Solar Panel Under-performance and Daily Yield Predictions based on Weather Forecasts

Daniel Lundstrom

October 2020

Abstract

We consider a data-set describing the performance of 22 solar power plant inverters along with one on-plant weather sensor, reports from these sources given every fifteen minutes over the course of 34 days. We establish a standard of normal performance for each inverter based on irradiation readings using polynomial regression, and define low-performing data-points relative to this standard. Prediction of under-performance is attempted via classification using residual lags, without success, indicating recent history alone is not useful in predicting under-performance of an inverter. Characteristics of the data are identified that strongly indicate under-performance of an inverter is non-random. We then develop a means of predicting total plant daily yield given the irradiation measurements of a day.

1 The Data

DATE_TIME	PLANT_ID	SOURCE_KEY	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD
23-05-2020 07:30	4135001	iCRJl6heRkivqQ3	3199.3750	314.41250	142.37500	7234735
23-05-2020 07:30	4135001	ih0vzX44oOqAx2f	3148.0000	309.37500	145.50000	6240688
23-05-2020 07:30	4135001	pkci93gMrogZuBj	3240.0000	318.47143	146.85714	7225347
23-05-2020 07:30	4135001	rGa61gmuvPhdLxV	2869.5714	281.55714	129.57143	7167271

Figure 1: raw data from power generator sensor

The raw data, originally from an unspecified solar power plant in India, is reported in two parts, the first being reports from a data generation sensor (Fig. 1), and the second being from an on-plant weather sensor (Fig. 2). The data was obtained from kaggle.com and is titled "Solar Power Generation Data". The power generator sensor receives reports from twenty-two different inverters in the plant, which are identified by their unique SOURCE_KEY. Each inverter is hooked up to multiple solar panels and receives power in the form of a direct current, reported as DC_POWER. The inverter converts direct current to alternating current, reported as AC_CURRENT. The inverter produces a yield of energy, reporting the total yield over it's life in TOTAL_YIELD and the day's total yield

in DAILY_YIELD. These reports arrive every fifteen minutes and are reported over the course of thirty-four days. The time and date of the report is reported as DATE_TIME. All data comes from the same plant, with a single common PLANT_ID.

DATE_TIME	PLANT_ID	SOURCE_KEY	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
2020-05-18 01:15:00	4135001	HmiyD2TTLFNqkNe	20.98966	20.38562	0.0000000000
2020-05-18 01:30:00	4135001	HmiyD2TTLFNqkNe	20.98384	20.37683	0.0000000000
2020-05-18 01:45:00	4135001	HmiyD2TTLFNqkNe	21.02701	20.39318	0.0000000000
2020-05-18 02:00:00	4135001	HmiyD2TTLFNqkNe	20.96934	20.33388	0.0000000000

Figure 2: raw data from weather sensor

The weather sensor reports the temperature of the power-plant in degrees Celsius as AMBIENT_TEMPERATURE , the temperature of the weather sensor itself in Celsius as MODULE_TEMPERATURE, and a measurement brightness as IRRADIATION. It reports the SOURCE_KEY of the weather sensor and the PLANT_ID, which are constant. Lastly, the date and time is reported in DATE_TIME.

DC_POWER	AC_POWER	SOURCE	DAY	TIME	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION	YIELD
11600.125	1132.0375	I	7	765	30.86417	51.88567	0.7466439	243.143
7913.429	774.0143	P	7	780	31.39427	54.49144	0.8004696	224.625
7922.286	774.4714	D	7	780	31.39427	54.49144	0.8004696	232.607
8425.429	823.5143	E	7	780	31.39427	54.49144	0.8004696	248.982

Figure 3: data post-processing

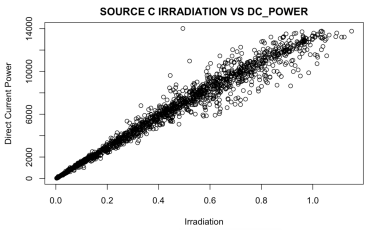
Processing the data, we first attach the relevant weather sensor data to the power generation data, namely AMBIENT_TEMPERATURE and IRRADIATION. We do this by taking every data point of the power generation data set and attaching the temperature and irradiation readings received at that time. We then calculate the 15-minute yield given by the gain in total yield since the last reading of each inverter. We relabel each SOURCE_KEY A-V, relabeling the SOURCE, and separate the day and time, numbering days 1-22 and measuring time in minutes since midnight. We discard irrelevant ID data, and redundant YIELD data. We also discard MODULE_TEMPERATURE as it is irrelevant to the physical function of the inverters and reflects the module alone, while the ambient temperature reflects the environment of the inverters.

