Solar Case Study

Gabriel Adams & Spencer Ebert

Brigham Young University

Abstract.

1 Background

1.1 Goals of Analysis

Solar power has become more prominent in the past decade and has led to many households installing it. This increase in solar power is a result of the decrease of cost to install (it pays for itself in less time), and the environmental benefits to clean energy. Solar power companies want to be able to predict power output for upcoming years for their customers and also understand how solar panels degrade over time. Predicting power output is difficult for many factors including the season, angle of the panels, daily weather, etc. The purpose of this analysis is to look at solar output trends throughout the year. Understanding power output for customers and how solar panels degrade over time can help with future sales and improve on their solar panel quality. The two main goals of this analysis are 1) analyze how the solar panels are degrading over time and 2) obtain projections of power output for this particular customer for the next year.

To answer the goals of this analysis, we looked at solar output data for an individual customer from 2015-2017. For each individual day kilowatt hours (kWh) was measured. A graph of the data is shown in Figure 1.

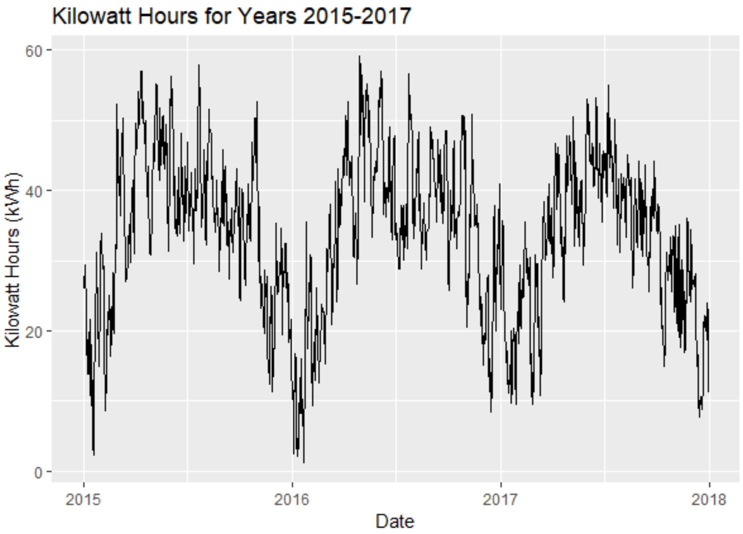

Fig. 1: Time Series of Data

At first glance of the data, there appears to be a seasonal trend from year to year with variability. When analyzing the data it is important to take into account the nonlinear relationship between day and kWh because otherwise the model won't predict well and won't capture the seasonality. The data also looks like it is correlated from day to day (one day influences the next day and so on). Figure 2 shows the autocorrelation of kwh from the solar data.

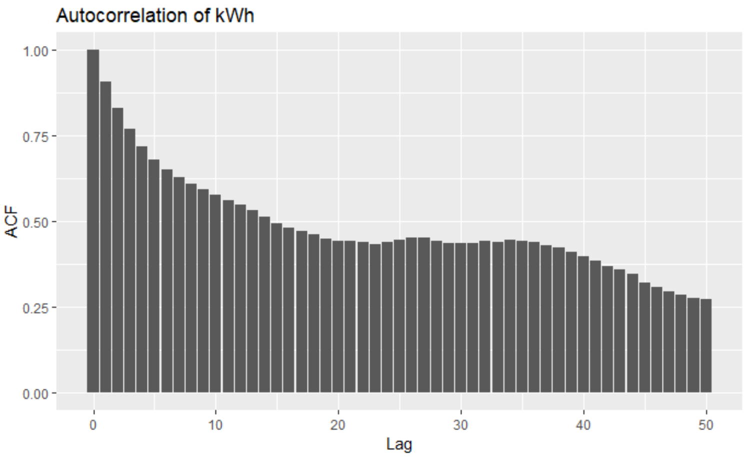

Fig. 2: ACF

It is clear in this ACF plot that there is autocorrelation for values of kwh and it appears that the correlation gradually goes down over time. A normal linear model assumes that there is no correlation between the residuals so in this case a normal linear model won't capture that correlation. Ignoring the correlation doesn't change the regression estimates but it does result in the confidence and prediction intervals being inaccurate and also future predictions being off. In this context, it makes sense that the residuals are correlated. The power output for each day should be correlated with past days. If there is one sunny day (higher kWh) then there is a greater chance that the next day will be sunny (similar kWh). The days affect each other.

