on the weather parameters. The data set to use is first the small data set with 1-day resolution and then the big data set with 1-min resolution. In a next step a PCA should be performed in order to gain better understanding about the correlations between features, and features and the label. Maybe the regression algorithms work better on the PCA-transformed data set than with the original features.

As an additional step, if the above-mentioned analysis does not give the desired results, a LSTM recurrent neural network could be applied to the full dataset (with 1-min resolution). A time series prediction also seems appropriate as it was found that some of the features change their correlation with the soiling rate cyclically over the time of 1 year.

5. Benchmark model

As mentioned above, there is a benchmark model available from (Wolfertstetter, 2017), that provides a very good base for comparison. The benchmark model is implemented as linear regression and as evaluation metric the Pearson coefficient (PCC) was used. One of the features showed the biggest correlation with the soiling rate with a PCC = 0.29. Other statistical significant features correlate with the soiling rate with PCCs of 0.25, 0.17 and 0.13 All other features did miss the relative statistical significance threshold (RSS) that has been defined as

$$RSS = \frac{|PCC| \sqrt{\frac{Nsg-2}{1-PCC^2}}}{ST_{p,Nsg}}$$

PCC … Pearson coefficient
N_{sg} … number of data points
$ST_{p,Nsg}$ … value from the 2-tailored student's t-distribution
p … level of significance: p = 98% in this model

Table 1: pair-wise comparison of each feature with the label (feature excerpt)

feature	Pearson Coefficient with the soiling rate	RSS (t-test) 98%
f04	0.29	3.1
f06	-	<1
f08	0.17	>1
f09	-	<1
f10	0.25	>1
f11	-	<1
f12	-	<1
f13	-	<1
f15	-	<1

6. Evaluation Metrics

The Pearson coefficient reveals only linear dependencies and thus cannot be used as indicator for non-linear dependencies. The R2 score seems appropriate for the analysis. Other possible metrics are the explained variance score and the mean squared error.

7. Project Design

First the small dataset with 1-day resolution is used. The idea is, after data preparation, to perform the calculation of the Pearson coefficients of each feature with the label and compare these to the ones found in (Wolfertstetter, 2017). Then a PCA will be performed and compared to the existing PCA from (Wolfertstetter, 2017) as shown in Figure 3. All steps are done using Python, scikit-learn and related packages. The results concerning the Pearson coefficients should be very similar to the results from (Wolfertstetter, 2017).

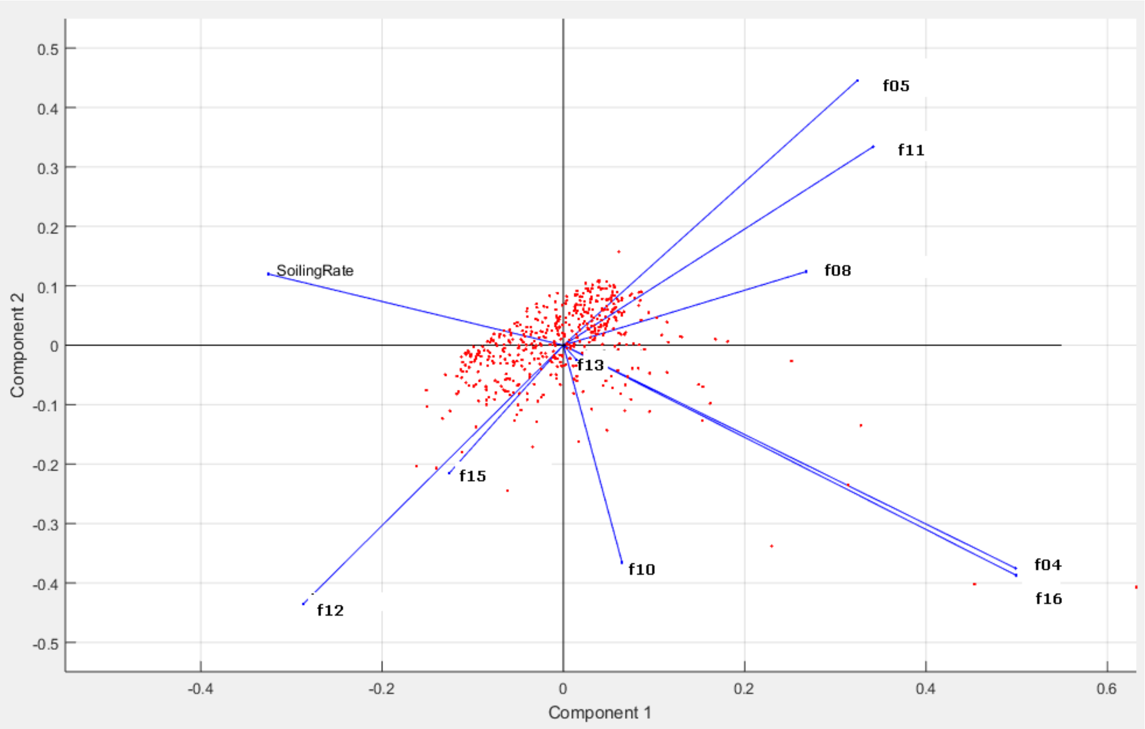
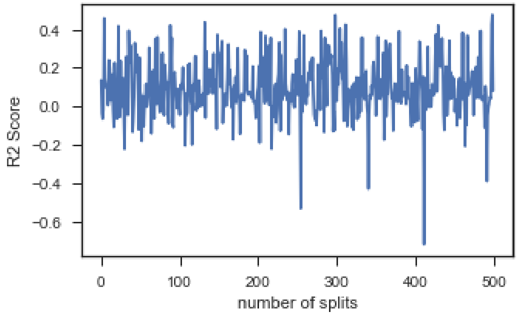
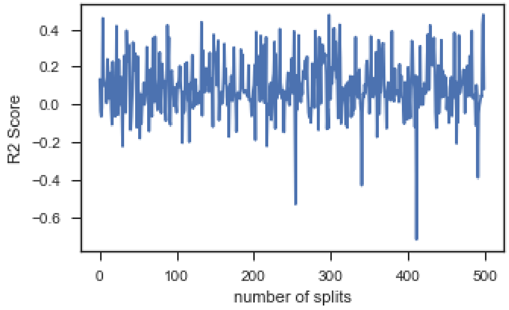
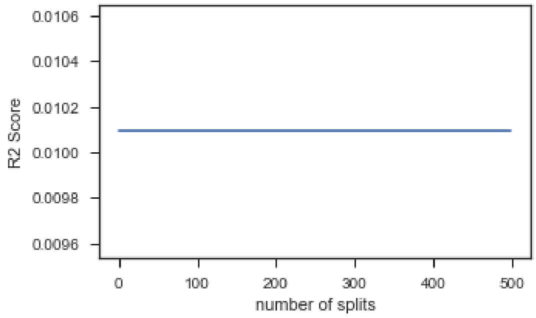
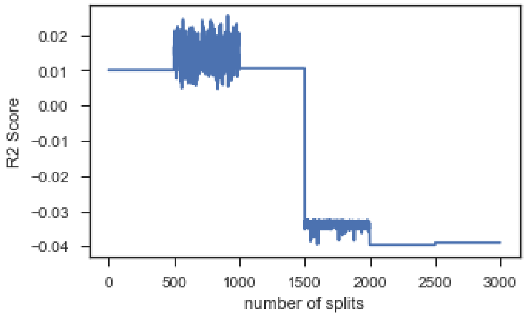


