Predicting Solar Power Generation

Alexander McIntosh

08/08/2022

Contents

1 Introduction

As the unintended consequences of traditional energy collection become untenable, the world is turning to alternative, renewable energy sources. Solar power has become a particularly popular method of energy extraction. However, a notable drawback of solar power is its lack of day-to-day consistency. Inconsistency is a significant challenge for interfacing with an external power grid.

In the analysis to follow, using only the data collected at the solar power plant, the power output of the plant will be predicted based only on data available to the plant. First, suspicious data are examined for usability. Then, the clean data are explored for patterns, correlation, and causality. Finally, four models are engineered using linear regression, ARIMA, and random forest machine learning. The success of each model was measured by the model's ability to predict three days worth of data, separated before the models' construction. The metric for success was the Root Mean Squared Error, or RMSE. The lower an RMSE value, the better the predictive capabilites of the model.

For brevity and readability, much of the code and many of the data transformations used on the data to produce graphs are not included in this report. For more information about how the graphs and results were generated, see https://www.github.com/mcintalmo/solar-plant-power-generation. The link includes the R Markdown that generated this report and the notebook used for exploration.

Table 1: First 5 Observations Of Power Generation Data

date_time	plant_id	source_key	dc_power	ac_power	daily_yield	total_yield
2020-05-15	4135001	1BY6WEcLGh8j5v7	0	0	0	6259559
2020-05-15	4135001	1IF53ai7Xc0U56Y	0	0	0	6183645
2020-05-15	4135001	3PZuoBAID5Wc2HD	0	0	0	6987759
2020-05-15	4135001	7JYdWkrLSPkdwr4	0	0	0	7602960
2020-05-15	4135001	McdE0feGgRqW7Ca	0	0	0	7158964

To explore the raw data, see *Solar Power Generation Data* from user Ani Kanal on Kaggle.

https://www.kaggle.com/datasets/ef9660b4985471a8797501c8970009f36c5b3515213e2676cf40f540f0100e54

Or use command

```
kaggle datasets download -d anikannal/solar-power-generation-data
```

2 Methods and Analysis

This analysis was conducted entirely within R. Packages included tidyverse and broom for data organization and transformation. Lubridate and hms allowed for time stamp transformation. Tsibble, forecast, fable, and feasts provided the backbone of analysis of the data as a time series and use of the ARIMA model. Ranger, randomForest, and caret all contributed functions for developing a machine learning model. kableExtra and bookdown were used only in the report to generate more aesthetically appealing tables and visualizations.

2.1 Data Collection

The data for this analysis was collected by two solar plants in India. The first plant, id: 4135001, is near Gandikotta, Andhra and the second plant, id: 4136001, is near Nasik, Maharashtra. The data were collected at 15 minute intervals for 34 days. 22 inverter sensors and one weather sensor at each plant brought the total to 46 sensors collecting data.

The inverter sensors collected the direct current (DC) and alternating current (AC) output of the group of solar panels it was monitoring. Additionally, the inverters tallied the daily yield in DC output from midnight to midnight and maintained the total yield the sensor had observed. Total yield increased until a sensor was replaced, at which point it would begin again at 0.

The weather sensors recorded the ambient temperature, the temperature of the module, and the irradiation level. Owing to the sensors' reading of 0 at night, it is assumed any irradiation is owed entirely to the sun's radiation.

2.2 Data Cleaning

2.2.1 The Raw Data

The data are spread out over four comma separated files; two power generation files and two weather sensor files, one each for each plant. First, the data are read into memory, using lubridate to parse the time stamp, coercing the plant_id to a character, and renaming variables for usability. Note, one of the power generation files has a different date format than the other three files. This suggests human input or conversion was involved in the collection of the data.

2.2.2 Loading the Data

The generation data consist of 136476 observations of 7 variables: The time stamp, the plant id number, the inverter key, the measured DC power from the panels, the measured AC after inversion, the daily yield of

Table 2: Power Generation Source Keys

Plant ID: 4135001	Plant ID: 4136001
1BY6WEcLGh8j5v7	4UPUqMRk7TRMgml
1IF53ai7Xc0U56Y	81aHJ1q11NBPMrL
3PZuoBAID5Wc2HD	9kRcWv60rDACzjR
7JYdWkrLSPkdwr4	Et9kgGMDl729KT4
adLQvlD726eNBSB	IQ2d7wF4YD8zU1Q
bvBOhCH3iADSZry	LlT2YUhhzqhg5Sw
iCRJl6heRkivqQ3	LYwnQax7tkwH5Cb
ih0vzX44oOqAx2f	mqwcsP2rE7J0TFp
McdE0feGgRqW7Ca	Mx2yZCDsyf6DPfv
pkci93gMrogZuBj	NgDl19wMapZy17u
rGa61gmuvPhdLxV	oZ35aAeoifZaQzV
sjndEbLyjtCKgGv	oZZkBaNadn6DNKz
uHbuxQJl8lW7ozc	PeE6FRyGXUgsRhN
VHMLBKoKgIrUVDU	q49J1IKaHRwDQnt
wCURE6d3bPkepu2	Qf4GUc1pJu5T6c6
WRmjgnKYAwPKWDb	Quc1TzYxW2pYoWX
YxYtjZvoooNbGkE	rrq4fwE8jgrTyWY
z9Y9gH1T5YWrNuG	V94E5Ben1TlhnDV
zBIq5rxdHJRwDNY	vOuJvMaM2sgwLmb
ZnxXDlPa8U1GXgE	WcxssY2VbP4hApt
ZoEaEvLYb1n2sOq	xMbIugepa2P7lBB
zVJPv84UY57bAof	xoJJ8DcxJEcupym

Table 3: First 5 Observations Of Weather Data

date_time	plant_id	source_key	ambient_temperature	module_temperature	irradiation
2020-05-15 00:00:00	4135001	HmiyD2TTLFNqkNe	25.2	22.9	0
2020-05-15 00:15:00	4135001	HmiyD2TTLFNqkNe	25.1	22.8	0
2020-05-15 00:30:00	4135001	HmiyD2TTLFNqkNe	24.9	22.6	0
2020-05-15 00:45:00	4135001	HmiyD2TTLFNqkNe	24.8	22.4	0
2020-05-15 01:00:00	4135001	HmiyD2TTLFNqkNe	24.6	22.2	0

power, and the total yield that the sensor has ever recorded. Each plant has 22 units measuring the inverter variables, shown in table 2

The weather data consist of only 6441 observations of 6 variables: The time stamp of the observations at 15 minute intervals, the plant id at which the observation was taken, the weather sensor key (identical across the plant), the ambient temperature in Celsius, the module temperature in Celsius, and the irradiation. The difference in the number of observations owes to the number of sensors. Each plant has only 1 sensor recording weather data, shown in table 4.

2.2.3 Sanity Check

First, let us take a look at a visualization of the generation data by plant.

Going by figure 1, it may be seen that the distribution of DC Power observations is considerably more spread out for plant 4135001 than 4136001.

In table 5, it is shown that the max and mean values differ by a factor of 10. Over half of the recorded observations at plant 4136001 are 0. A median of 0 is not unreasonable, if we expect the sun to be down for

Table 4: Weather Data Source Keys

Plant ID: 4135001	Plant ID: 4136001
HmiyD2TTLFNqkNe	iq8k7ZNt4Mwm3w0

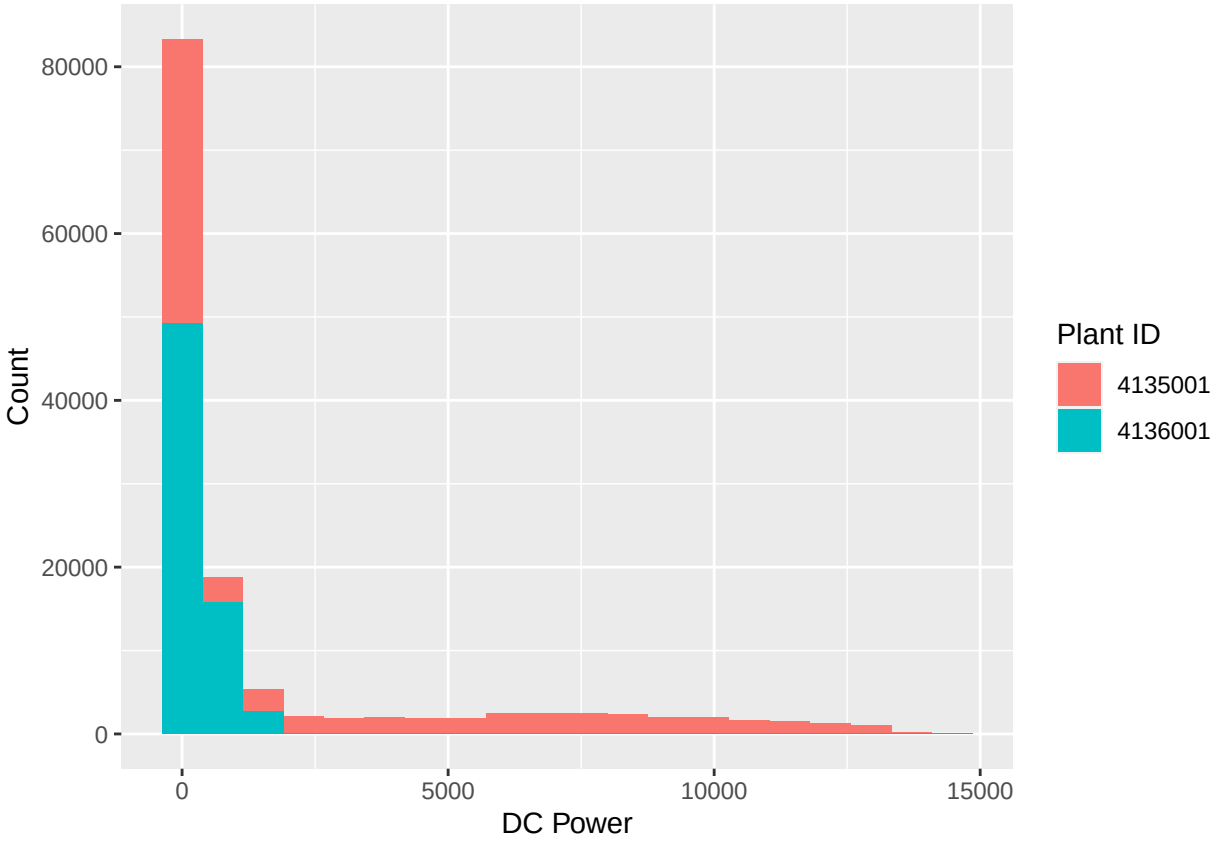

Figure 1: DC Power Distribution by Plant

Table 5: DC Power Summary by Plant

plant_id	min_dc	max_dc	median_dc	mean_dc	sd_dc
4135001	0	14471	429	3147	4036
4136001	0	1421	0	247	371

more than half of the day.

However, these data were collected in May in India, which resides in the Northern hemisphere. It would be expected that the sun is out for more than half of the observations. To better understand what might be happening, let us compare the DC output of the solar panels to the AC output after conversion in figure 2.

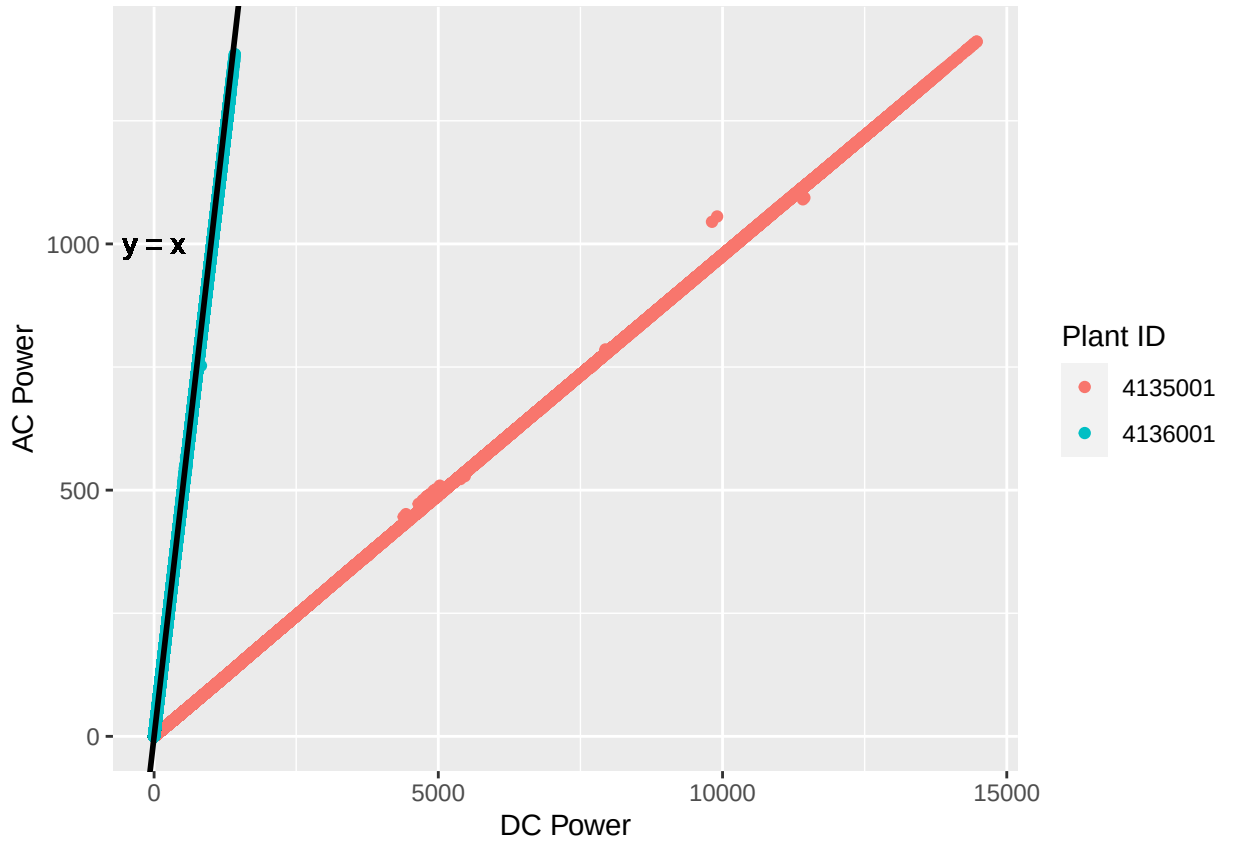

Figure 2: DC to AC Power Conversion

Plant 4136001 is reporting that the conversion rate from DC power to AC power is about %100. However, 4135001 is reporting a 90% loss of power during conversion, despite collecting 10 times as much power. Even the worst DC to AC adapters are about 80% effective. This suggests that a conversion error changed the DC Power variable from plant 4135001 by a factor of 10. The conversion error is assumed for the remainder of the analysis, and the DC power values for plant 4135001 are divided by 10.

2.2.4 Missing Observations

As noted previously, there were 136,476 power generation observations made. However, if an observation is made every 15 minutes at 44 sensors for 34 days, we expect to see 143,616 observations.

The remaining 7,140 observations are missing, and will be considered NA for time series analysis. Can a pattern to the missing values be found? To start, we will fill in the missing time values and combine the weather data.

Figure 3 shows the number of missing values from power sensors over time. Notably, missing values at plant 4135001 seem to go across all 22 sensors. Additionally, For much of the last week of may, plant 4136001 had about 4 power generation sensors that were offline.

Table 6: First 5 Observations in Solar Data

plant_id	date_time	generation_source	dc_power	ac_power	daily_yield	total_yield	weather_source
4135001	2020-05-15	1BY6WEcLGh8j5v7	0	0	0	6259559	HmiyD2TTLFNqkN
4135001	2020-05-15	1IF53ai7Xc0U56Y	0	0	0	6183645	HmiyD2TTLFNqkN
4135001	2020-05-15	3PZuoBAID5Wc2HD	0	0	0	6987759	HmiyD2TTLFNqkN
4135001	2020-05-15	7JYdWkrLSPkdwr4	0	0	0	7602960	HmiyD2TTLFNqkN
4135001	2020-05-15	adLQvlD726eNBSB	0	0	0	6271355	HmiyD2TTLFNqkN

Table 7: Solar Data Dimensions

Observations	Variables
143616	11

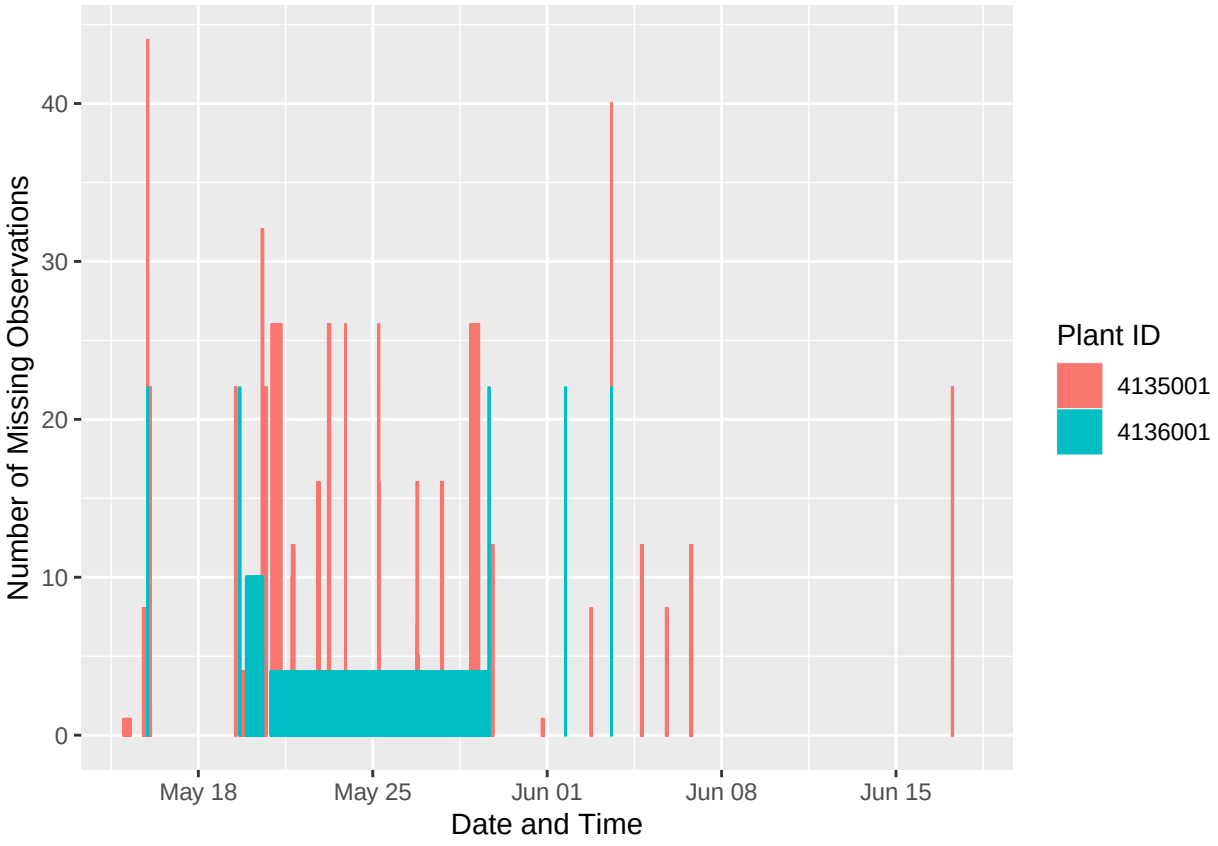

Figure 3: Time Series of Power Inverter Failures

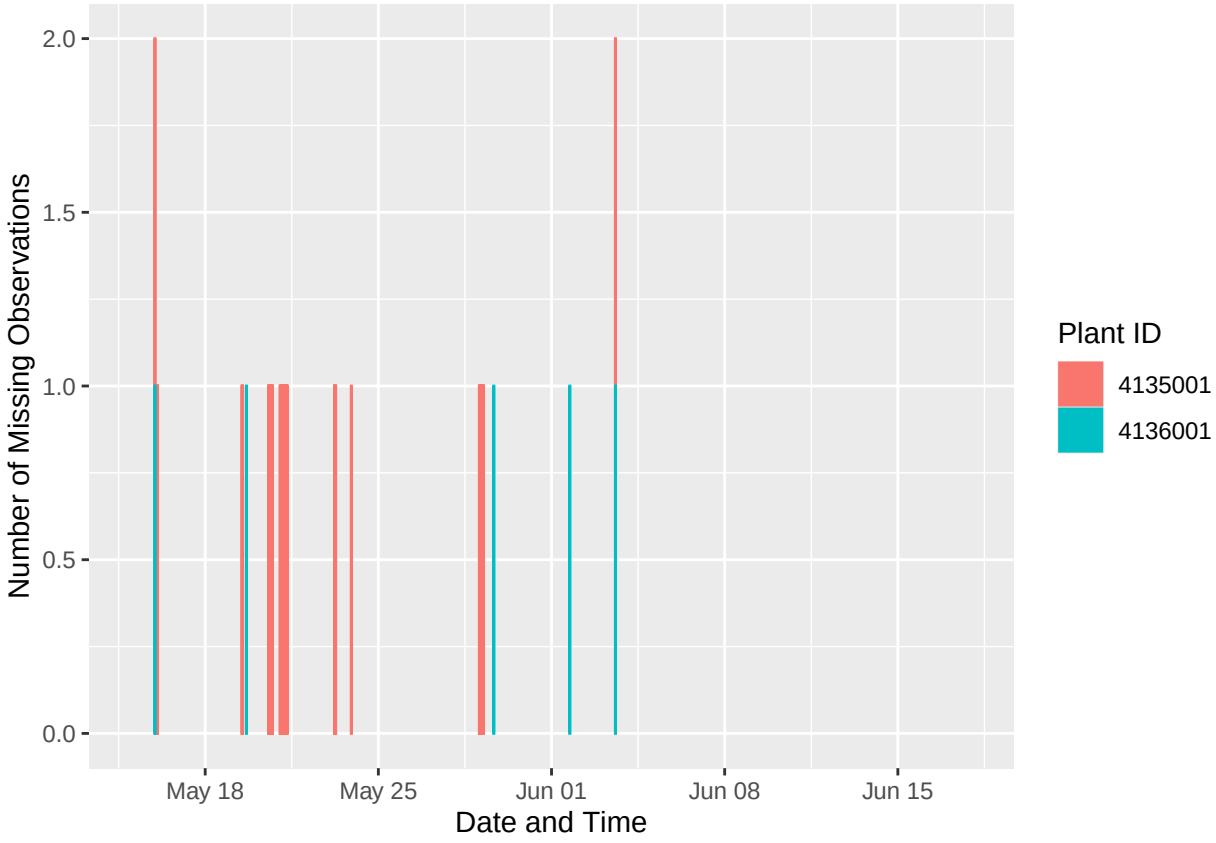

Figure 4: Time Series of Weather Sensor Failures

Figure 4 shows the number of missing values from power sensors over time. Note that many of the missing values of one or both weather sensors align with missing values of the generation sensors as well. This might suggest maintenance or a failure of a kind that affected the entire plant.