To help overcome these obstacles we fit a model that takes into account the nonlinear relationship and the correlation structure as discussed in the next section.

2 Model Fit

In order to account for correlation and to properly answer our research questions, we fit the following model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\boldsymbol{\epsilon} \sim AR(1)$$

which means

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \vdots & \phi^{T-2} \\ \vdots & \vdots & \ddots & & \vdots \\ \phi^{T-1} & \cdots & & & 1 \end{pmatrix}\right)$$

$\boldsymbol{Y}$ is a vector of the response variable (KWH). $\boldsymbol{X}$ is a matrix containing information from our explanatory variables. Its first column is a column of 1's, corresponding to the intercept. The next 11 columns correspond to 11 months of the year (February through December). January is included in the intercept. A 1 in any of these columns indicates that the specific observation was taken during that month. A 0 indicates otherwise. The last column of $\boldsymbol{X}$ corresponds to day, which is a integer from 1 to 1096 indicated which ordinal day the measurement was taken. $\boldsymbol{\beta}$ is a vector of coefficients. The first element of $\boldsymbol{\beta}$ is the intercept. The next 11 elements correspond to the effects of the months (February through December), while the last element corresponds to the effect of ordinal day.

Because our observations come from a time series, the residuals will be correlated. To account for this correlation, we determine that they follow an auto-regressive process of order 1. Essentially, this means that residuals that are close in time will be highly correlated with each other. As residuals go farther apart, this correlation will lessen. We can further describe this relationship as:

$$\epsilon_t = \phi\epsilon_{t-1} + \omega_t$$

where

$$\omega_t \sim \mathcal{N}(0, \sigma^2)$$

ϵ_t is the residual for each observation, while ϵ_{t-1} is the residual for the previous observation. ϕ is the correlation (or effect) of the previous observation. ω_t is random noise.

Using this model will help us answer our research questions and will properly account for the correlation present in the data. By building correlation into the structure of the model itself, our standard errors and intervals will be more accurate. We will be able to determine if the system is deteriorating over time. If it is, we will be able to quantify this deterioration. We will also be able to predict power output into the future using this model, although we will need to be cautious due to extrapolation.

To properly use this model, several assumptions will need to be justified. Our model should account for nearly all the correlation in the data. We will check this using an ACF plot. After accounting for correlation, residuals should be normally distributed and a Residuals vs. Fitted Values plot should show equal variance about the line. We will also need to check if there is a linear relationship between day and kwh. We will verify these assumptions in the following section.

3 Justification & Performance

While determining which model to use for this analysis, we considered several different options. In order to answer the research question regarding an overall downward trend, we would need to somehow quantify that trend with regards to time. Due to the cyclical nature of the data, we hoped our model would be able to capture the "waves" in the data. Finally, because our data are from a time series, correlation would be an issue. Our model would have to properly account for that correlation. In order to capture the overall downward trend, we decided to include a "day" variable in our model. In order to capture the wavy nature of the data, we decided to include "month" as a factor variable. In order to account for correlation, we investigated using AR(1), MA(1), and ARMA(1) processes. To decide between these three, we aimed to minimize AIC. We chose AIC because it has a emphasis on predictive abilities, which is a focus of our analysis. Using an AR(1) model resulted in an AIC of 6398.7, a MA(1) model had an AIC of 6791.2, and an ARMA(1,1) model had an AIC of 6400.7. Because the AR(1) model had the minimum AIC, it is the model we chose to use for this analysis. Additionally, an AR(1) model captures correlation that gradually wears out over time. This type of correlation is present in our data and so it makes sense to use an AR(1) model here.