We remove all entries where the irradiation is zero as this data does not reflect inverter performance and may skew regression by having a mass of zero-power, zero-yield, zero-irradiation data points. The date amounts to 38,376 data points. There are minor holes in the data; missing data points for various times and inverters, but it seems the weather sensor data is able to account for every time a positive yield is reported by an inverter. We decide to leave the holes in as they are minor and the methods we proceed with are robust against a few missing data points. When calculating yields or lags, data points for an inverter with no temporally prior data point are discarded as the yield cannot be calculated.

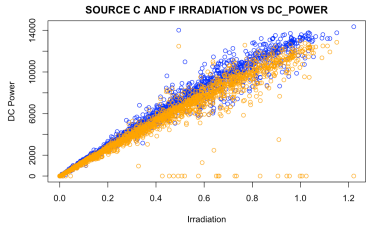
2 Identifying Solar Panel Under-performance

Our first goal is to identify and predict when the solar panels of an inverter under-performs. The benefit of this is to be able to spot declining performance, indicating potentially a need for repair, or potentially finding contributing factors to performance-decline that can be mitigated. If performance-decline can be predicted and we know contributing factors then we can potentially prevent performance decline and increase productivity.

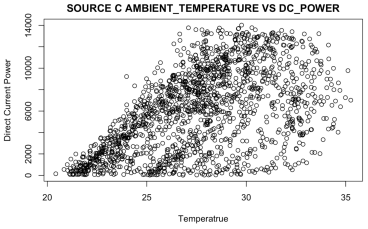
To measure the performance of an inverter, we will use DC_POWER as opposed to AC_POWER or YIELD. DC_POWER is the most fitting to use because it is the power directly fed into the inverter by the panels and represents the contribution of the solar-panels to YIELD. AC_POWER is a function of the performance of the solar panels and the inverter itself, as the ability of the inverter to convert DC to AC is a factor. YIELD is also a function of the inverter, as the inverter may consume some of the DC power to function, and convert the rest to yield.



To measure performance we seek to establish a normal level of DC_POWER, and from this standard of normalcy, identify those that perform significantly below the norm. Our model is that DC_POWER = f(IRRADIATION, AMBIENT_TEMPERATURE, SOURCE). This model is justified from a physics standpoint.Solar panels harvest energy in the form of light rays, converting it to solar power. The more light, the more energy gathered. Temperature is included because it is common for many engineered objects to exhibit error or under-performance when they overheat, and perform best at cool temperatures. This may be the case for solar panels at the temperature ranges we are considering. We do not include data from the other inverters because they are connected to a different circuit and their panels receive different light rays, i.e. their function should not effect the function of the solar panels of the inverter in question.



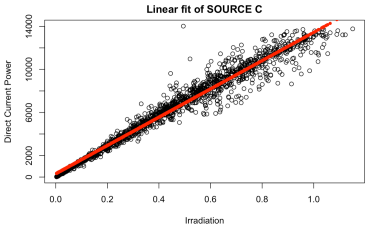
A plot of IRRADIATION vs DC_POWER at the top right shows that the relationship is highly linear. This is expected, since the purpose of solar panels is to harvest energy from sunlight. The more sunlight, the higher the IRRADIATION reading, the more DC_POWER the panels will be able to generate. In the figure above, IRRADIATION vs DC_CURRENT is given for SOURCE C in blue and SOURCE F in yellow. We can see that source C solar panels perform differently than source F, either because they are less efficient, or because more, larger, or more efficient solar panels are attached to C than to F. Plotting AMBIENT_TEMPERATURE vs DC_POWER, we fin that this looks much less linear, with a very large degree of variation in the moderate and higher temperatures.