Figure 3: PCA performed by (Wolfertstetter, 2017).

After that the above mentioned supervised learning algorithms will be applied to the data set and the most appropriate one is chosen to predict the soiling rate. In case a learner works well on the data set, this algorithm is then used with k-fold cross-validation and grid search in order to maximize its R2 score on the validation and test data set. Learning curves are printed in order to understand the behavior of the learner.

Pre-tests showed that the algorithms do not reach a good R2 score with the 1-day test set. The figure below shows the R2 scores for Linear Regressor and SVM with rbf kernel. Equal setups were used of each algorithm: First a train_test_split without random_state is performed on the data set and the the sklearn learner is fitted using default settings. The split and fit is repeated and the R2 scores shown in the plot. It is seen that without randomizing the R2 score varies a lot, between -0.7 and +0.4. The mean R2 score is about 0.01. Once the data set is splitted in a randomized way the R2 variance drops significantly, to about 0.01. This indicates a time dependency in the data set and the standard cross-validation (*train_test_spit*) from sklearn does not seem appropriate. A possible solution for the split might be the *TimeSeriesSplit*, that is also provided by sklearn.

Table 2: R2 scores for Linear Regression and SVM regressor. "number of splits" indicates how much the train_test_split and the regressor.fit() functions were executed. If no random sampling is done, the R2 score varies a lot. With random sampling the R2 score goes to 0 basically.

	Linear_Regression()	SVM('rbf')
without random sampling in train_test_split		
with random sampling in train_test_split		

time [hh:mm]	feature1	feature2	...	feature n	label
00:00					calculated soiling rate 1
00:01					
00:02					
...					
23:58					
23:59					
00:00					calculated soiling rate 2
00:01					
00:02					
...					
23:58					
23:59					
			...		
00:00					calculated soiling rate n
00:01					
00:02					
...					
23:58					
23:59					

Figure 4: features with 1-min resolution and corrected, calculated soiling rate 1-day resolution

In a next big step the dataset with 1-min resolution will be used for analysis. If the results are not sufficient using the regression learners mentioned before, a LSTM network is to be built and applied to the big data set. One issue here remains the use of the label. While the cleanliness could be used as label available in 1-min resolution, it is a highly biased measurement value. The use of the calculated soiling rate in 1-day resolution as label and the features in 1-min resolution seems interesting. Figure 4 shows the general structure of such a data set. It is still unclear if this is possible.

8. Citation

(Wolfertstetter, 2017) „Effects of Soiling on Concentrating Solar Power Plants “, Dissertation, 2017, not yet published, in German language

http://www.sciencedirect.com/science/article/pii/S1876610214007115

http://elib.dlr.de/77590/