We now justify the assumptions listed in the previous section. We first verify that correlation has been properly accounted for in our model. To verify this, we look at an ACF plot (Figure 3). We can see that auto-correlation has been greatly reduced. Observations are perfectly correlated with themselves while different observations do not appear to have any notable correlation. We can see in a histogram of standardized residuals (Figure 4) that the residuals, after decorrelation, are normally distributed. By looking at a Residuals vs. Fitted Values plot (Figure 5), we can verify equal variance. The spread of the residuals is fairly constant as the fitted values increase.

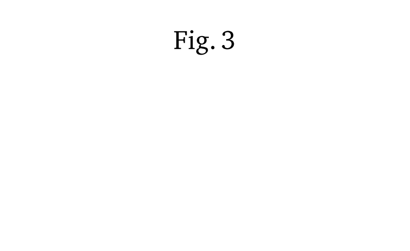

Fig. 3

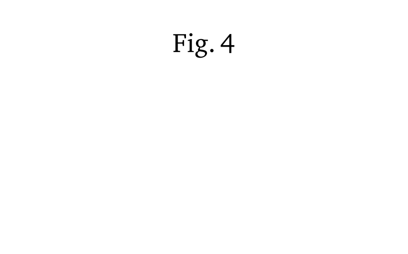

Fig. 4

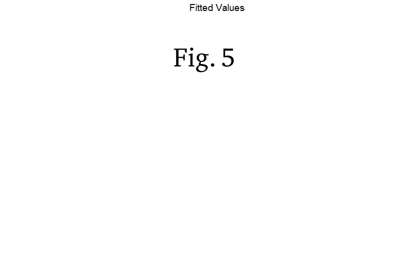

Fig. 5

To check the linearity assumption, we looked at added variable plots when fitting a normal linear model to the data (the correlation doesn't affect the linearity assumption). It's in Figure 6

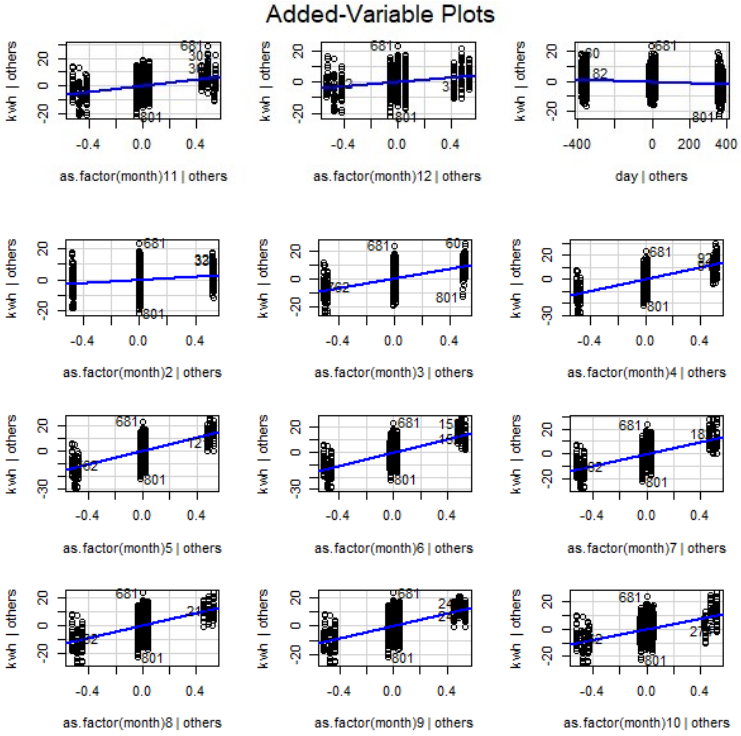

Fig. 6: AV Plot

From these plots it looks as though there is a linear relationship of kwh with the other variables so we are confident in this assumption.