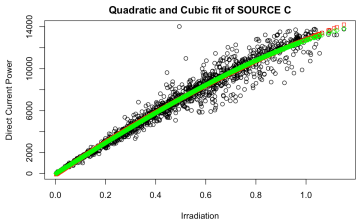


We begin with a linear regression of DC_POWER, and gain the results:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          385.286    167.872   2.295   0.0219 *
IRRADIATION        13156.513     73.452 179.118   <2e-16 ***
AMBIENT_TEMPERATURE    -3.171      6.584  -0.482   0.6301
```

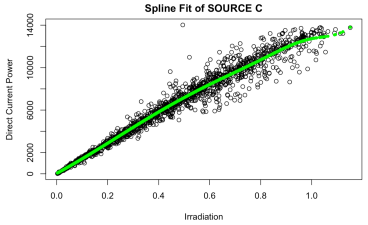
Interpreting the results we see that irradiation is highly significant in predicting AC power, with a significantly large coefficient, while temperature is insignificant. The average R^2 value of the linear fit over all sources is 0.972, which is very high. However, in spite of this, there are some reasons why this fit seems unsatisfying. The first is that the plot is obviously not linear; it has a concave down trend that is seen especially in the higher irradiation values. The second reason is that graphically there are some signs that this is not a correct fit. The fit is so biased for low irradiation values that over-predicts every point with irradiation under than 0.1. Graphically, you can see it under-predicts a dense bulk of values between 0.3 and 0.6 that are the main trend, graphically seen as a large black "hump" sticking above the red line. We seek to capture the slight concavity of the data, so we look at higher fits.

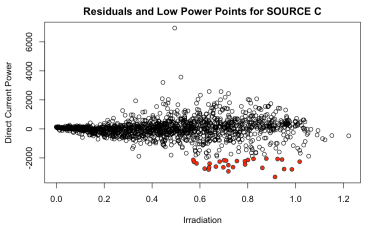




We explore polynomial regressions to fit DC_power, with polynomials in irradiation up to degree 8, quadratic factors in temperature, and a temperature*irradiation factor in case temperature has greater effect when it is brighter. We find that any term including temperature has a nominal effect on the fit, as does any polynomial fit of IRRADIATION higher than 2. For instance, the the the left fits irradiation with polynomials of degree 2 in red and 3 in green. The average R^2 value over 22 inverters for the quadratic fit is 0.9762, while the same for the cubic fit is 0.9764. The 3 polynomial fits almost directly on top of the red one and boasts nominal R^2 gains. We conclude it seems that a polynomial regression fit of DC_POWER using IRRADIATION of degree 2 is the most preferable of the polynomial fits. This is because it resolves our issue of not taking into account the concavity of the data and resolves the graphically seen issue of not well-fitting bulk-trends in the data distribution. An additional reason to settle on the quadratic fit is that no higher fit here boasts the significant R^2 gains either.

We consider that the data may not be quadratic, and, for this project, we look to practice cubic spline fitting and evaluate it's results, so we attempted a cubic spline fit. We try an irradiation vs DC power spline fit with knots at 0.8 to capture the bend for high irradiation; knots at 0.4 and 0.8; and a third fit with knots at 0.4, 0.6, and 0.8. This gets an average R^2 value of 0.9766, which is a nominal gain. In addition, the fit looks a bit forced compared to the quadratic fit graphically, so we end up keeping the quadratic fit.