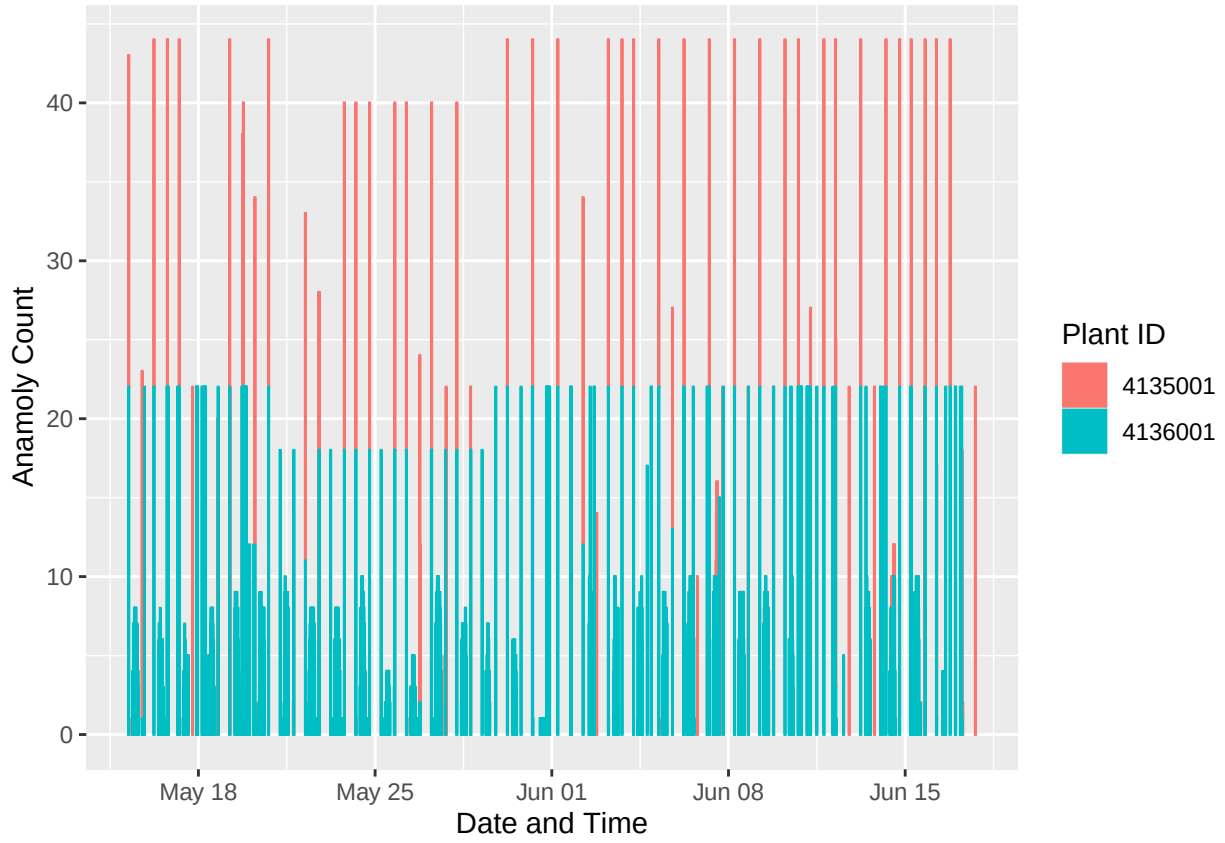

Figure 5: Sensor Aanomalies Over Time

Figure 5 displays anomalies in the sensor data. Here, an anomaly is defined as an observation in which the irradiation measured by the weather sensor is greater than 0, but no DC power is generated. This is incredibly unlikely to happen in a functioning solar cell. Without more information, it is hard to tell exactly what is happening in these cases. For example, is it true that the panels are not actually producing DC power, or is it the case simply that the power sensor is not detecting the generated power.

It is time to get more granular. Are there specific sensors that might be contributing disproportionately to missing and anomalous data?

Figure 6 shows the total number of combined anomalies and missing observations broken down for each source. It is clear that plant 4136001 experiences many more problems, with four sources contributing many problems. Even the best 4136001 sensor performs worse than the worst 4135001 sensor.

Figure 7 clearly shows that plant 4136001 has many more anomalies occurring. Most sensors are reporting 2 to 4 times as many anomalies.

Figure @{fig:combined-missing} shows an interesting phenomena. Plant 4136001 has the worst offending sensors by a large margin. Nearly 4 times as many observations were missed by 4 outstanding sensors. The rest of the sensors at plant 4136001 have missed relatively few observations, nearly none in many cases. What could be causing this discrepancy? Let us explore if there is a relationship between missing observations and anomalies.

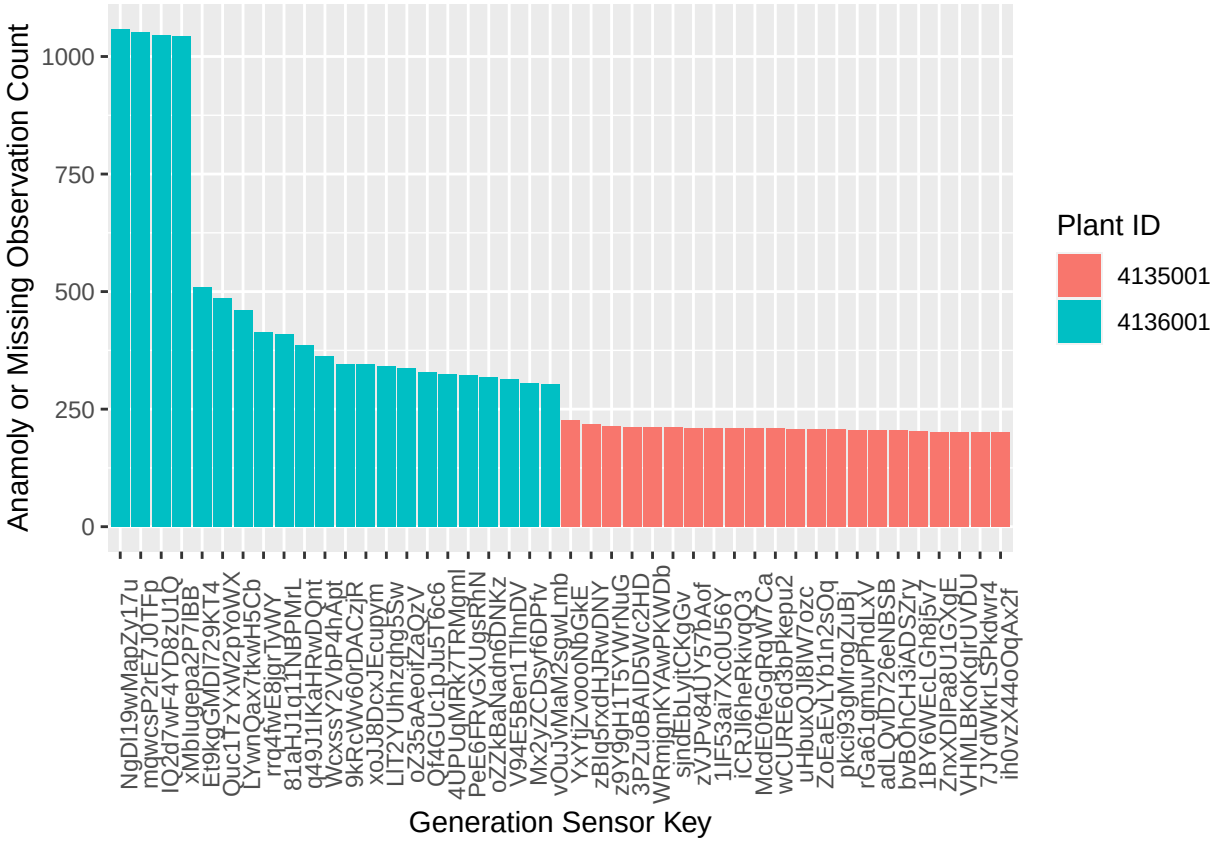

Figure 6: Combined Sensor Anamolies and Missing Observations by Source

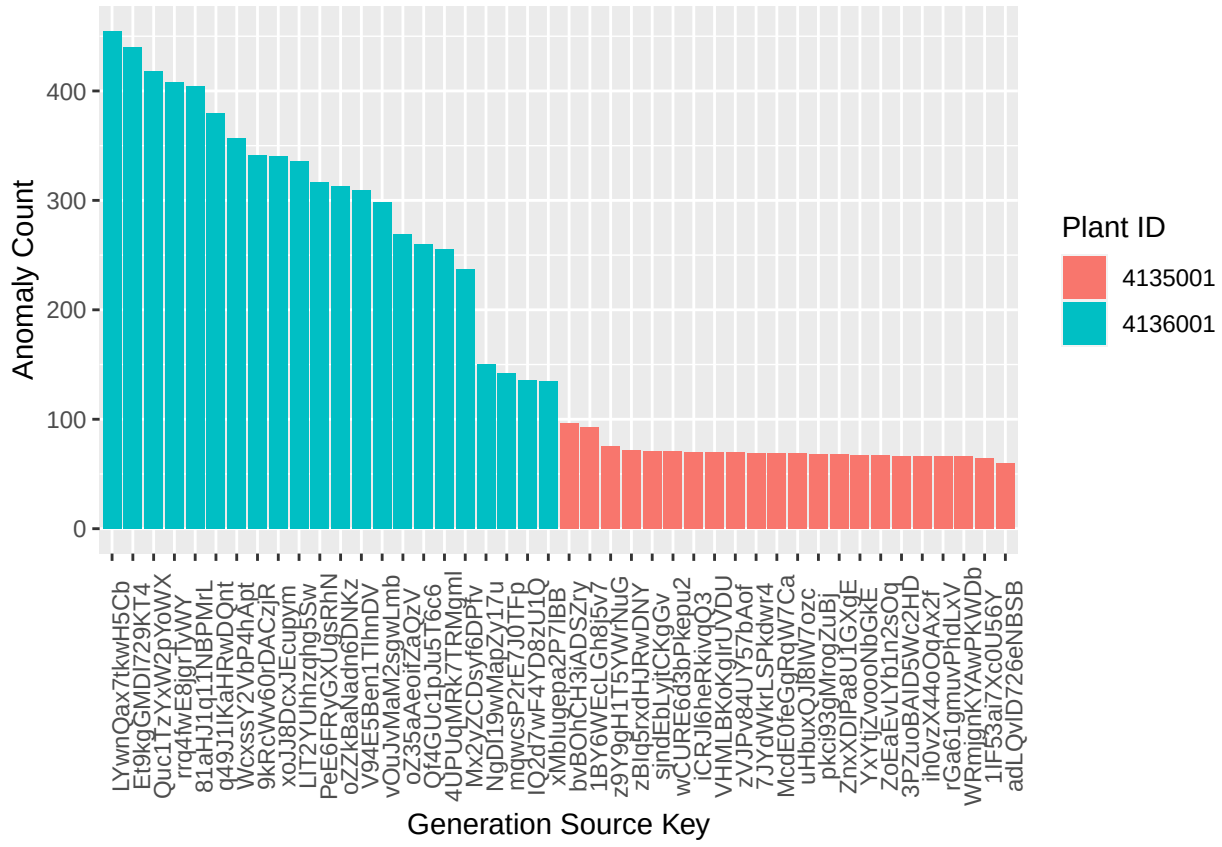

Figure 7: Anomaly Count Per Sensor

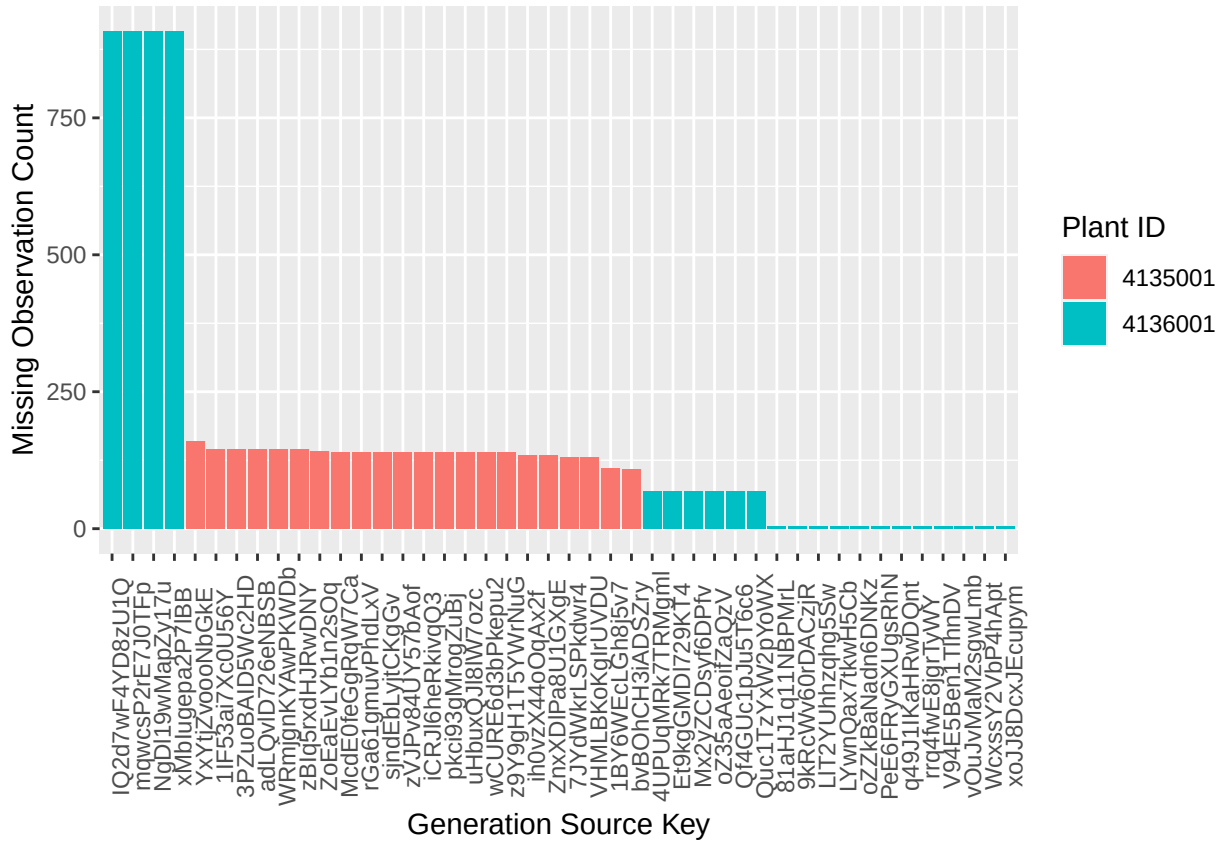

Figure 8: Combined Missing Observations Per Sensor

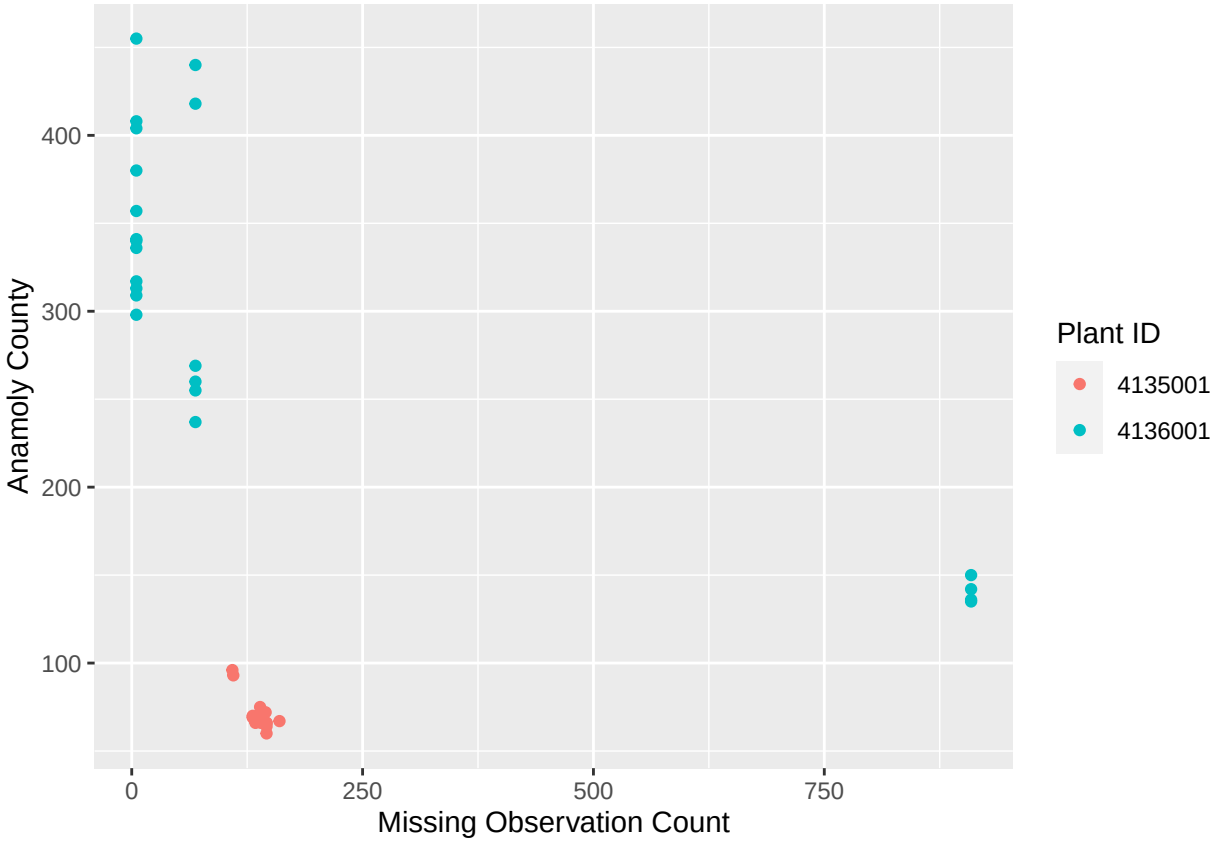

Figure 9: Relationship Between Sensor NA and Anamoly

Figure 9 shows the sensor grouping as a result of relating missing counts to anomaly counts. The plant since 4136001 cluster shows a healthiest collection of data, while plant 4135001 clusters suggest that less missing data from a set does not necessarily mean that the data is more trustworthy.

That is, as the missing observation count decreases at plant 4136001, the number of anomalies increases. This relationship suggests that despite an observation being made, one or more values were corrupted. Only in the case of a very high number of missing values does the number of anomalies approach the numbers of plant 4135001.

To increase the understanding of the value of data collected at both plants, let us now turn our attention to the total yield.