We also wanted to measure how well our model fit the data so we computed R^2. Computing R^2 in this context is a little more tricky because we have correlated data. To do this, we decorrelated our Y and X matrices by premultiplying them by the inverse of the lower cholesky decomposition of the correlation matrix R. (L is the lower cholesky decomposition matrix.)

$$L^{-1}Y \sim N\left(L^{-1}X\beta, \sigma^2 L^{-1}LL'L'^{-1}\right) \tag{1}$$

By using this formula our variance matrix ends up being just $\sigma^2 I$ so we can now fit a normal linear model with $L^{-1}Y$ as the adjusted kwh (response) and $L^{-1}X$ as the adjusted variables for month and days (explanatory variables). After fitting the linear model we got $R^2 = 0.7279$ which means that our model accounts for 73% of the variability in the data.

In order to assess the predictive capabilities of our model, we ran a cross-validation study. We fit a model to the first two years of data (2015 and 2016) and used it to predict solar power output in 2017 (Figure 7). The predicted line is shown in red, the actual line is shown in black, and the prediction interval is in blue. Through this process, we determined that the RPMSE of our model is 7.4273. This means that our predictions for KWH are typically 7.4273 units ”off” their actual values. We determined that the average prediction interval width is 32.3097. This means that our model will typically predict KWH within $\frac{32.3097}{2} = 16.1594$ units. Our intervals are quite wide on the scale of the data. This makes sense, however, due to the extreme fluctuations present in the data. In order to capture that noise, the intervals need to be wide. Finally, we determined that our model has coverage of 0.9671. This means that our 95% prediction intervals are performing like 96.7% prediction intervals. Basically, they are slightly overperforming. This is not too worrisome, however, so we will continue with our analysis.

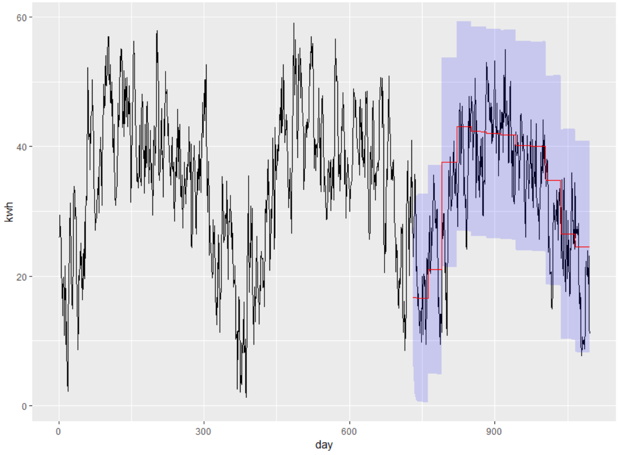

Fig. 7: Cross Validation

4 Results

By using a linear model with month as a factor and an autoregressive 1 correlation model we were able to achieve the goals of this analysis.

4.1 Degradation Over Time

To understand the degradation over time we looked at the coefficient for Day in our model. All the estimated coefficients along with corresponding 95% confidence intervals are found in Table 1.

	Estimate	Lower	Upper
Intercept (includes January)	19.0497	15.2490	22.8504
February	3.8504	-0.0663	7.7671
March	19.3555	14.9257	23.7853
April	24.3966	19.7954	28.9979
May	24.7228	20.0880	29.3576
June	25.6901	21.0167	30.3636
July	25.4852	20.8147	30.1556
August	23.1473	18.4627	27.8319
September	21.7675	17.0551	26.4800
October	17.5384	12.8537	22.2232
November	9.3724	4.7591	13.9856
December	6.0685	1.8152	10.3218
Day	-0.0040	-0.0082	0.0002

Table 1: Note: For the month coefficients January is the intercept and then the other months are the difference from January.

Our correlation estimate ϕ was 0.7924 meaning that each day was correlated with another day by ϕ^{Δ} (Δ is difference in days). The 95% confidence interval for ϕ is from 0.7494 to 0.8292.