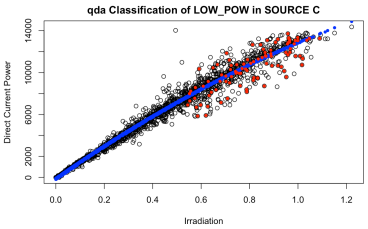




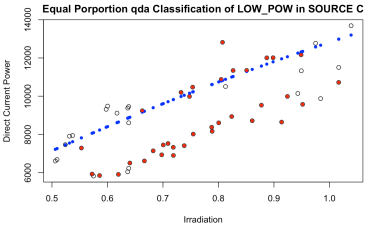
To identify points in the data that under-perform, we accept the quadratic regression as a base-line and consider the residuals. Computing the variance of the residuals, any point that is at least three standard deviations below zero we will consider a under-performing point, or what we will call a low power point. This is illustrated to the right, where we plot residuals and mark low power points in red.

3 Predicting Low Power Points

Our strategy for predicting low power points is to use the recent performances of DC_POWER. We create 5 lag variables that represent the residuals of the previous five DC_POWER reports for that inverter. These five lags must be of times immediately prior temporally to the data-points itself; any time-gaps we report as zero. We also make an indicator variable IS_LOW with a value of "YES" if the point was classified as a low power point, and "NO" otherwise.

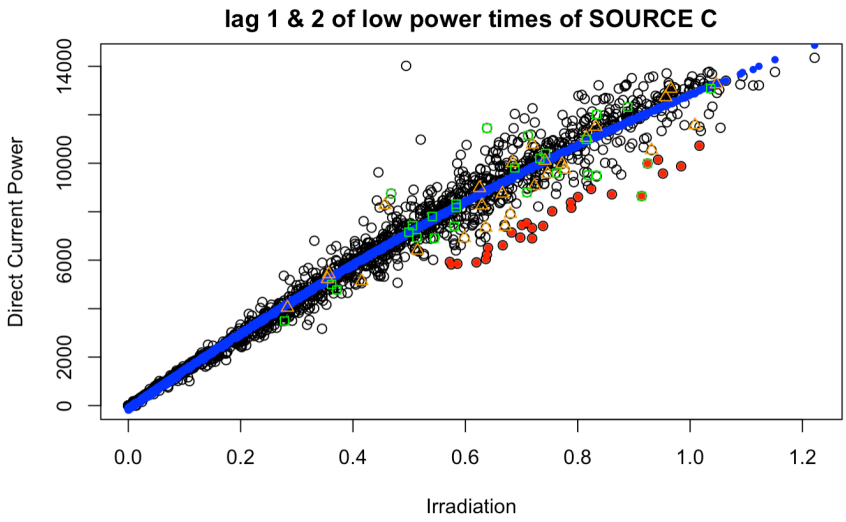


We then apply various classification techniques on the IS_LOW variable using the lags as predictors. Logistic regression classified all points as "NO". I believe this is due to the small number of data points marked "YES", as there are an average of 21 data-points classified "YES" for an inverter, and and average of 1653 data-points total per inverter. LDA did not perform much better, classifying few points as "YES", and most incorrectly. QDA identified far too many points as "YES" and most were incorrect, while KNN with k=3 and no removal would correctly identify zero to a few points as "YES", while k=5 identified none as "YES".



To investigate further whether low power power points have any difference in lag distribution, we repeated the above tests with an equalized portion of low power points and non-low power points. We also only considered points where the irradiation is greater than 0.5, as that is where the low power points are located. For logistic regression, lda, qda, and knn, we get classification error rates of 0.68, 0.63, 0.74, and 0.76, respectively. Overall, it appears that the lags of the low power points may be somewhat different from the average point, but not enough to be easily distinguished. Although we have a correct classification rate of up to 0.76 after equalizing the proportions, this is consistent with low classification rates for normal proportions, and does not indicate we can predict low power points with any reasonable success.

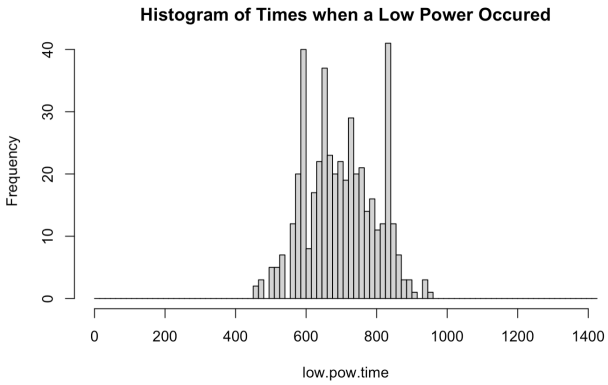
I believe that the low ratio of low power points to total points contributed to the failure of these methods for equal proportion. Investigating the failure of these techniques, I plotted lag 1 and lag2 besides the low power points, with lag1 points in orange, and lag2 points in green. It appears that most of the lag points do not "dip" low, or have lower