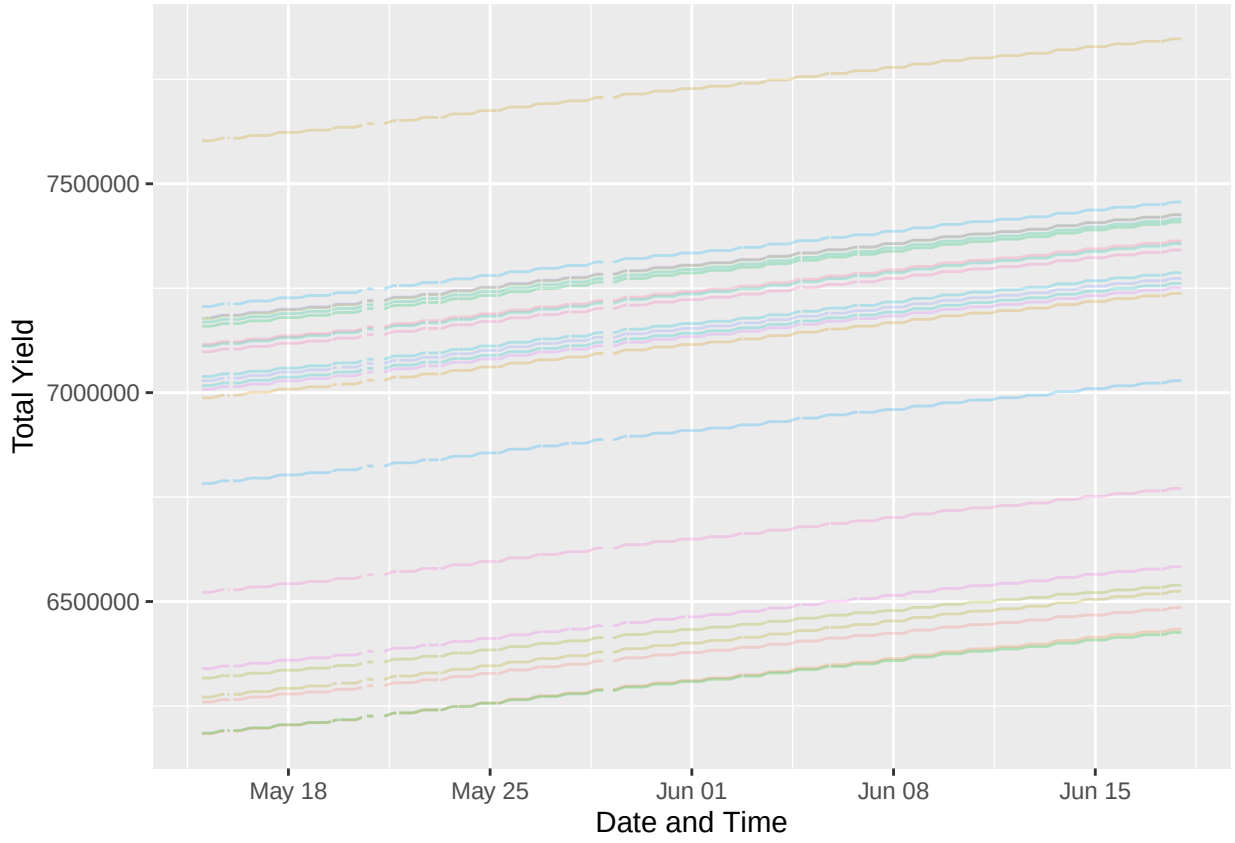

Figure 10: Plant 4135001 Total Yield Over Time

Figure 10 shows a slowly growing total yield from every source. The total yield of each power generation source increases at about the same rate, and missing values appear to propagate across the sensors of the plant.

Figure 11 reveals a few things. First, the different sensors have been running for very different amounts of time. To have total yields that high, many of these sensors have likely been recording data for years longer than at plant 4135001. This may account for the numbers of failure.

While the evidence does not clearly indicate what might cause strange sensor data, the data suggest frequent problems with the sensors. This evidence, combined with the evidence provided by analyzing the DC power output of the panels, justifies removal of plant 4136001 from the data set for model development.

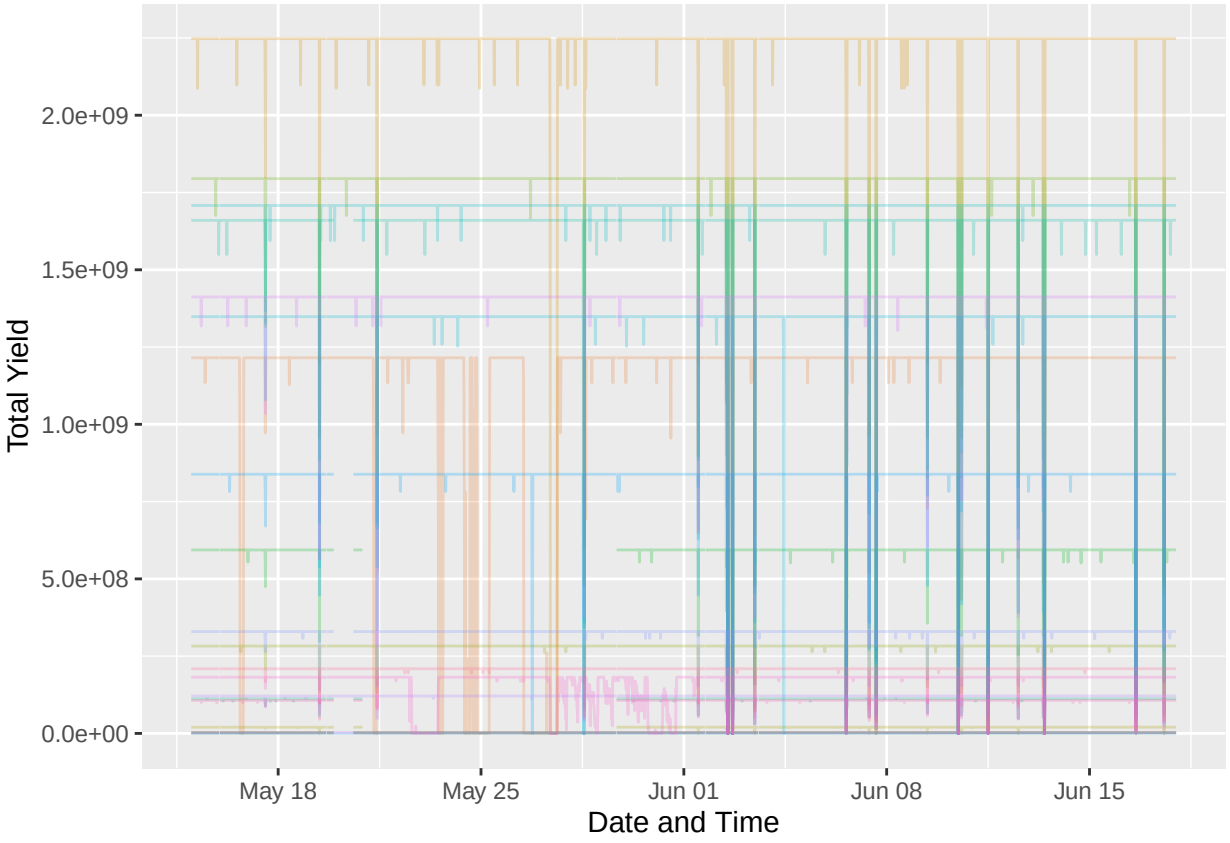

Figure 11: Plant 4136001 Total Yield Over Time

Table 8: Summary of Power Generation Data

Variable	min	max	median	mean	sd
dc_power	0	1441	44.9	320	409
ac_power	0	1405	43.4	313	399
daily_yield	0	9163	2559.9	3307	3186

2.3 Separating Validation Data

To best determine the overall efficacy of predictive models, the data sets were broken into a validation set and a training set. There were 34 days in the entire data set. So, the last three days were partitioned to be a validation set, against which predictions would only be made after the models were finalized. Later, to build the models themselves, the training data was separated further into a train set and test set, using days 30 - 31 for comparison.

2.4 Data Exploration

The data has been cleaned. A validation data set has been partitioned. To get some rough ideas of how a model may be constructed, let us explore the data for patterns.

2.4.1 Generation

While table 8 gives an understanding of what the entirety of plant 4135001 power generation looks like, it is important to look at each sensor individually.

Figure 12 shows consistent medians across many of the sensors. Sensors 1BY6WEcLGh8j5v7 and bvBOhCH3iADSZry trend towards slightly lower values, but only slightly.

Figure 13 shows similar results to examining the DC data, unsurprisingly. Again, sensors 1BY6WEcLGh8j5v7 and bvBOhCH3iADSZry show signs of lowered efficacy.

14 shows an interesting reversal. Sensors 1BY6WEcLGh8j5v7 and bvBOhCH3iADSZry, though showing a lower median DC output, have a slightly higher median daily yield.

Figure 15 appears to be showing values all over the place. However, what figure 15 really reveals is which of the power generation sensors have been running the longest, and as such have generated a larger total yield.

Figure 16 takes into account the cumulative yield of the sensor at at the start of the observations. This normalized view shows general consistency, again with general under performance from 1BY6WEcLGh8j5v7 and bvBOhCH3iADSZry.

Figure 17 overlays the DC power output of each generation source during the first day. Before sunrise and after sunset, about 5:30 AM and 6:30 PM, no DC power is is generated. This is to be expected. The power output grows, peaks, and drops off. This is to be expected, if the angle of the sun is expected to impact the solar panel output. However, the data is much noisier than to be simply described by the angle of the sun.

Most sources follow similar paths, suggesting a plant-wide variable, like irradiation, is a driving force. However, we also see that one source drops to 0 output around 9:15 AM while the others only experience a dip. Sensor faults will have an impact on predictions.

Figure 18 shows the average DC power from each power generation source for the entire month. The result is predictably periodic, peaking at about 1:15 PM each day. Any trends up or down are not evident upon visual examination.

Figure 19 shows the cumulative daily yield of each sensor over the month. This graph more clearly demonstrates the difference in DC output from day to day, and that each solar power source outputs roughly the same amount every day. The smoothness of the peaks might suggest that changes in weather are also generally smooth, if not also periodic.

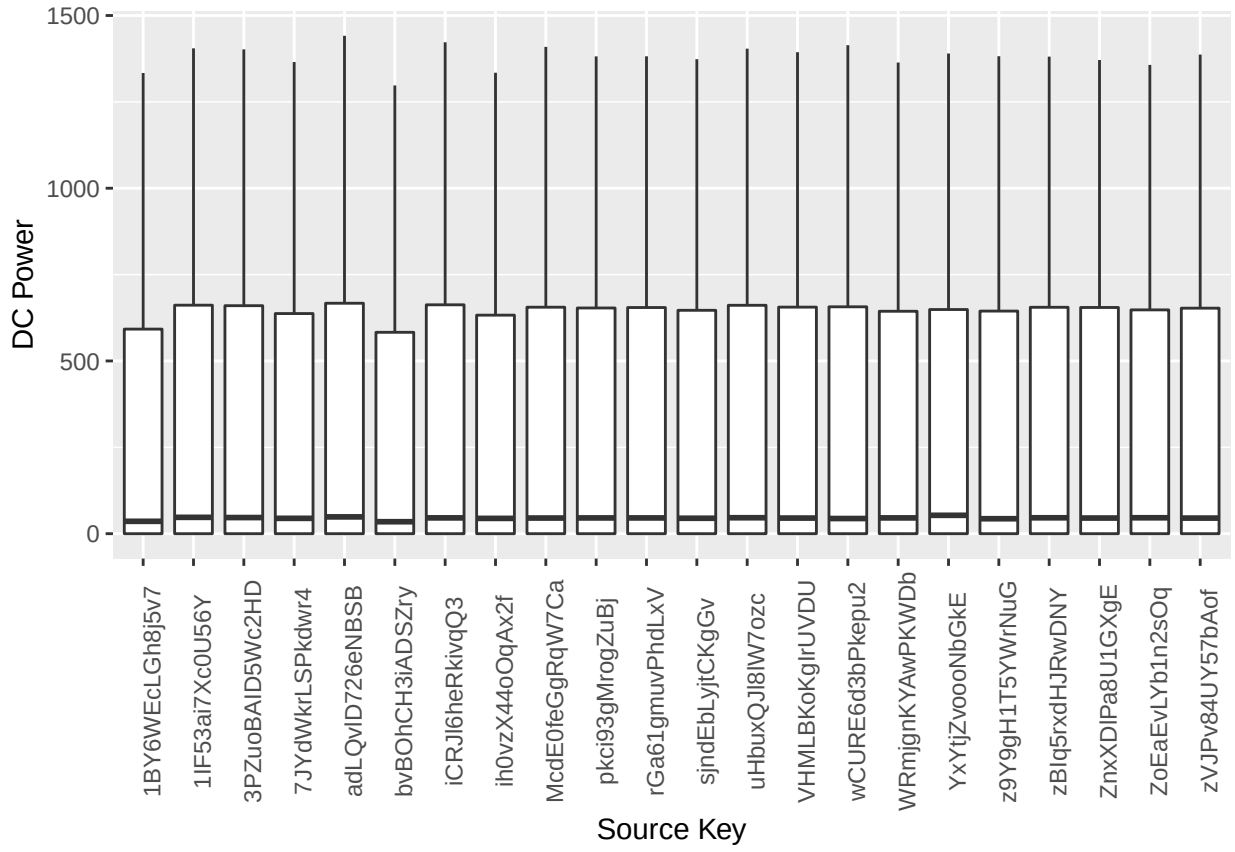

Figure 12: DC Power Output by Source

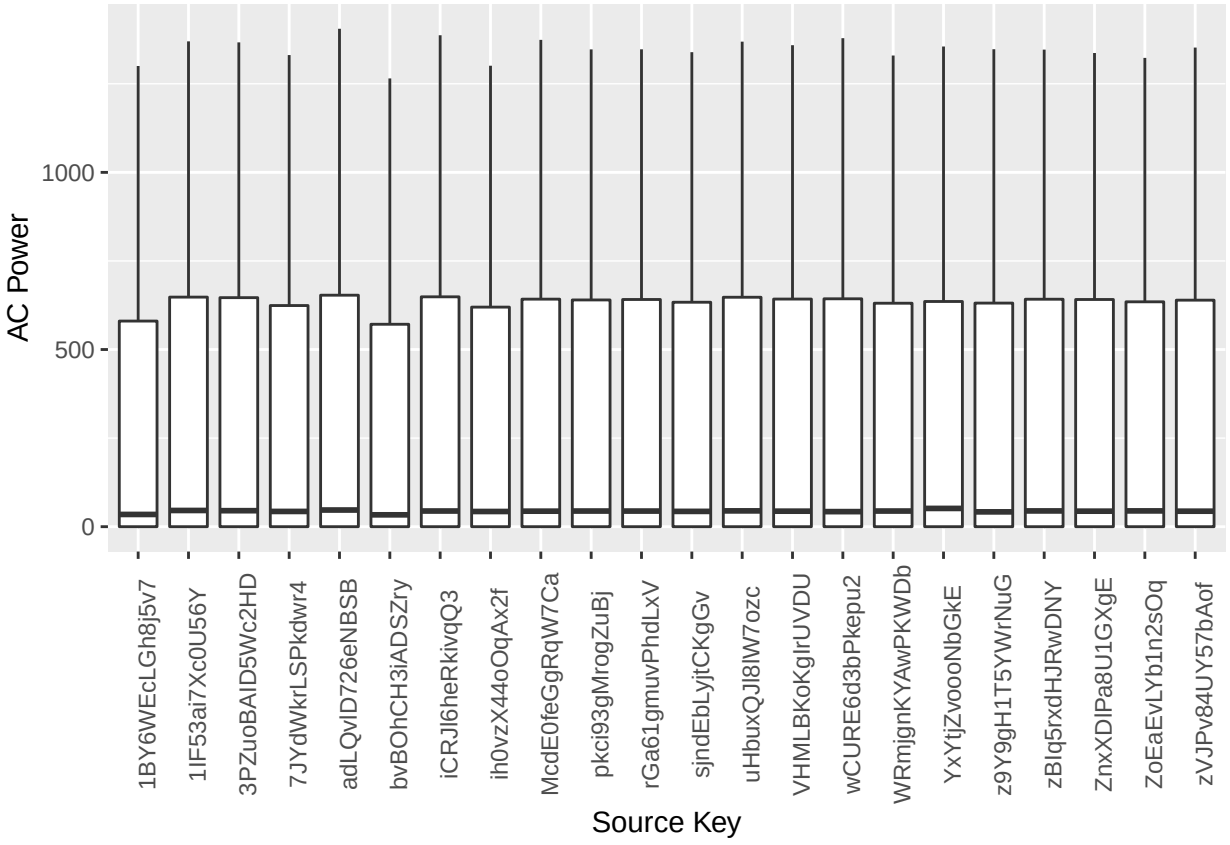

Figure 13: AC Power Output by Source

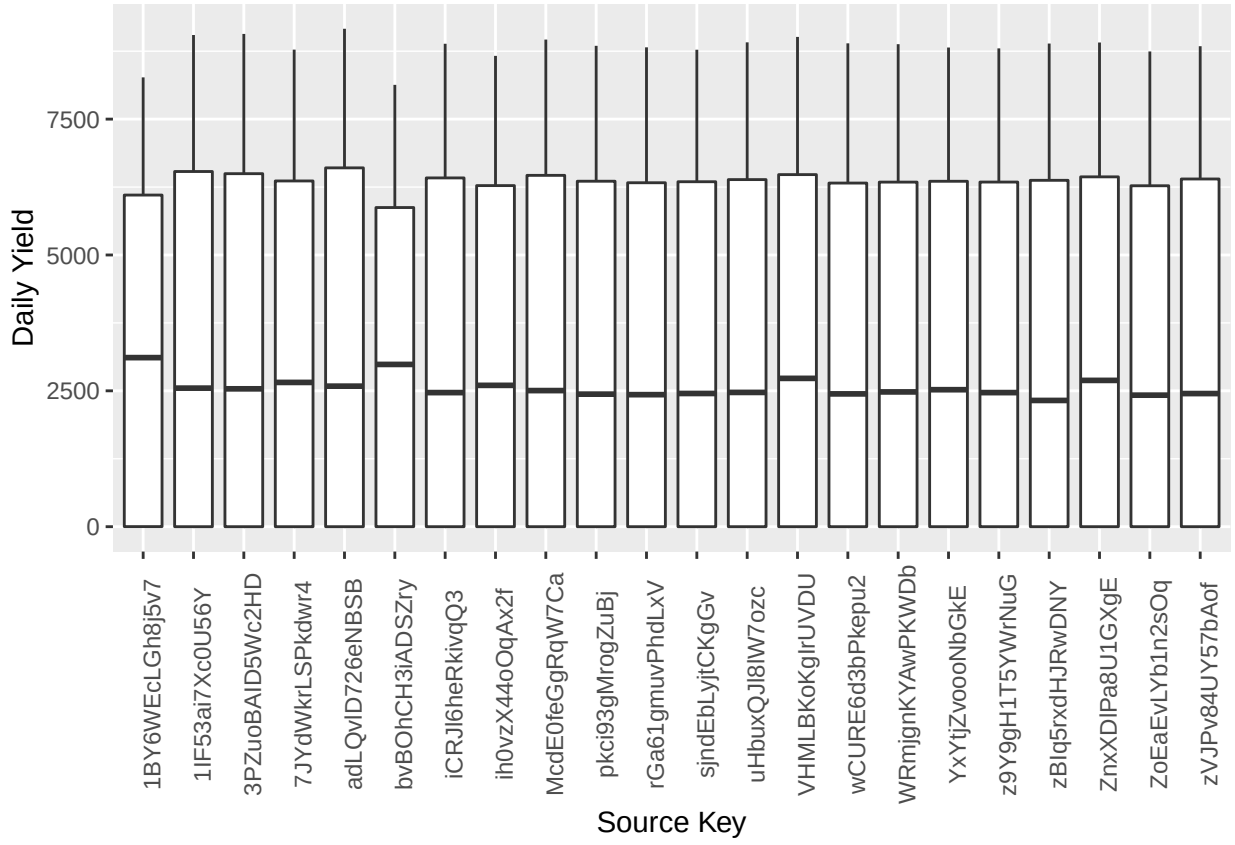

Figure 14: Daily Yield by Source

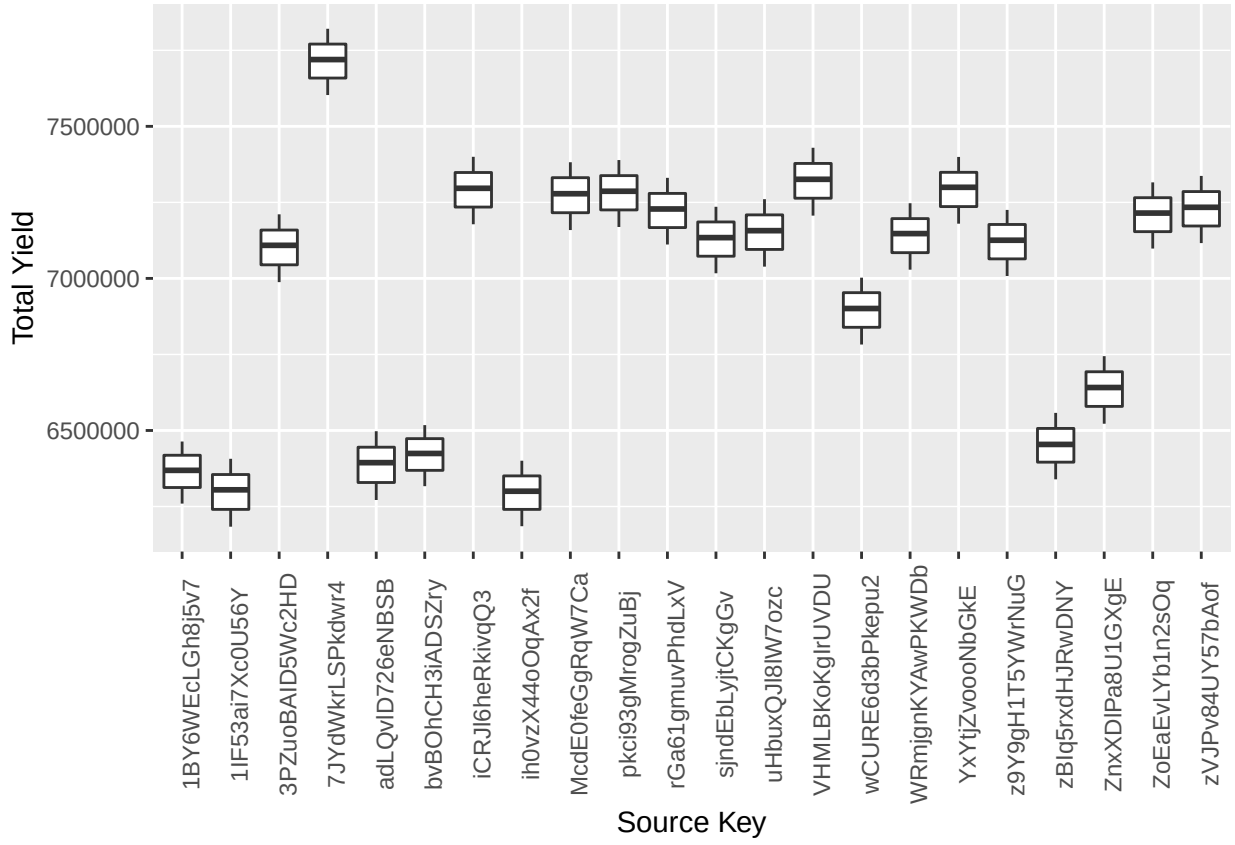

Figure 15: Total Yield By Source

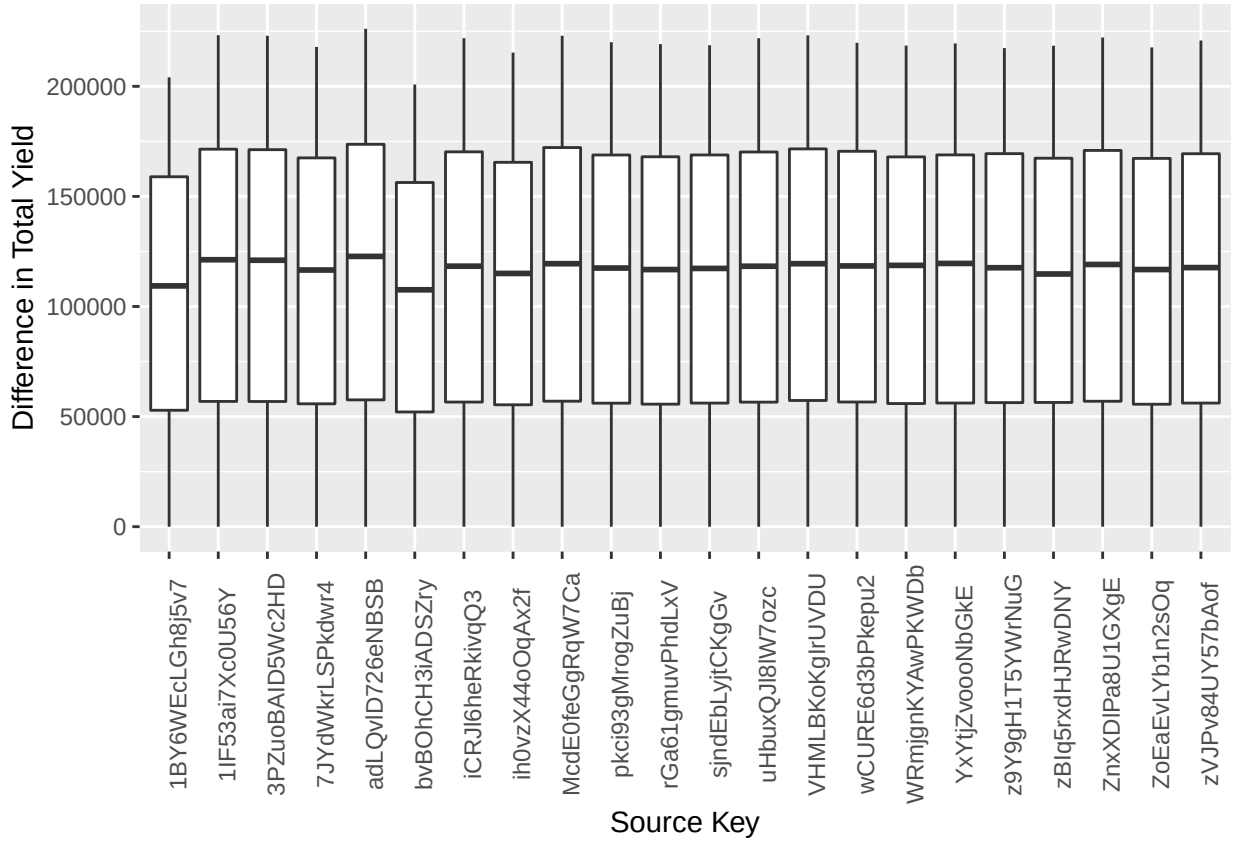

Figure 16: Difference over Month of Total Yield by Source

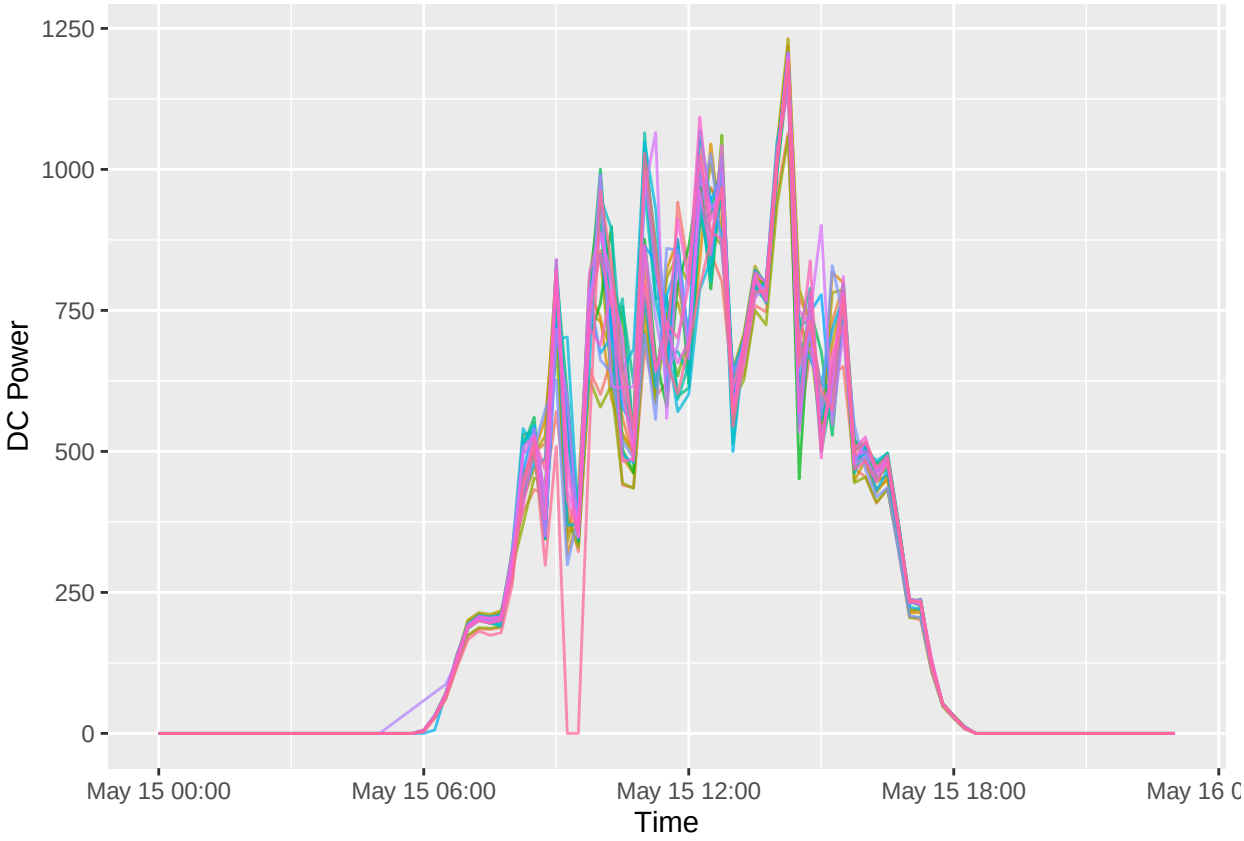

Figure 17: DC Power Output Over One Day

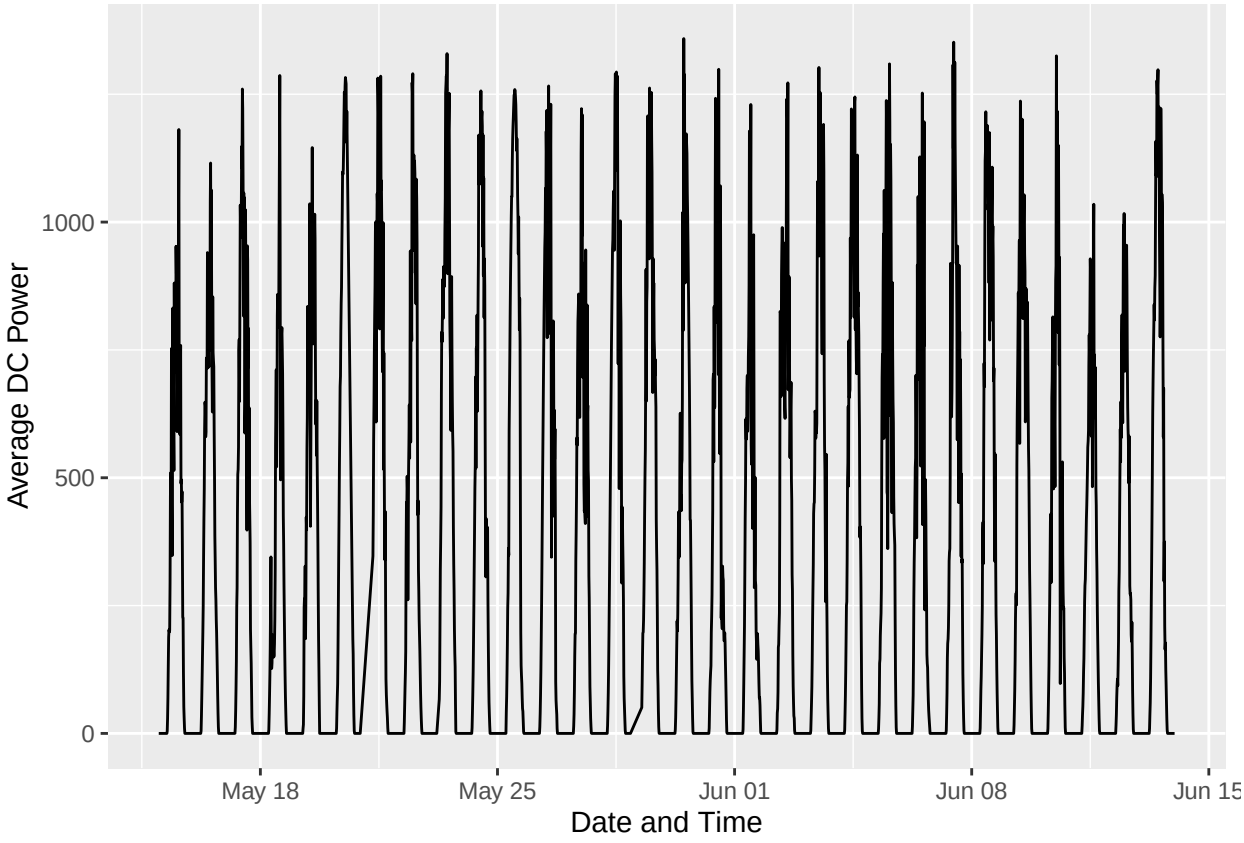

Figure 18: Average DC Power Output Over One Month

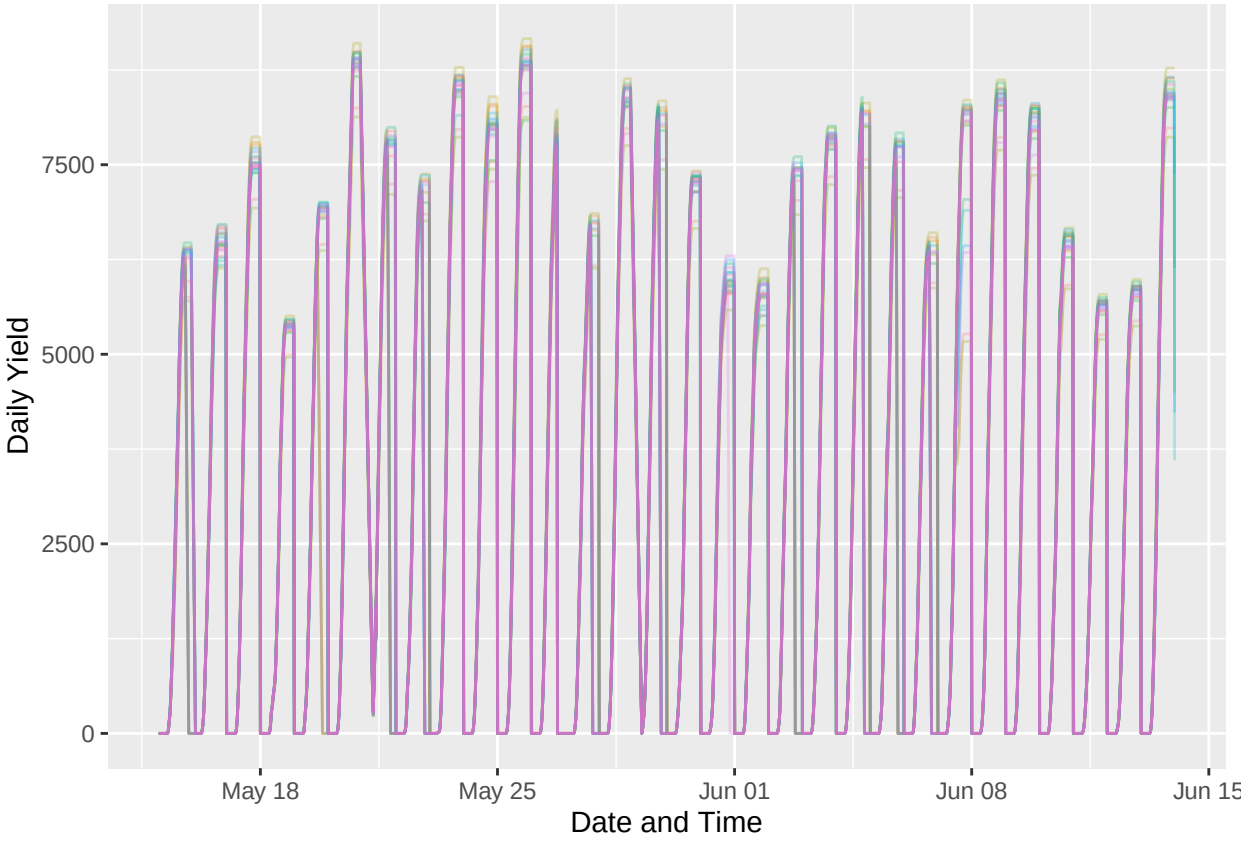

Figure 19: Daily Yield Over One Month

Table 9: Summary of Weather Sensor Data

Variable	min	max	median	mean	sd
irradiation	0.0	1.15	0.026	0.233	0.306
module_temperature	18.1	65.55	24.764	31.334	12.562
ambient_temperature	20.4	35.25	24.812	25.659	3.463

2.4.2 Weather

It is not surprising that the module and ambient temperature have roughly the same median temperature, 24.7 degrees Celsius. However, the module temperature shows a much greater spread in temperature, especially in higher directions. The module also gets colder. This suggests that the module has some sort of cooling system in an attempt to mitigate the higher temperatures that it reaches. The existence of a cooling system suggests that the module has an ideal operating temperature at which it most efficiently generates power. Figure 20 visualizes the greater spread.

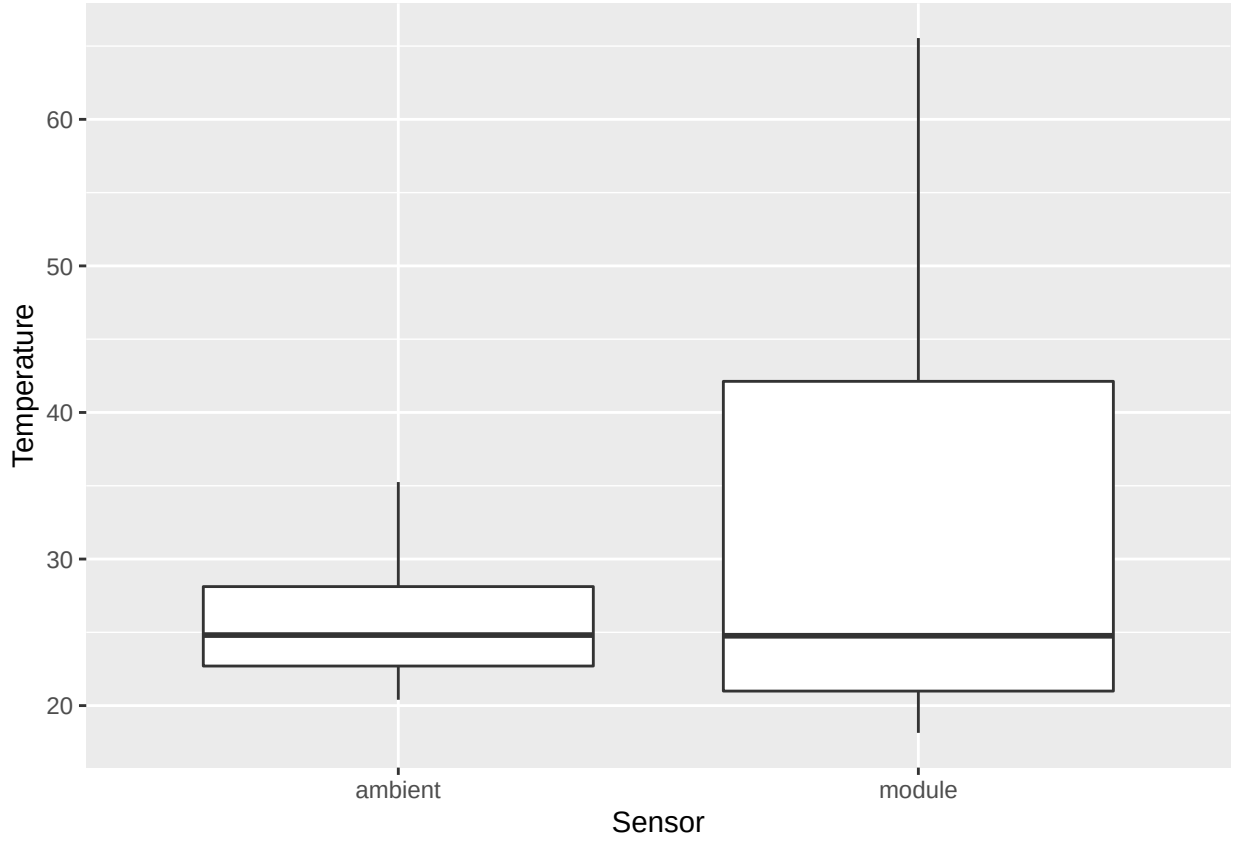

Figure 20: Temperature Readings

Figure 21, a graph visualizing the distribution of the observed irradiation, does not reveal much information. It is not surprising that the most common observed irradiation is 0, as whenever the sun is down, the measured irradiation ought to be 0.