residuals, but instead most have average or above average performance. This would indicate that the methods failed because low performance is not strongly indicated by previous performance. This is further supported by not outstanding classification results when the proportions were equalized. The above plot also suggests that the performance of solar panels is not declining in time, for if that were so, low power points would be significantly likely to have poor performing lags, which is not the case.

4 Contributing Factors to Low Performance

Although our attempts to classify low power points via lag did not produce significant results, there are significant reasons to believe that the distribution of low-power points is neither uniform nor dispersed. To the right is a histogram of the times when a low power point occurred. At 10:00 AM, 11:00 AM and 2:00 PM, an improbably large number of low power points accrue, indicating that some sort of phenomenon occurs at these times. At first it appears unlikely that cause is weather related because weather is not that consistent in it's schedule. However, perhaps the sun gets to a certain angle in the sky at these times, causing a fluctuation in function of the solar panels if they were at a fixed angle. Perhaps the solar panels rotate depending on the time of day, and this fluctuation indicates that.



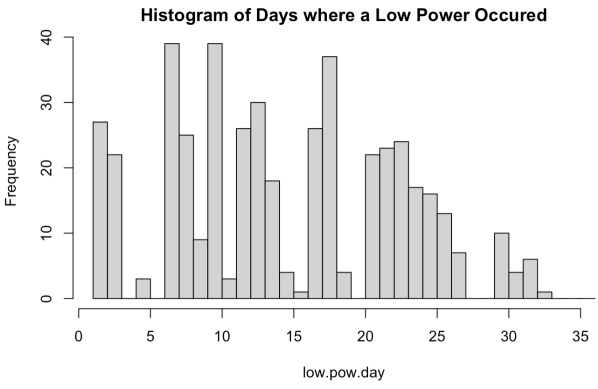
To further investigate this, we investigated the peak low-power times of 600, 660, and 840, the three outstanding times indicated by the histogram. For the time of 600, we found the residual differences after quadratic fitting for every time, then we compared the

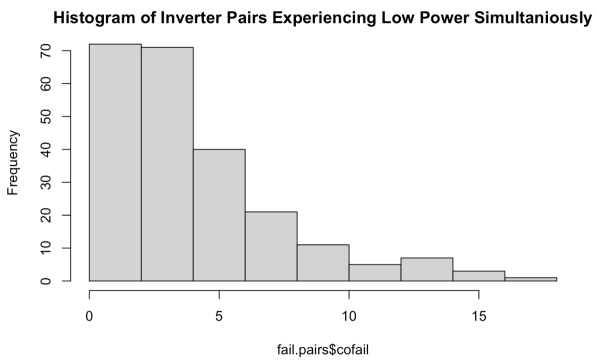
residuals for 600, the time immediately before, and the time immediately after. We did this procedure for each of the three outstanding times. If either the fixed angle or scheduled adjusting hypotheses were true, they hypothesize a physical circumstance that causes a certain time to perform lower than the times around it. In this case, there should be a dip in the residuals going into that time and a rebounding of the residual after. We took an average of the difference between the prior time and predicted dip time, and again the future time and the predicted dip time, and evaluated whether they had a tendency to be negative.