Figure 22, in which nighttime observations have been removed, reveals an interesting quirk. There is a secondary mode for irradiation at about 0.5. This may have to do with how sunlight passes through the atmosphere before reaching the solar panels. At certain angles, enough sunlight is scattered that the irradiation

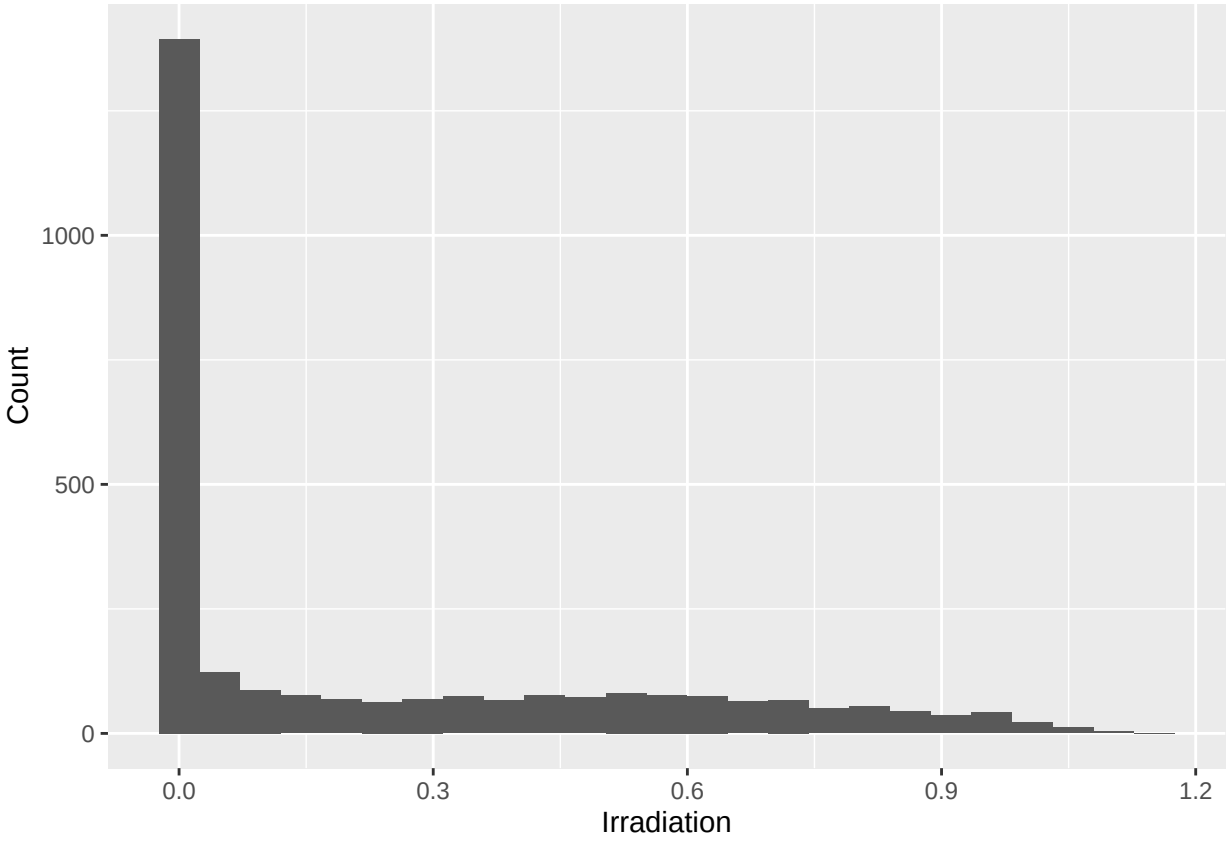

Figure 21: Irradiation Distribution

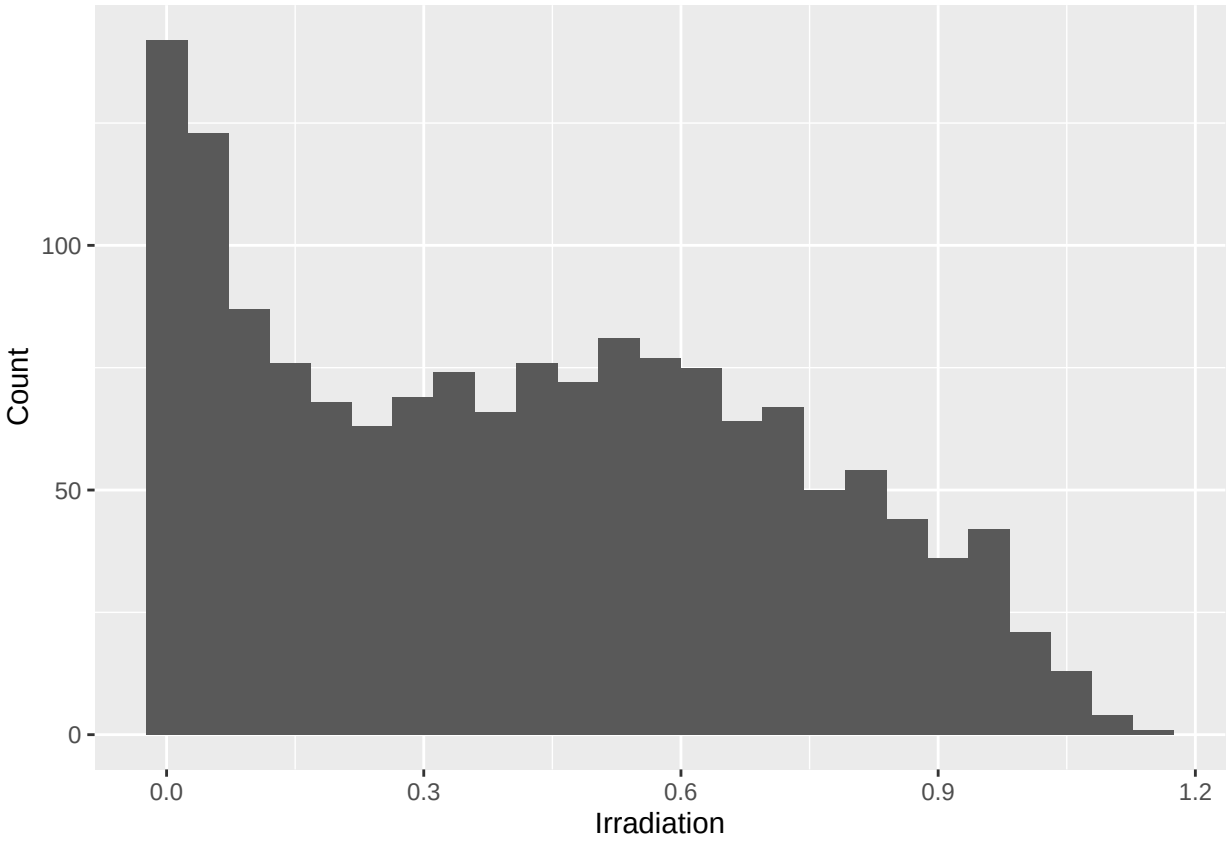

Figure 22: Irradiation Distribution (Day Time Only)

remains low. Once the angle of reflection has been breached, irradiation increases as the angle of the sunlight approaches normal, or 90 degrees. Regardless, this graph reveals there is more to the irradiation reaching the solar panels than simple geometry.

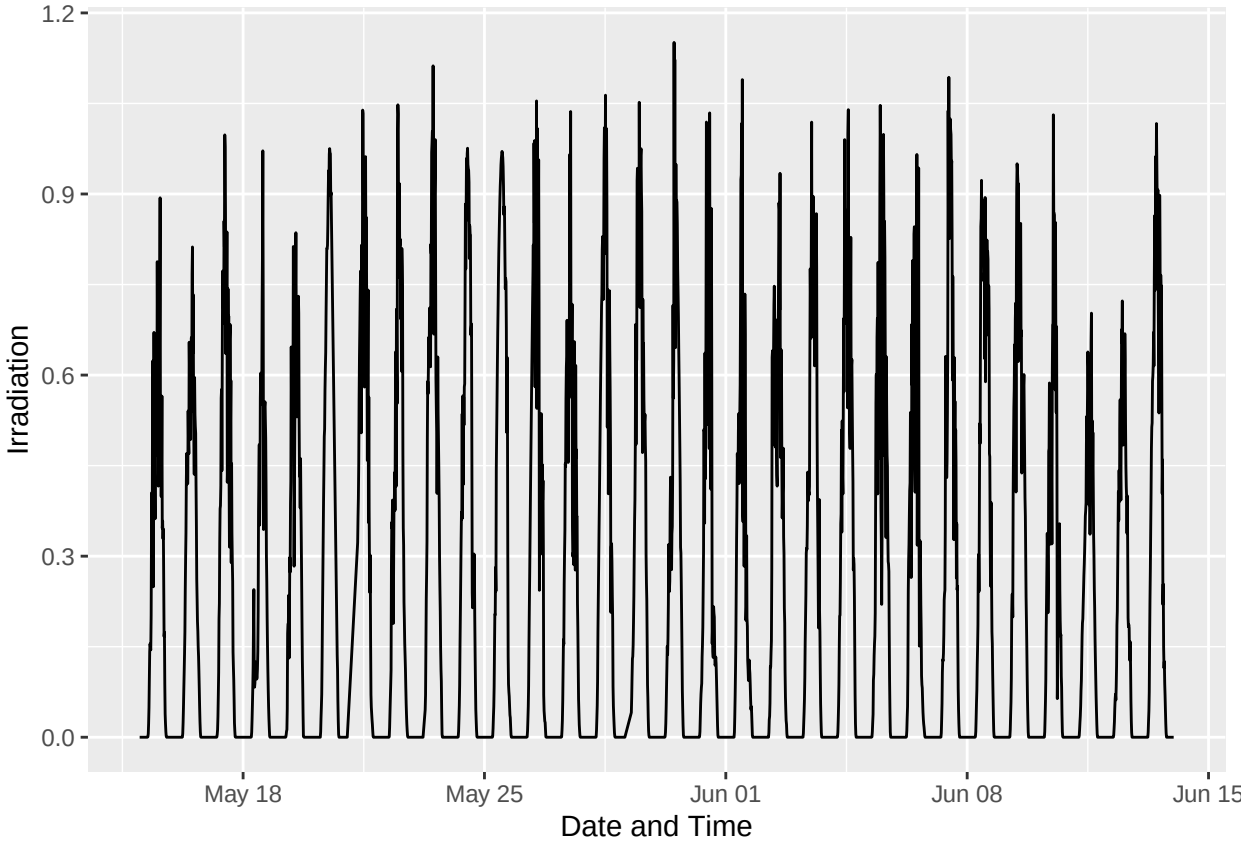

Figure 23: Irradiation Over Time

Figure 23 displays that irradiation over time follows a pattern that looks very similar to the graph generated of DC power output over time. It is periodic, with a frequency of 1 day, peaking at approximately 1 PM.

Figure 24 shows that the temperature data also follows what might be expected after seeing the summary table. The module temperature varies much more than the ambient temperature. However, note that the ambient temperature actually lags behind the module temperature. The relationship between irradiation and temperatures will be further examined below.

Note that in figure 25, the scale of the irradiation has been multiplied by 40 and shifted up 20 degrees to better visualize the relationship with module temperature and its lag. Irradiation leads module temperature by about 15 minutes and ambient temperature by a number of hours. However, the module temperature appears to be some composition of the ambient temperature and irradiation. This might suggest that the weather data can largely be described by the irradiation, or by the irradiation and the ambient temperature.

2.5 Correlation

The correlation between the DC Power and AC Power was found to be 1. They are highly correlated. This aligns with expectations, that a certain percent of power is lost in during conversion. The loss can be determined by fitting the data to a line.

The fit shown in table @rec(tab:ac-dc-line-fit) suggests a 97.7% conversion rate from DC to AC power, only

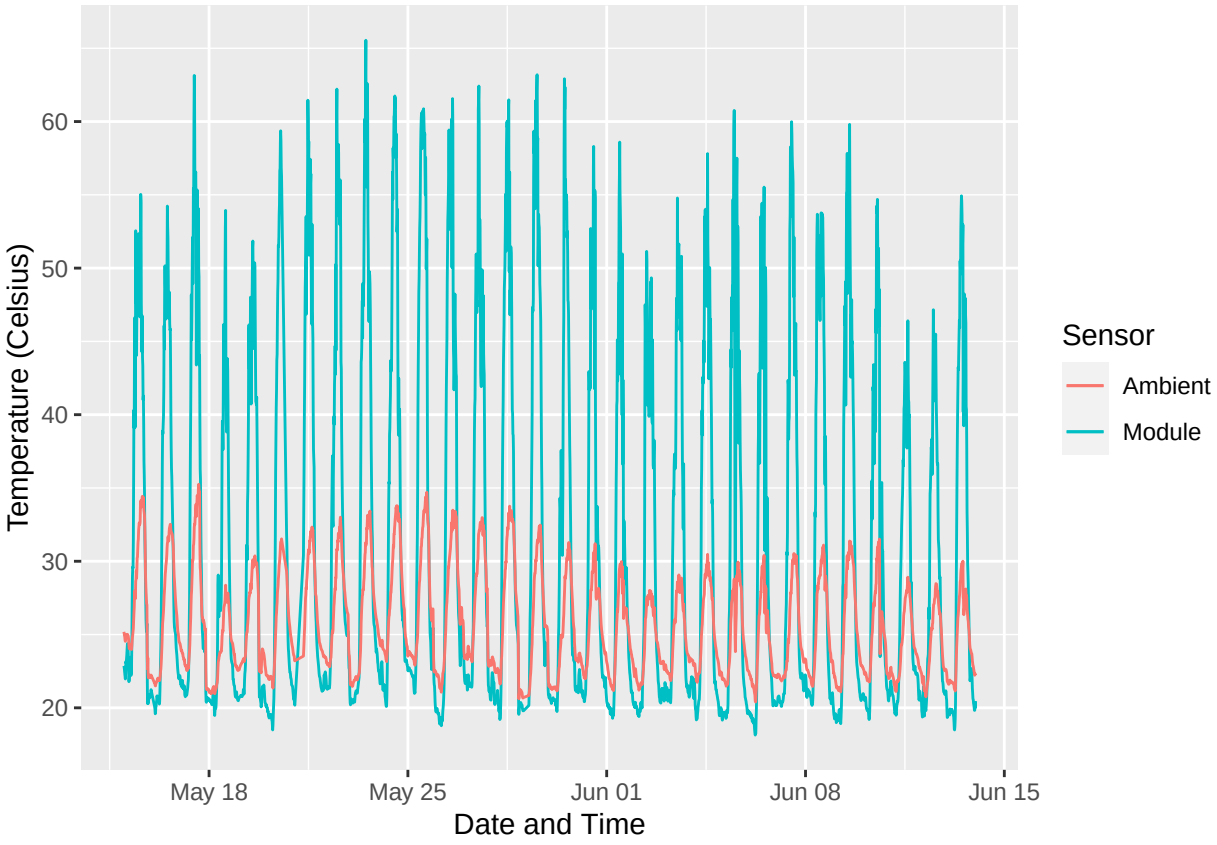

Figure 24: Temperature Over Time

Table 10: AC Power per DC Power Line Fit

term	estimate	std.error	statistic	p.value
(Intercept)	0.277	0.006	50	0
dc_power	0.977	0.000	91637	0

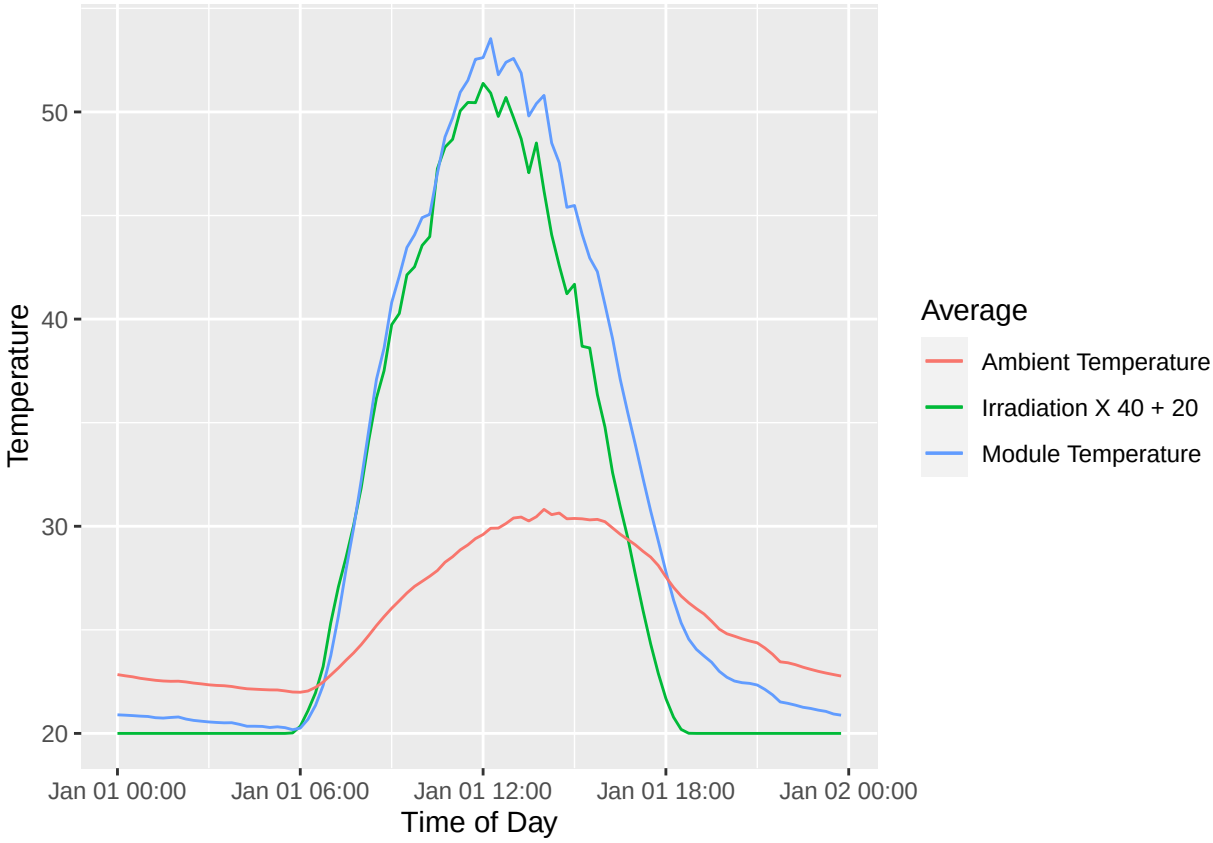

Figure 25: Averaged Weather Data

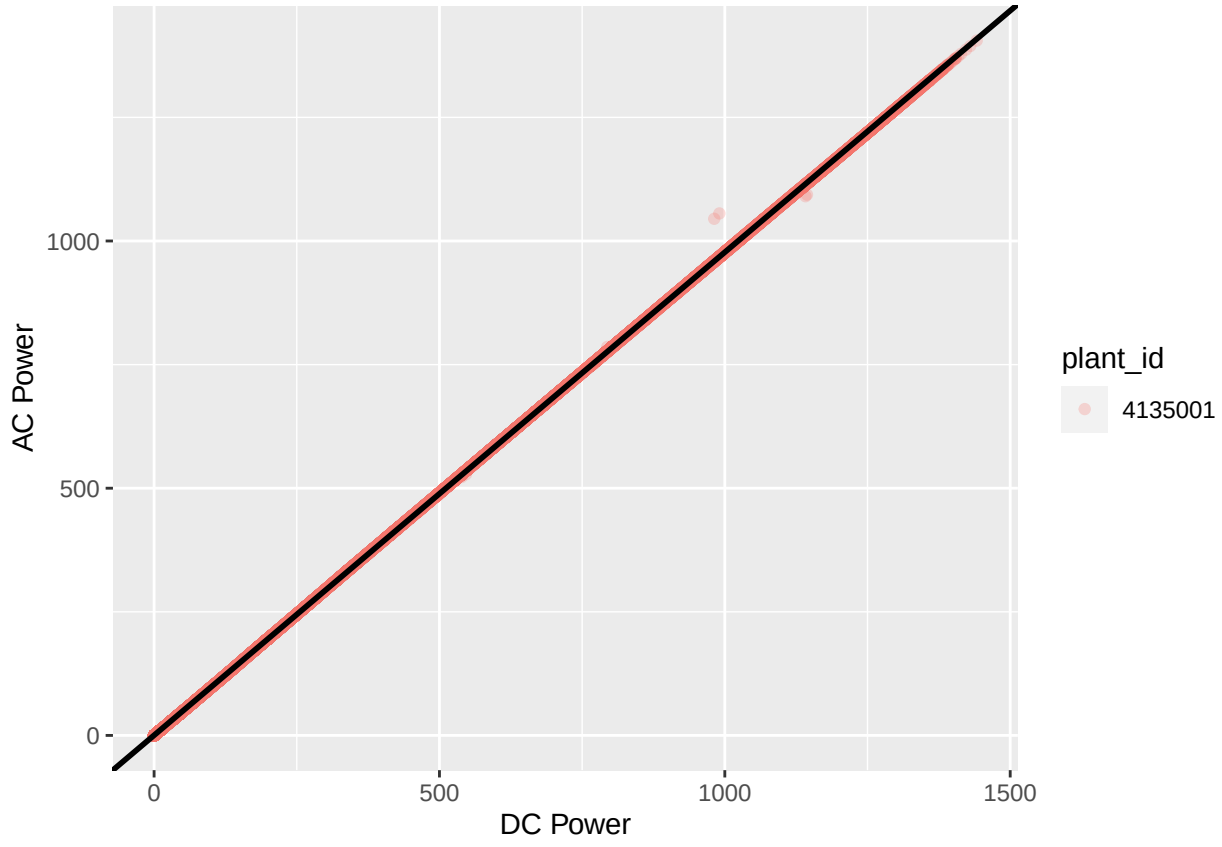

Figure 26: AC Power Produced per DC Power In

Table 11: Lag of Maximum Cross Correlation Between Irradiation and Module Temperature

lag	correlation
15m	0.975

2.3% is lost. Such a small loss suggests that the plants are using high quality transformers. We can see in figure 26 that the fit is good.

2.5.1 Lag

While exploring the weather data, a lag was noted between irradiation, module temperature, and ambient temperature. This could represent some relationship between the three. Because of the physical properties of temperature and heat flow, it is likely that irradiation drives both the module temperature and the ambient temperature. It is unlikely that the module temperature has an effect on the ambient temperature. It is possible that the ambient temperature acts as a heat or cold sink for the module.

For a better understanding of what may be going on with weather at the power plant, the values of the weather sensors are treated as time series.

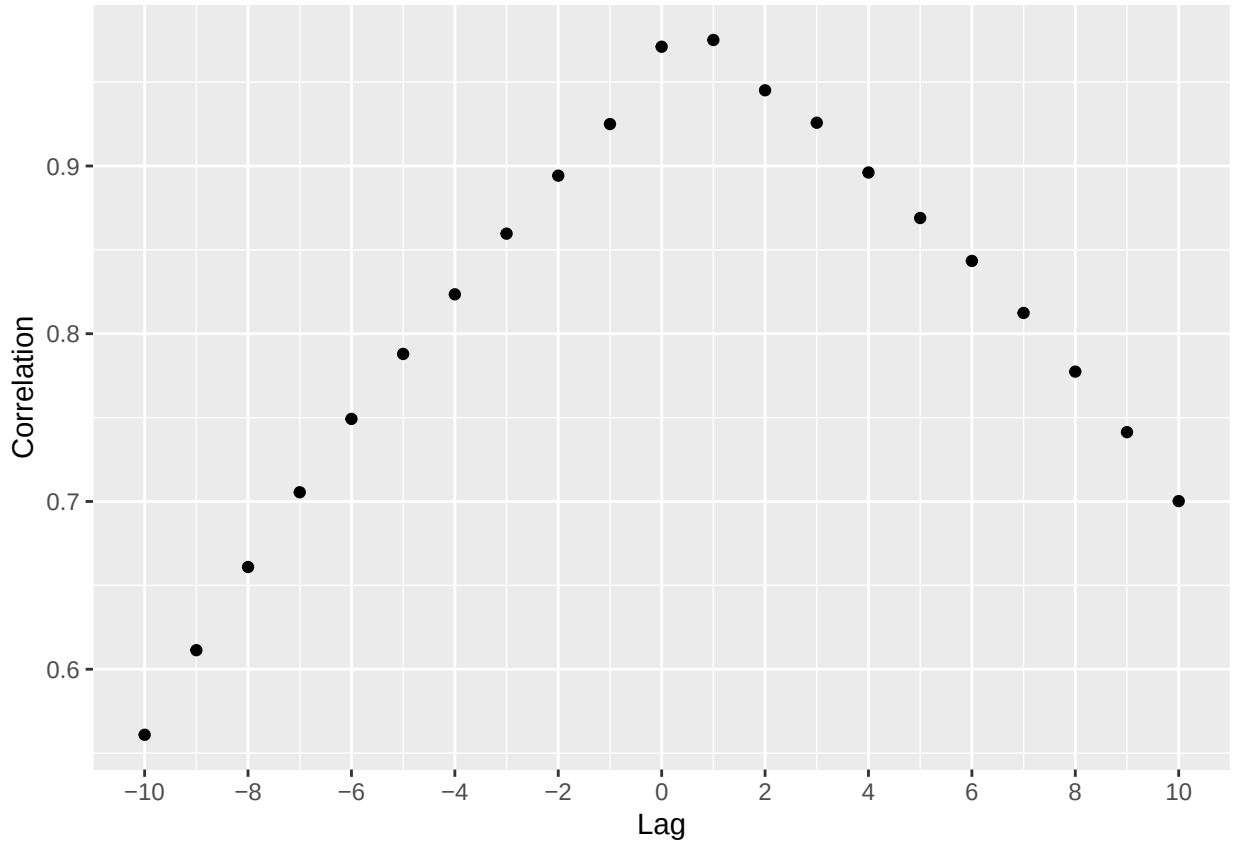

Figure 27: Cross Correlation Between Irradiation and Module Temperature

A strong correlation is found between the irradiation and the module temperature after a 15 minute lag, as shown in figure 27 and table 11. This can be explained by the radiation heat transfer as a result of the electromagnetic radiation of the sun. It does not take long for higher irradiation to increase the temperature of the module. However, referring back to the average day temperatures graph, it appears that as the radiation

Table 12: Lag of Maximum Cross Correlation Between Irradiation and Ambient Temperature

lag	correlation
75m	0.885

increases, temperature increases without lag As the radiation decreases, the temperature decreases after about 30 minutes.

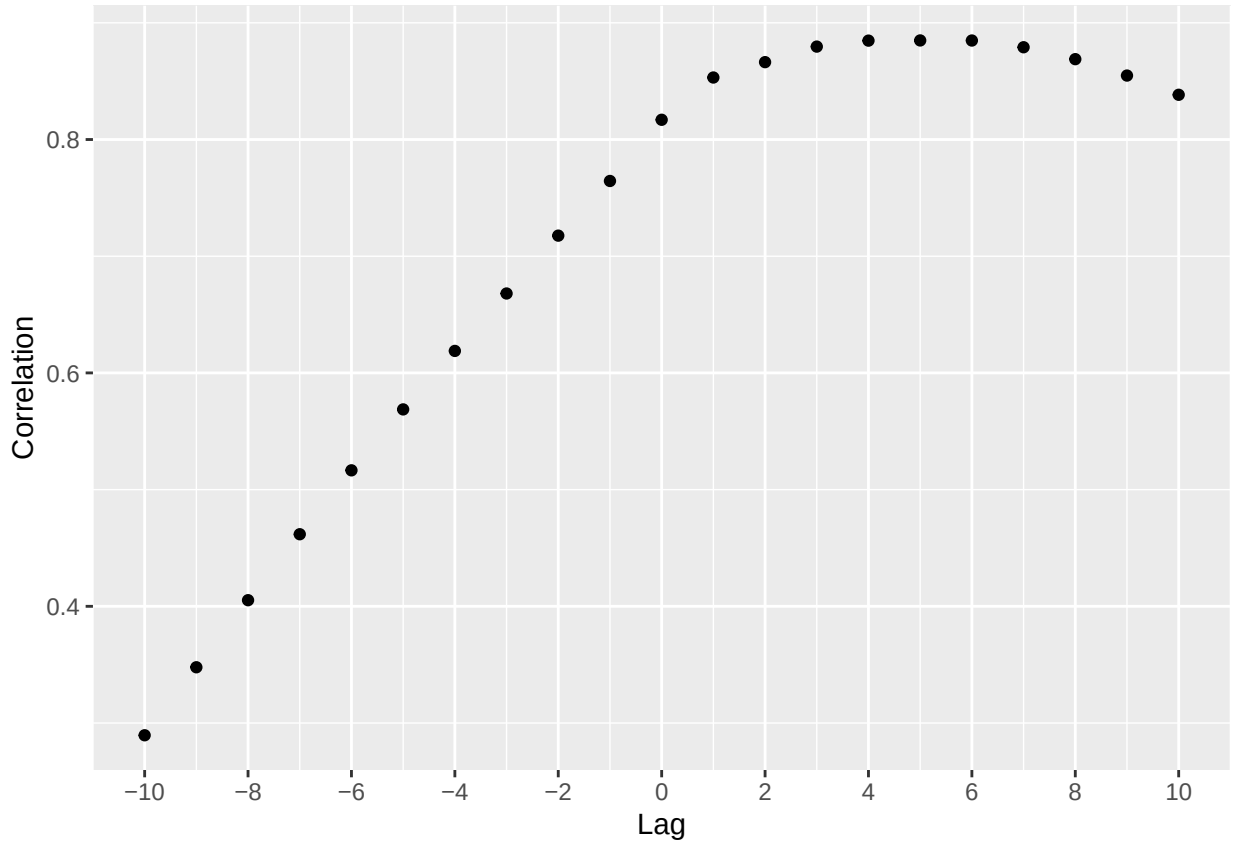

Figure 28: Cross Correlation Between Irradiation and Ambient Temperature

The connection between irradiation and ambient temperature is less clear. The correlation of 0.877 after an hour and fifteen minutes suggested that irradiation has some effect on the ambient temperature. The increased lag time has to do with how much less interaction there is between rays from the sun and air molecules vs a stationary module. However, there are sufficient confounding and unobserved variables related to weather and weather prediction that it is beyond the scope of this analysis to dig deeper.

Interestingly, the ambient temperature lags the module temperature, and by 45 minutes, as shown in table 13. At first glance, this does not make much sense. It is unlikely that a module could be driving the ambient temperature.

However, if one keeps in mind the confounding variable of weather, the relationship between irradiation and module temperature explains this. It is far more likely that the ambient environment acts as a heat sink when the module becomes very hot. Though the physical interactions of heat transfer might be modeled, they are outside the scope of this analysis.

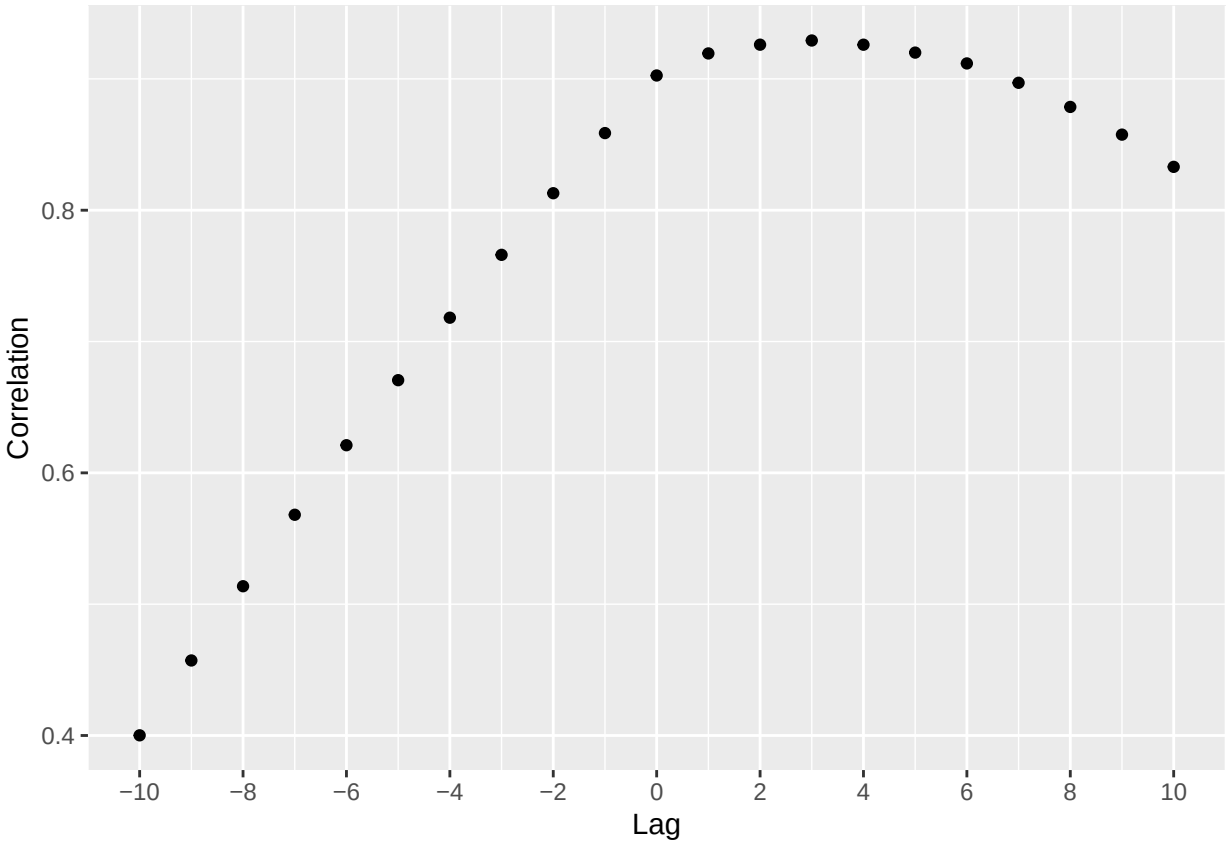

Figure 29: Cross Correlation Between Module Temperature and Ambient Temperature

Table 13: Lag of Maximum Cross Correlation Between Module Temperature and Ambient Temperature

lag	correlation
45m	0.929

mu_hat	naive_rmse
321	402

Table 14: Model Results

method	RMSE
Average by Source	401

2.6 Building The Linear Regression Model

2.6.1 Splitting the Train and Test Sets

A validation data set has already been separated from the data. However, to evaluate the efficacy of the model as it is constructed, it becomes necessary to further partition the data to avoid over training. The final two days of the solar data are put into a test set, all other data remain in a train set.

2.6.2 Baseline (Naive RMSE)

The RMSE is the metric by which success is measured. To get an understanding of how much the model is improving, the naive RMSE will be calculated as a baseline. That is, the predicted dc_power will simply be the mean of all dc_power observed.

The average of the dc_power over the entire data set is 321. Predicting the average dc_power results in an RMSE of 402. That is substantial error. Time to explore ways to reducing reduce it.

2.6.3 Generation Source Effect

To start, the effect of stratifying by generation source is examined. The results are shown in table 14.

A slight improvement, but only slight. About a quarter of a percent.

2.6.4 Irradiation Effect

In exploration, it was observed that irradiation and DC power output were highly correlated. A linear regression method was used to predict DC power output based purely on irradiation, using the train set to predict against the test set. The result is shown in table 15.

An RMSE of 44.4 is, unsurprisingly, a substantial improvement over the naive RMSE.

Figure 30 shows the line that resulted from linear regression, and suggests that a linear relationship fits rather well. How much can the RMSE be improved upon using a linear model, beyond using solely irradiation as a predictor?

2.6.5 Module Temperature Effect

It is not unreasonable to consider other weather effects to construct the model. We can get a better understanding of how temperature effects the DC power output by sweeping away the irradiation effect and plotting the resulting residual against the module temperature. The results are shown in figure 31.

Figure 31, relating module temperature and the DC power output, suggests that the module has an optimal operating temperature between 20 and 54 degrees. The relationship does not appear to be a straight line,

Table 15: Irradiation Effect

method	RMSE
Irradiation Effect	44.4

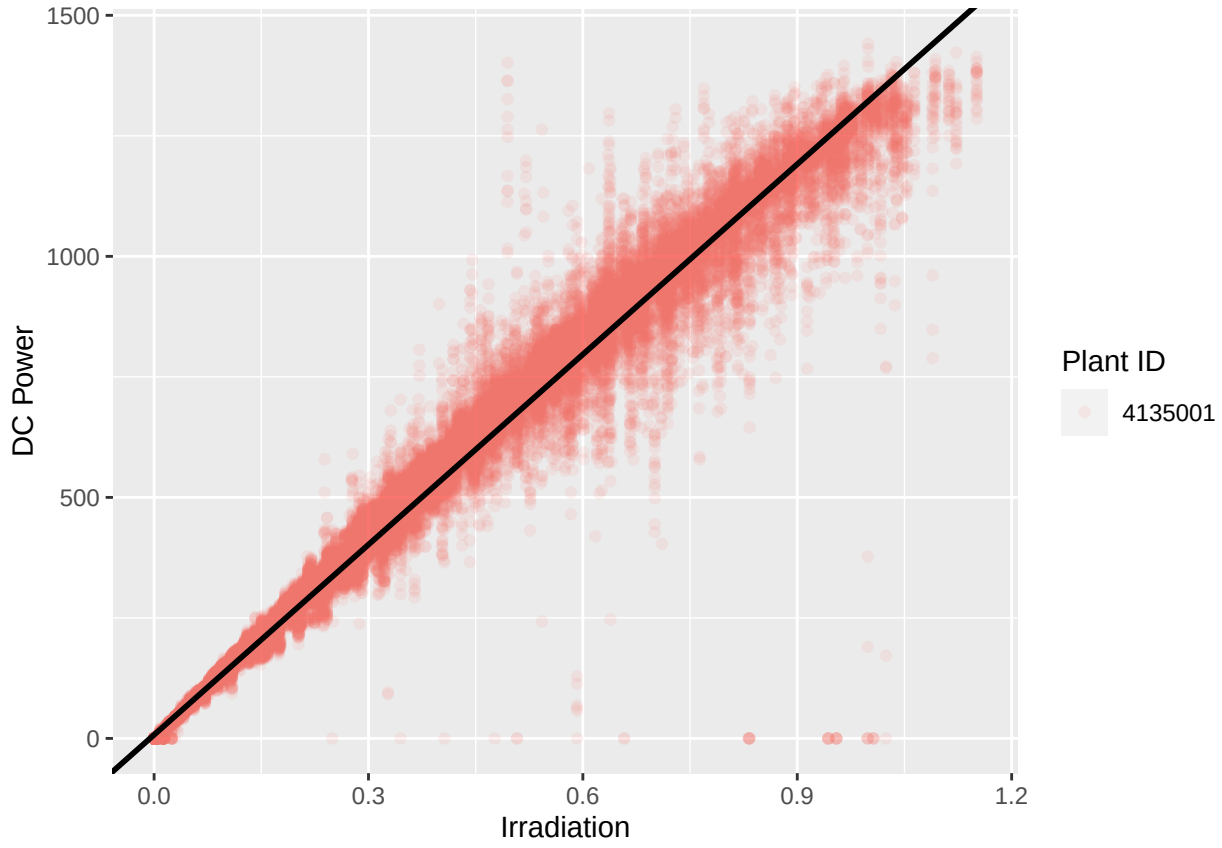

Figure 30: Irradiation Fit

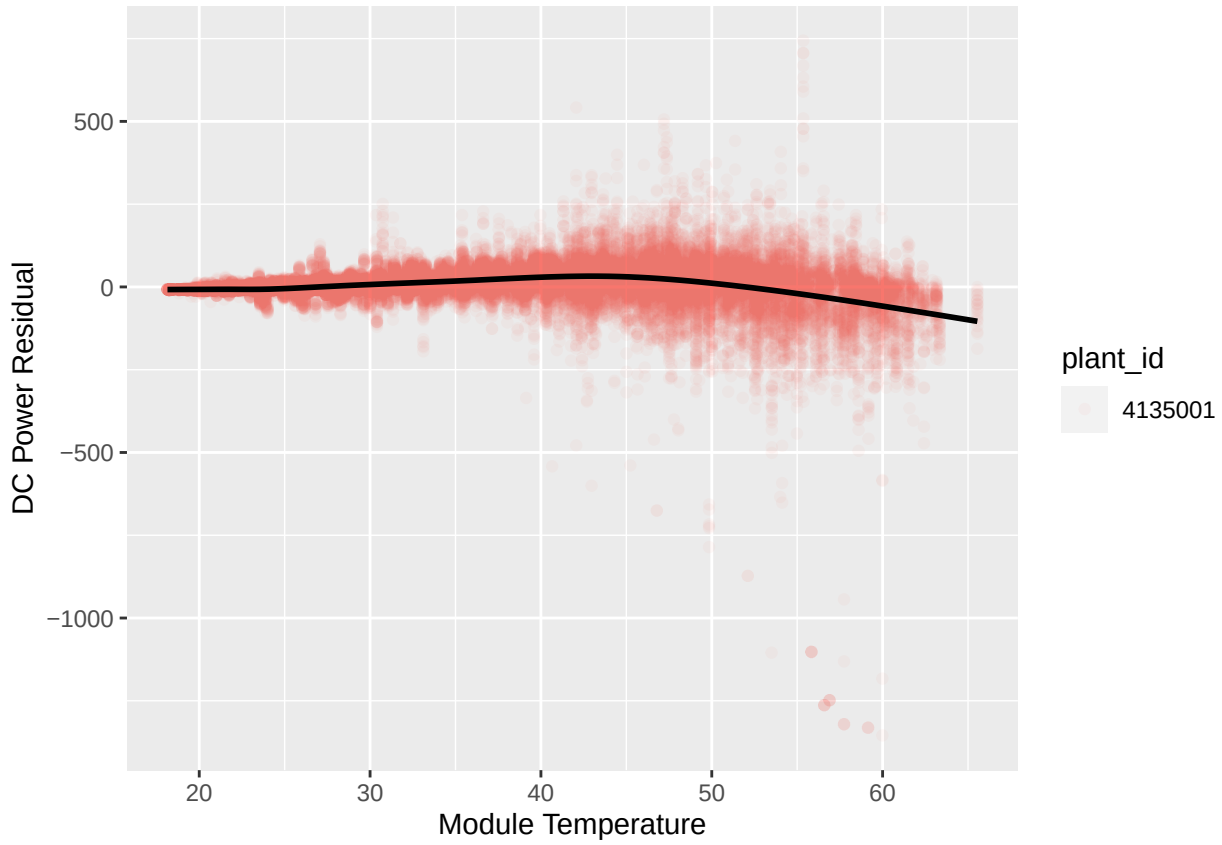

Figure 31: Module Temperature Effect

Table 16: Module Temperature Effect

method	RMSE
Irradiation + Module Temperature Effect	40

method	RMSE
Irradiation + Irradiation Lag Effect	40.6

however. Perhaps the module temperature effect is best represented by parabola.