For the 600 time we found that there was a significant tendency for the time-before residual difference to be negative, with an average difference of -205, but a less strong tendency for the time-after residual difference to be negative with a average difference of -72. This indicates that generally, the 600 time has a dip in efficiency. Most sources in the time-before comparison had negative average residual differences, while many in the time-after had positive residual differences, suggesting that perhaps some solar panels are affected by the scheduled affect, but not all. For the 660 time, the residual differences were 63 and 88, indicating that the DCat that time performed at above average efficiency at that time compared to directly before and after. Lastly the 840 time had residual difference values of -94 and -102, indicating a dip in efficiency.

Based on this information, it looks like some sort of times interruption in efficiency happens at times 600 and 840, while the 660 time was an anomaly. This result supports our hypothesis of fixed angle panels with an critical angle, but may or may not support our hypothesis of adjusting solar panels. If the panels are fixed, then there would likely be two critical angles, one when the sun moves into a prime angle, and another when the sun sets out of a prime angle, which is what we found. If the panels adjusted their angle, we would need to know if we should expect more than two adjustments, and thus more than two dips, throughout the day, or if two would suffice. I expect more than two adjustments are necessary, since the angle difference between the 600 and 840 times is a difference of 4 hours. Since India is near the prime meridian, the angle difference is approximately $\frac{4\text{hours}}{24\text{hours}} * 360^\circ = 60^\circ$, a larger angle.

To the right we have a histogram illustrating the frequency with which a day had low power points. There are multiple days with a high frequency of low power points and multiple days with no low power points. If wither of the phenomenons hypothesized in the previous paragraph did occur, it seems unlikely that they would skip some days. We know that low power points are less likely to occur during low irradiation levels; perhaps the days with no low power points were days with low irradiation. These question are not addressed here, but could be investigated up in a following paper.





Finally we enumerated all possible pairs of inverters and counted the number of times those two inverters had low power points on the same time of the same day, and produced a histogram of the results. For comparison, the average number of low power points is 21 per inverter. The histogram indicates that a significant number of pairs of inverters have low power points together often. One pair of inverters had simultaneous low power points eighteen times, which is extremely unlikely if their performances were uncorrelated. Investigating further, it was discovered that inverters B,C,D, and E had simultaneous low power points eight time. Taking the correlation of the residuals of these four inverters, each correlation pair is above 0.7, while their correlation to other inverters can be quite low or even negative. This suggests that whatever phenomenon that causes a decline in performance is affecting groups of them consistently, but not others.