To account for the parabolic shape of the data, a placeholder variable was added to the train and test sets that was simply the squared module temperature. Then, a fit was made based on irradiation, module temperature, and the module temperature squared. The results may be seen in table 16.

40.0, a %9.9 improvement over the prediction made purely on the irradiation effect.

Earlier, it was discovered that module temperature lagged irradiation by 15 minutes and was highly correlated. Can a similar RMSE be achieved by simply lagging the irradiation by 15 minutes?

The `lag` function from dplyr allows the creation of a placeholder variable that hold the irradiation value from 15 minutes prior (the previous data entry). A second placeholder variable contains the squared lag value. Note, the lagged time series results in an NA. Since this is at night, the NA is be replaced with the lowest predicted dc_power.

40.61, about %1 worse than if module temperature was incorporated instead.

2.6.6 Ambient Temperature Effect

Now, the residual DC power is compared to the ambient temperature, similar to how module temperature was examined. The result of sweeping out the irradiation and module temperature squared effect are shown in figure 32.

Figure 32 suggests there is a negligible effect.

Table 17 reinforces what was assumed from figure 32. Adding ambient temperature to the models results in an improvement of less than half of a percent.

2.6.7 Clamping and Sensor Faults

Let's take a glance look at how our predictions compare to the test data. A summary is shown in 18.

Immediately, a source of improvement presents itself. It is not physically possible for negative DC power to be produced. Negative DC Power would suggest that the solar panels are drawing energy in. To account for this impossibility, a clamped prediction increases any results DC Power that is less than 0 to 0. The results of clamping may be seen in table 19

Other intuitive improvements may be made. For example, consider the anomalous data found during exploratory analysis. Do we improve the predictive model if we do not train anomalous data?

Table @ref(tab: clamped-no-anomaly) shows the result of removing anomalies from the training data and clamping the prediction.

Table 17: Ambient Temperature Effect

method	RMSE
Irradiation + Module + Ambient Temperature Effect	39.9

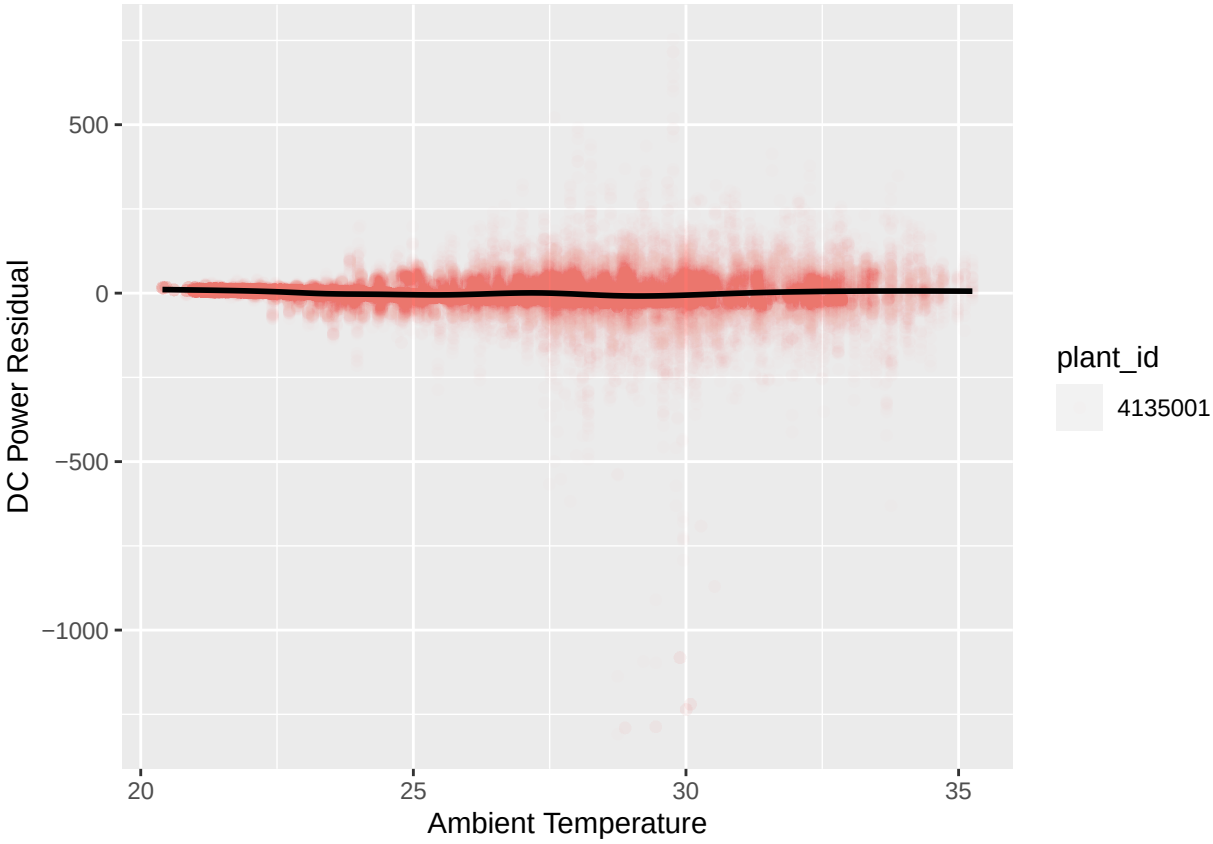

Figure 32: Ambient Temperature Effect

Table 18: Summary of Linear Fit Predictions

DC_Power	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Predicted	-15.1	-3.81	34.5	296	619	1316
Test Set	0.0	0.00	35.3	305	632	1374

Table 19: Clamping Effect

method	RMSE
Clamped	39.7

method	RMSE
Clamped + Anomolies Removed	39.2

39.15 is a considerable improvement of 90.3% over the naive RMSE. The final clamped RMSE of 39.15 is an 11.8 % improvement on the uni-variate linear regression prediction based on irradiation. Figure 33 shows the final fit. The colored lines represent the observed DC output at teach plant. The black line represents the final linear fit.

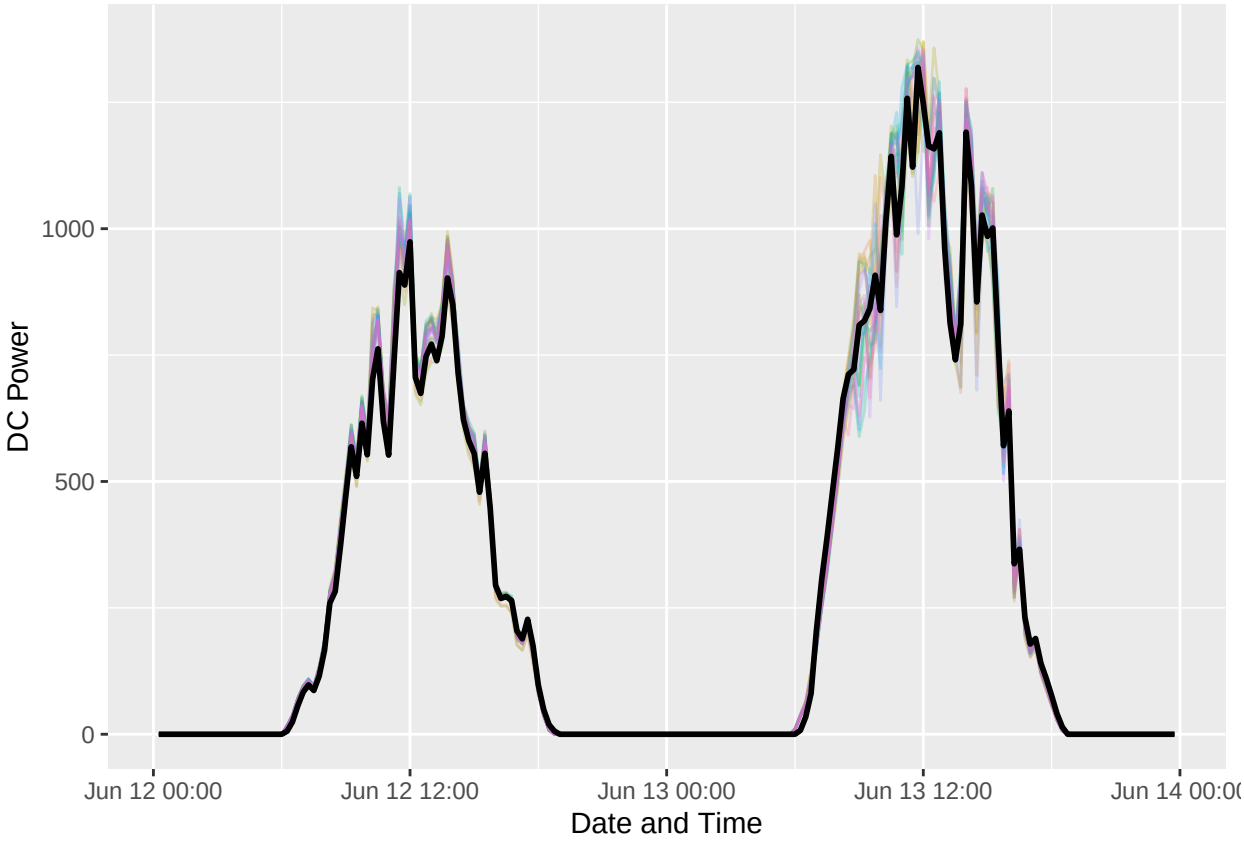

Figure 33: Final Linear Prediction on Test Set

2.7 Building the ARIMA Model

ARIMA stands for **A**uto-**R**egressive **I**ntegrated **M**oving **A**verage. It is a statistical analysis model that uses time series data to predict trends. Auto-Regressive: Predicts future values based on past values, regresses on its own lagged values Integrated: Differences raw observations to make the time series stationary Moving Average: Smooths out time series data by creating a subset of recent data to use as an average auto.arima (from forecast) and ARIMA (from fable) automatically determine the p, q, and d parameters, that is, the parameters used in the auto-regression, integration, and the moving average.

2.7.1 Averaged ARIMA

Below is the result of running an auro.arima on the train set, after averaging the observed DC power from each power generation source. Figure 34 shows the predicted result for the test set days.

```
## Series: .
## ARIMA(2,0,0)(0,1,0)[96]
##
## Coefficients:
```

Table 20: ARIMA RMSE

method	RMSE
ARIMA	179

```
##         ar1    ar2
##       0.57  0.081
## s.e.  0.02  0.020
##
## sigma^2 = 19702:  log likelihood = -15874
## AIC=31754   AICc=31754   BIC=31771
```

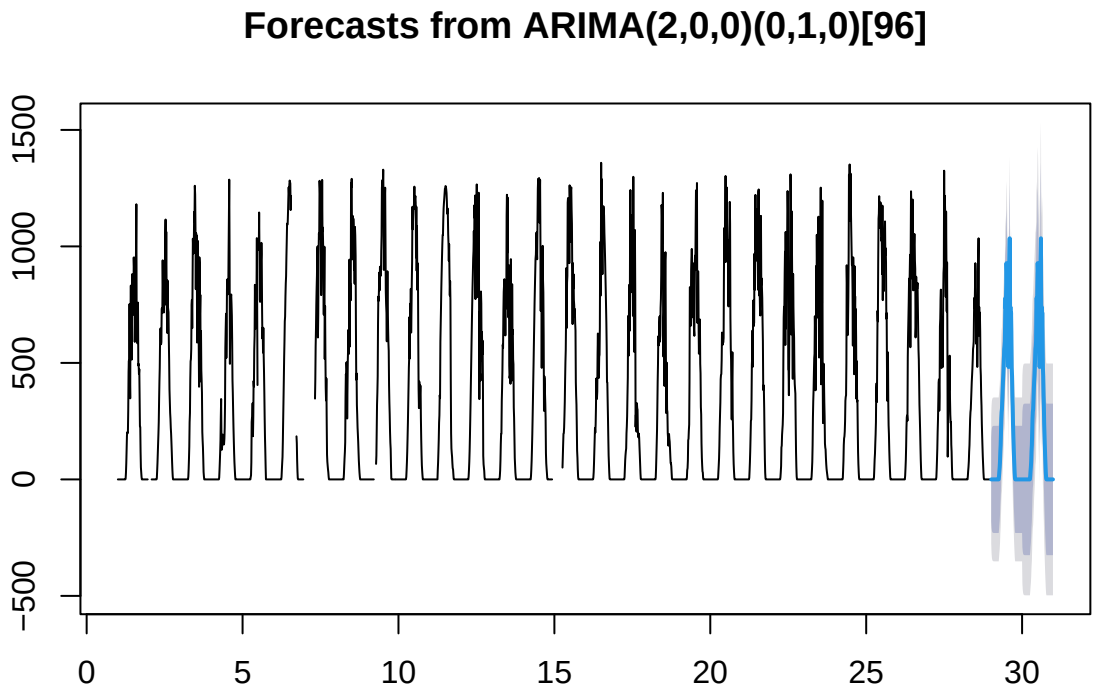

Figure 34: ARIMA Forcecast on Test Set

Table 20 shows the RMSE of the ARIMA model. An RMSE of 179 is much worse than the irradiation prediction, but it is important to keep in mind that this prediction was made using *only* DC power as a predictor. The fact that it improves so much over the average is impressive.

2.7.2 ARIMA By Generation Source

It is possible through the use of tsibbles and the fable package to be more granular with ARIMA. In particular, it may be separated to apply only to each group. The previous ARIMA model used only the average dc. Table 21 shows the result of applying an ARIMA fit to each group.

Stratifying by generation source significantly reduced the RMSE of applying the ARIMA. This is an

Table 21: Result of Applying ARIMA to Each Generation Source

method	RMSE
ARIMA by Source	40.7

Table 22: Random Forest RMSE

method	RMSE
Random Forest	39.4

improvement on the uni-variate linear regression, but does not hold up to the multivariate models. Note that for this result, clamping predictions to be greater than 0 was applied. Clamping only reduced the RMSE by 0.3. The individualized-by-source fits are shown in figure 35.

While there exist methods to include external regressors to an ARIMA fit, those models will be left for future work.

2.8 Building the Random Forest Model

First, a linear model was constructed out of the variables that the author could think to reasonably include. Then, an ARIMA model was applied to the DC power to get single dimensional results. Now, the floor will be opened to all observed variables by using random forest to make predictions.

Random forest will require the data to be presented in a slightly different format. For example, the date_time is split into the days since the first observations, the hour of the day the observation, and the minute of the hour the observation. Generation sources are treated as factors, rather than groups. Additional variables are created out of squaring the irradiation, module temperature, and ambient temperature.

AC power, daily yield, and total yield are all removed from prediction. Though their previously recorded values could be incorporated with lag, AC power, daily yield, and total yield are all direct results of DC power, and do not make realistic predictors if this model were ever applied to future days.

Additionally, date time, plant id, and weather source are all removed due to redundancy. The fit is shown below.

Now, let us take a look at what the importance of various variables in building the model. The importance of each model, extracted with varImp, is shown in figure 36. Figure 36 clearly shows that irradiation, irradiation squared, module temperature, and module temperature squared dominate the predictions. This is in line with the previously constructed linear model.

Interestingly, ambient temperature remains a relevant predictor. And despite the early assumption that the time of day help little importance, it still makes an appearance on.

Day, minute, and generation source are all of negligible effect.

Figure 37 displays the predicted DC power for each generation source on the test set. Other collections of variables applied did not improve the result. For example, despite the generation sources showing little importance, removing them from the random forest training data was detrimental to predictions on the test set. The result of the random forest fit to the test set is shown in table 22

3 Results

3.1 Model Root Mean Squared Errors

The RMSEs of the prediction from each model are shown in table 23.

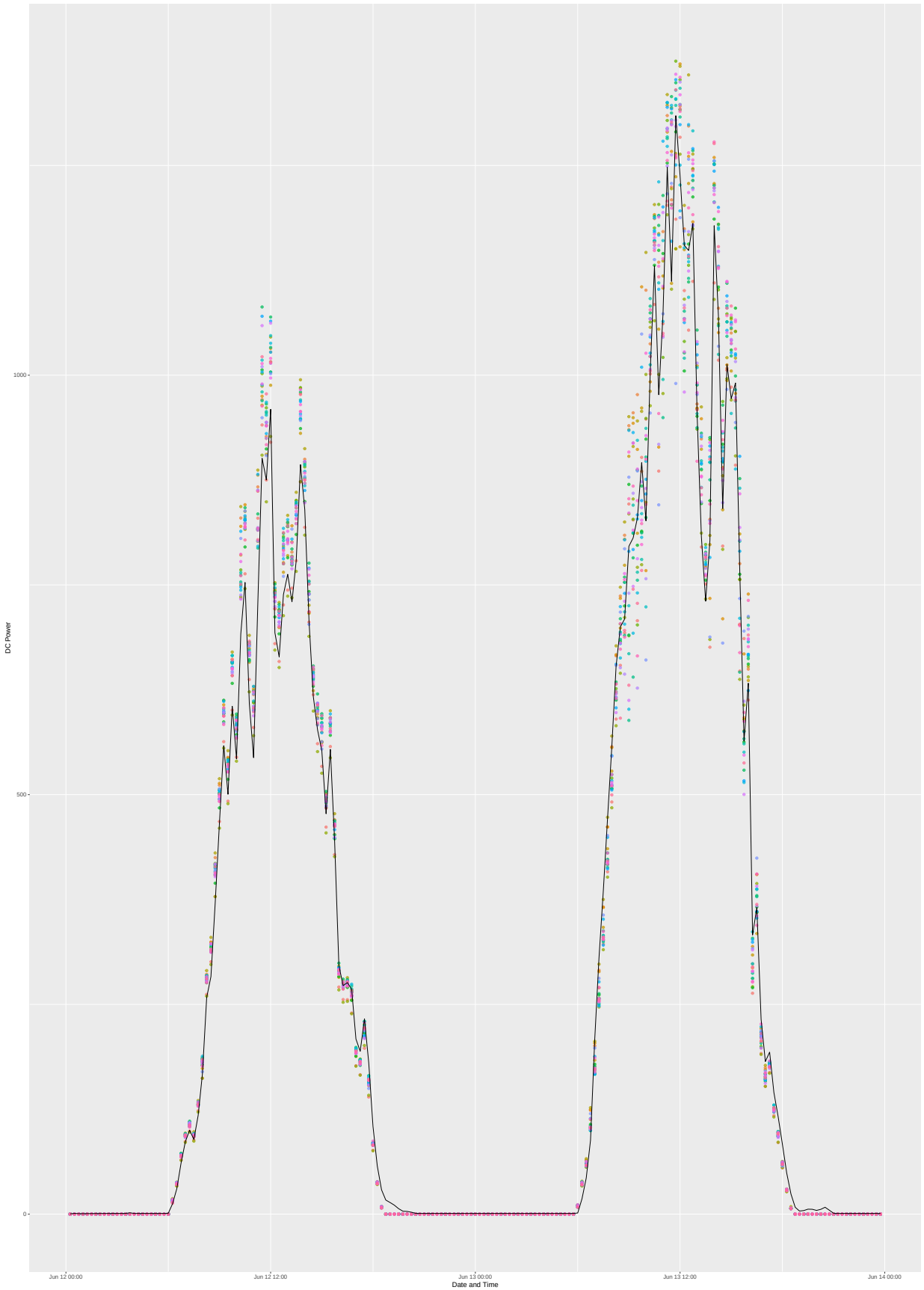

Figure 35: Arima Prediction for Test Days

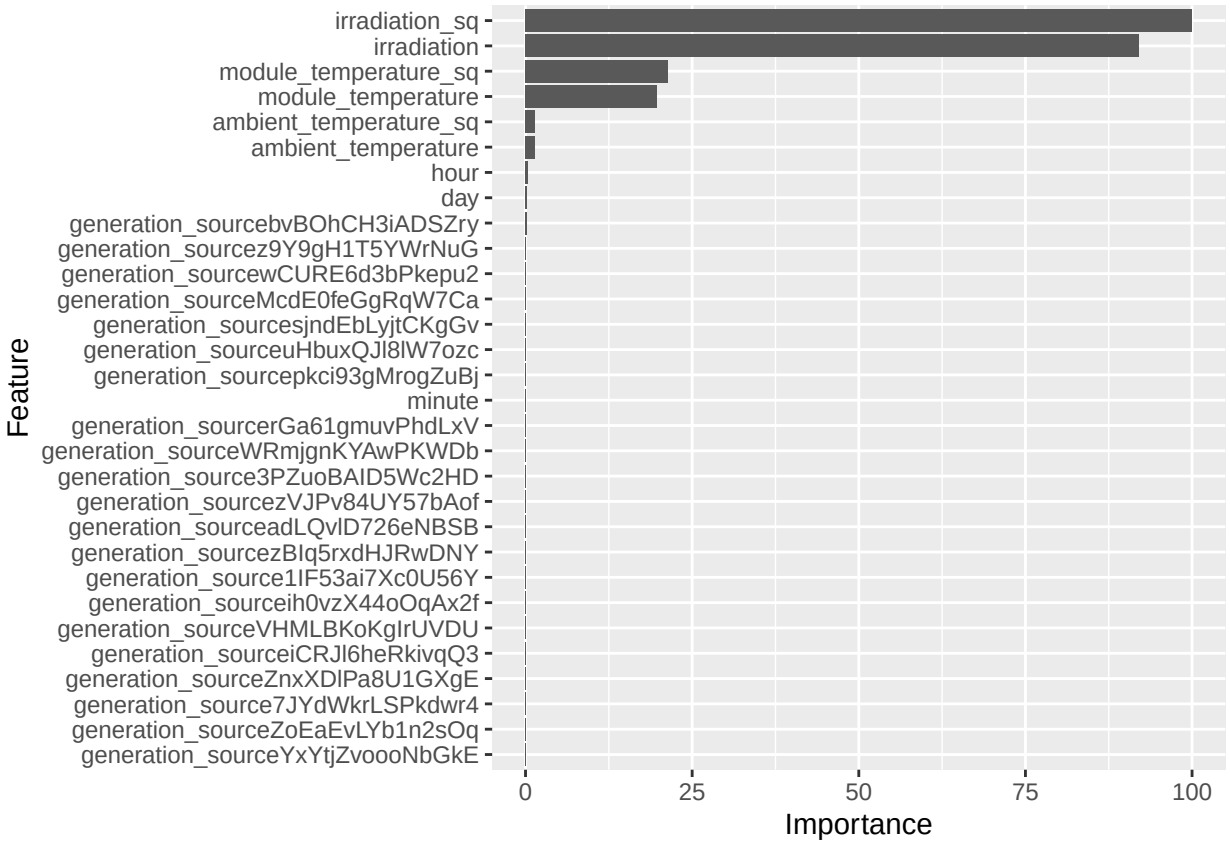

Figure 36: Importance of Features in Random Forest

Table 23: RMSE of Each Model on the Test Set

method	RMSE
Average	401.8
Average by Source	400.6
Irradiation Effect	44.4
Irradiation + Module Temperature Effect	40.0
Irradiation + Irradiation Lag Effect	40.6
Irradiation + Module + Ambient Temperature Effect	39.9
Clamped	39.7
Clamped + Anomolies Removed	39.2
ARIMA	179.1
ARIMA by Source	40.7
Random Forest	39.4

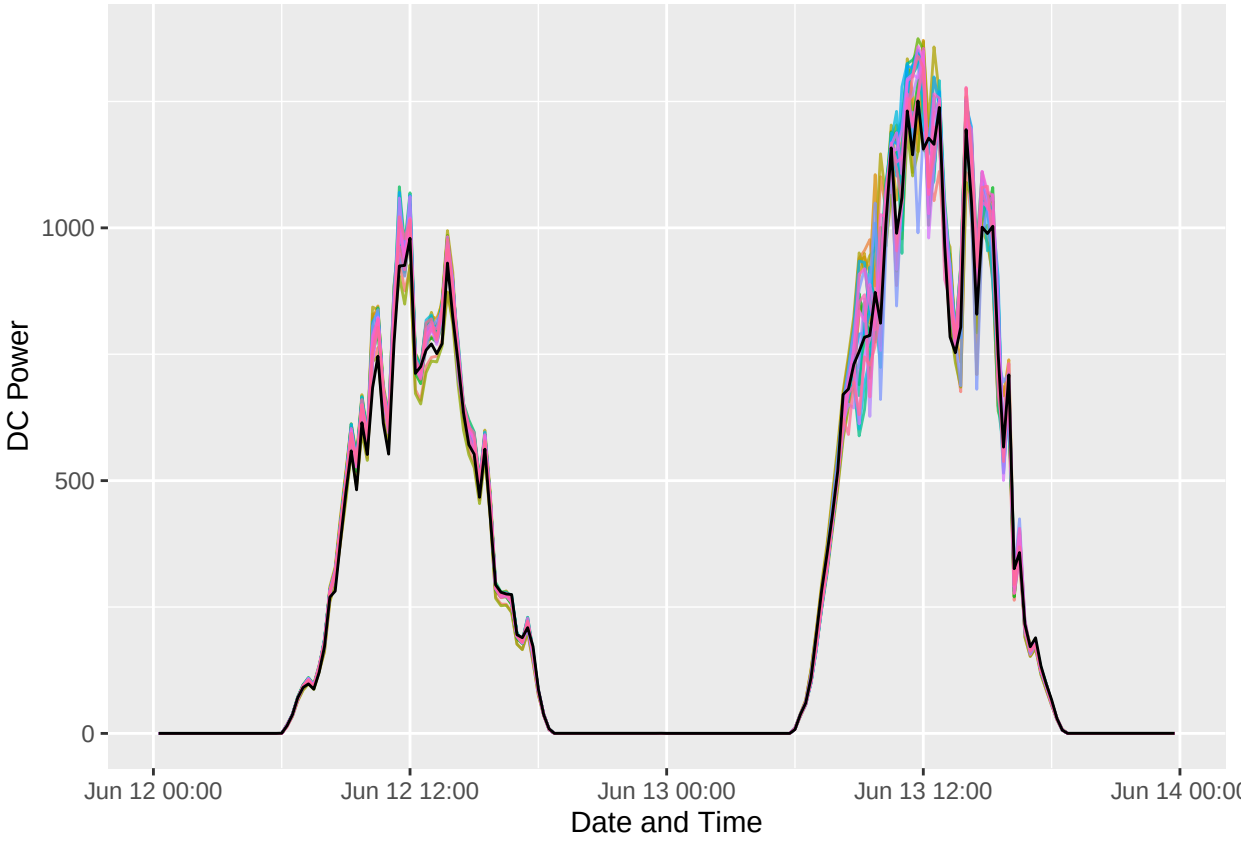

Figure 37: Random Forest Prediction on Test Set

Table 24: Validation RMSEs

method	RMSE
Linear Regression	62.3
ARIMA Forecast	58.3
Random Forest	59.1

3.2 Final Validation

Finally, the training data and test data were recombined into the solar data. The solar data was used to train a prediction to be applied to the validation data set, separated at the beginning of model development. The three models were the best fitting linear regression model, shown in figure 38, the ARIMA model based on the average DC power generated by the various sources, shown in figure 39, and the random forest model, shown in figure 40.

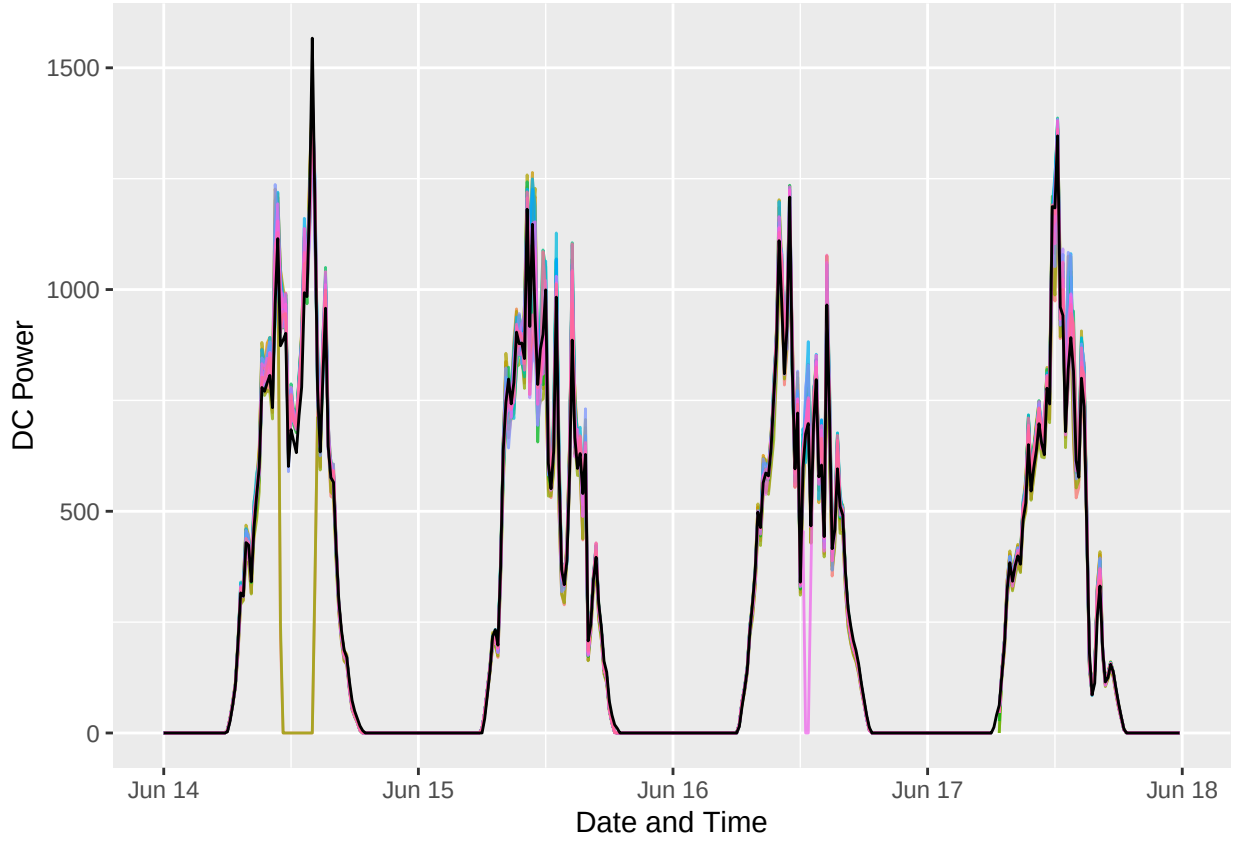

Figure 38: Linear Regression Predictions on Validation Set

The RMSEs of the various predictions are shown in table 24.

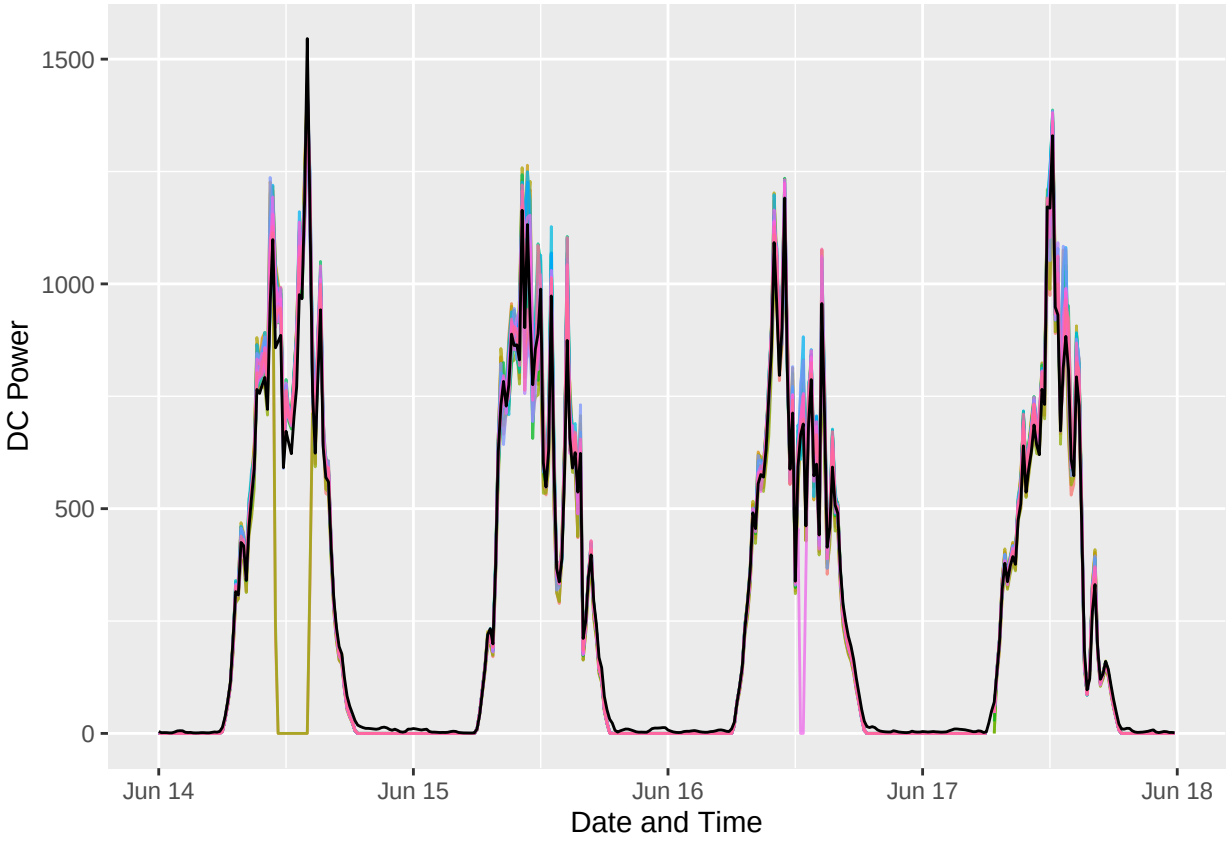

Figure 39: ARIMA Forecast on Validation Days

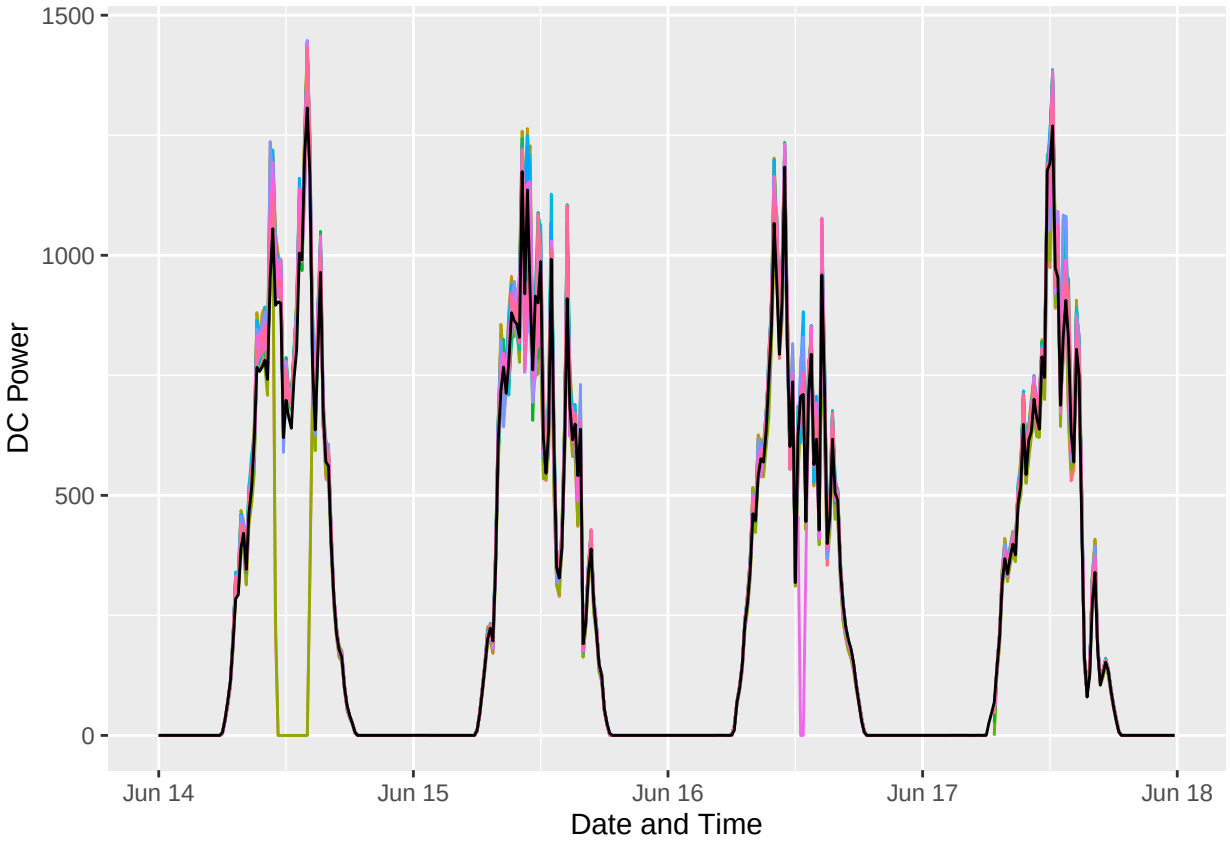

Figure 40: Random Forest Predictions on Validation Set

4 Conclusion

4.1 Summary

Exploratory analysis was conducted to see the patterns of the power generation data and the weather data. Sanity checks on the data revealed that the data produced by one of the plants would be very difficult to conduct analysis on. The problematic plant was removed from further explorations.

Correlations between different variables were explored. It was confirmed that AC power and DC power were highly correlated. The inverters at the plant displayed a 97.7% conversion rate from DC power to AC power. Module temperature was highly correlated to irradiation, especially after a 15 minute lag.

A validation set was separated containing the last 4 days of the data set. A linear regression model was tested and refined. An ARIMA model was applied to the data set. Finally, random forest was employed to achieve the best prediction. The random forest variable importance results reinforced the findings of the refined linear regression model.

Finally, the predictions were trained on all of the non-validation data and used to predict the validation data. The resulting RMSEs were higher for the validation set. The higher RMSEs suggest some amount of over training. However, it may be observed in the graphs of the fit that the validation data had missing values and sharp drops in DC power. Sudden plunges to 0 DC power suggest a sensor fault of some kind, which would negatively impact RMSE as they were not accounted for in the model.

4.2 Potential Impact

The refined linear regression model and the random forest predictions both produced reasonable improvements over the most basic linear regression model built solely on irradiation. A plant seeking to improve their predictive abilities on DC Power, and thus the AC power that the plant contributed towards an electric grid, could use either model to great effect.

4.3 Limitations

Unfortunately, the noisiness of missing or anomalous observations at plant 4136001 made predictions very difficult. Without cleaner data with which to work, making accurate predictions for the DC power output of plant 4136001 was outside the scope of this analysis.

The weather sensor data was fairly limited. As was noted in the importance of variables determined by the random forest, the hour of an observation has some observable impact on the DC power. Time is expected to be a confounding variable of weather, not a driving variable in itself. Thus, the hour being a non-negligible variable suggests that there were aspects of the weather or surrounding environment that were both influenced by the time of day and also impacted DC power.

The data included in this data set covered only 34 days. Multiple months or even many years of data could possibly reveal greater periodicity in the data. Additionally, since plant 4136001 appears to have been in operation longer, it may reveal when plant-wide outages appear.

4.4 Future Work

While data collected at plant 4136001 did not serve well to make predictions on the DC power, the data could serve well to develop a system to predict sensor faults before they occur. It could be possible that irradiation levels, temperature, or a host of other observable variables correlate to a missing observation or anomalous data.

Treating the data as a time series and accounting for lag revealed lagged correlation between irradiation, module temperature, and ambient temperature. Irradiation and ambient temperature may be predicted in advance, to some extent, from a meteorological organization. Module temperature, however, is specific to the plant. Thus, developing a model to predict module temperature from other future-predictable variables allow more actionable results.

More work could be done to incorporate predictable weather data. That is, data could be queried from other data bases to find information about wind speed, cloud cover, precipitation, etc. A more sophisticated model could then be constructed that requires only the predicted weather that day to make a reasonable estimation of the DC power the plant would produce. This is the best case scenario for the plant, as they could use publicly available weather data to predict the impact on the electrical grid each day.

In this analysis, three models were built: Linear Regression, ARIMA, and Random Forest. Other methods were explored, for example Principle Component Analysis, but were found to be less effective than even predicting the average. A powerful approach for future work would be to assemble an ensemble model combining models that consider a-temporal variables, like the Random Forest, and a model that acknowledges the time series qualities of the data, like ARIMA.