There are further problems in interpretation due to the ambiguity of the sensor data, namely, is an irradiation measurement a report of what the sensor experiences at the moment of reporting, or an average over the fifteen minute period? If it is a measurement at that moment, it may be wildly different from the time in-between measurements. For instance, a cloud could move over the plant in the 15 min interval between reports, then leave before the reading. This could explain some low power points. Also, where is the irradiation sensor compared to the inverters? The irradiation measurement reported by the inverter will reflect experienced irradiation of closer inverters more accurately that inverters further away. Having this data would help us interpret and explain the distribution of low power points.

5 Predict Yield given Weather Data

Our second goal is to predict a days yield given its weather data. This has the obvious use of taking weather projections and extrapolating power plant yield projections. Our model is as such: let Y_d be the plant yield on day d, and Y_{di} be the yield on day d for inverter i. Then:

$$Y_d = \sum_{i \in inverters} Y_{di}$$

Now denote the yield on day d, time t, of inverter i as y_{dit}. We can write

$$Y_{di} = \sum_{t \in times} y_{dit}$$

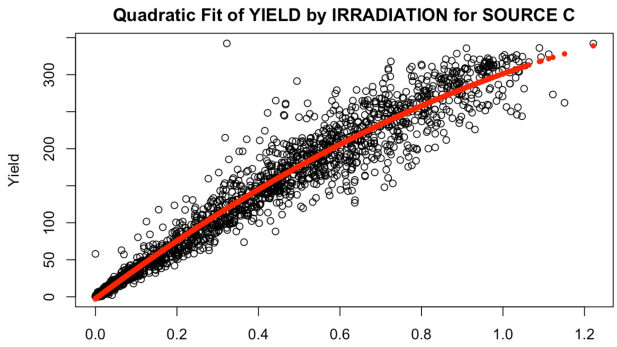
Which leads to:

$$Y_i = \sum_{i \in inverters} \sum_{t \in times} y_{dit}$$

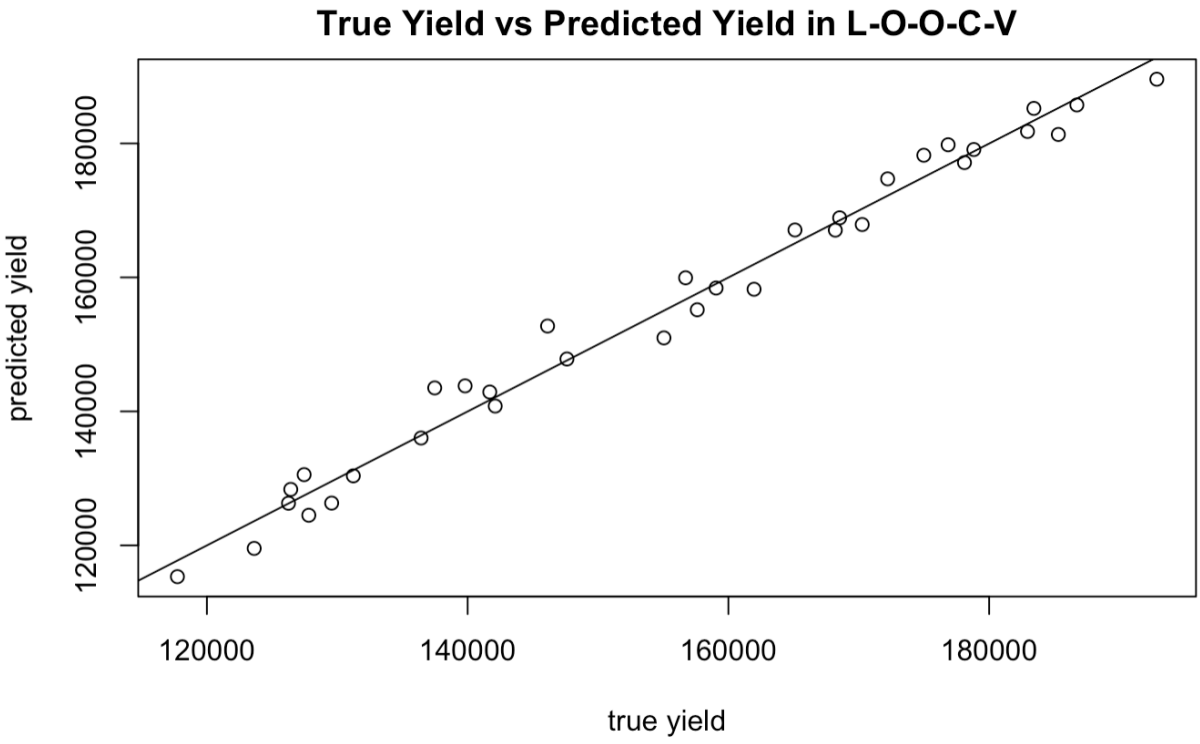
To predict the yield of an inverter based on irradiation and temperature, we will use a fitting method. Letting $(IRR, TEMP)_{dt}$ denote the weather report for a given day and time, we will fit y_{dit} with IRRADIATION and AMBIENT_TEMPERATURE to produce an approximation $f_i((IRR, TEMP)_{dt})$. Thus the predictor will have the form:

$$\begin{aligned}\hat{Y}_i &= \sum_{i \in inverters} \sum_{t \in times} \hat{y}_{dit} \\ &= \sum_{i \in inverters} \sum_{t \in times} f_i((IRR, TEMP)_{dt})\end{aligned}$$

To produce f_i, we notice YIELD has a highly linear relationship to DC_POWER. It's relation to IRRADIATION and AMBIENT_TEMPERATURE is very similar to DC_POWER's. After some investigation similar to that in the above regression problem, our model $f_i((IRR, TEMP)_{dt})$ is chosen to be a quadratic regression in IRRADIATION that is unique for each inverter.



To test the effectiveness of our model, we use leane-one-out cross-validation. We remove one day but save it's weather and yield data. We then train our models on the other 33 days, feed the model our removed day's weather stats, and use that as a predictor for the true yield of the day. The R^2 value for this model and validation method is 0.983.


True Yield vs Predicted Yield in L-O-O-C-V
predicted yield
120000
140000
160000
180000
true yield
120000
140000
160000
180000