The estimated day coefficient was -0.004. This means that for every day we expect kilowatt hours to go down by 0.004 with a 95% Confidence Interval from -0.00825 to 0.00016. The p-value for the effect of day is a little above 0.05 (0.0596) but it is still close enough that we are comfortable saying that there is a slight degradation for solar panel power output over time. Our degradation looks small here but makes more sense when you look at degradation over the year. A rate of -0.004 per day corresponds to -1.47 kwh per year. This degradation is small which is a good sign for solar power companies because that means they don't have to replace their solar panels as often. More data from other individuals would be useful to understand the degradation.

4.2 2018 Projection

To come up with predictions for 2018 it is important to still take into account the correlation from day to day and how those values are correlated in the future. This equation shows the assumed distribution of our kWh values including future kWh values (Gaussian Process).

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} = \begin{bmatrix} Y \\ Y(T+1) \\ \vdots \\ Y(T+K) \end{bmatrix} \sim N\left(\begin{bmatrix} X\beta \\ X^*\beta \end{bmatrix}, \sigma^2 \begin{bmatrix} R_Y & R_{Y,Y^*} \\ R_{Y^*,Y} & R_{Y^*} \end{bmatrix} \right)$$

Y are the kWh values in our data and Y^* are the future kWh values. R is the correlation matrix for our values and as can be seen here there is correlation between the future days and the past days. Our model takes this into account but if we want to predict future values we need to take this correlation into account. To compute our predicted kWh values for 2018 we used

$$E(Y^*|Y) = X^*\hat{\beta} + \sigma^3 R_{Y^*,Y}(\sigma^2 R_Y)^{-1}(Y - X\hat{\beta})$$

$E(Y^*|Y)$ are our predicted values for 2018. Basically, this formula doesn't just compute $X^*\hat{\beta}$ it takes into account the correlation because our Y^* is conditioned on Y. By using this equation we came up with predicted values shown in figure 8.

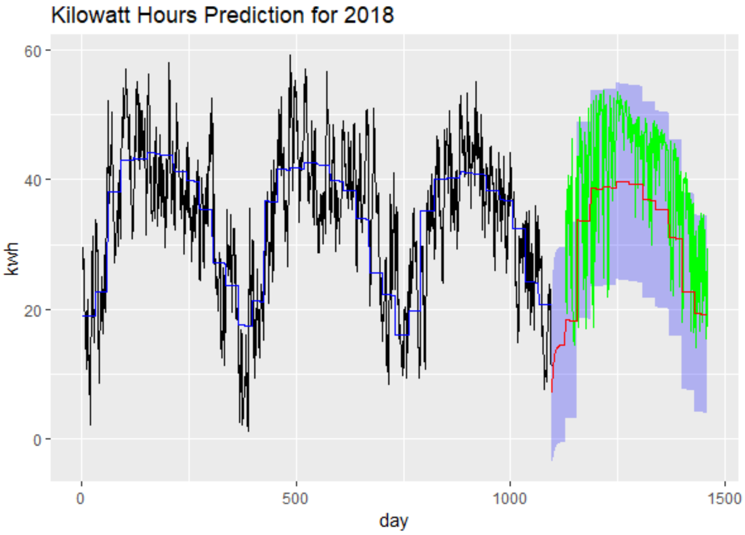

Fig. 8: The green line is the actual kWh values for 2018 and the red line is the predicted values. The light blue shaded area is the prediction interval for that year.

As can be seen in figure 8 it looks like our predicted values are lower than the true values for 2018. At first glance this is a cause for concern for our model, but when thinking about what happened that year it makes more sense. 2018 was one of the sunniest years in Provo and as a result led to greater solar power output. Our data in past years couldn't take into account this sudden change in weather so it makes sense that the model would under predict. The RPMSE for year 2018 is 10.48, which means that the prediction was off by about 10 kWh for each day on average. A larger RPMSE makes sense in this case because of the change in weather for the year.

5 Conclusion

This analysis adequately addressed the research goals of this study. We were able to predict power output for this individual's solar panels for 2018 and we were able to explore how the panels degrade over time. We found that there was a slight degradation from year to year by analyzing our coefficients and we computed the 2018 power output by using Gaussian Process Regression with an autoregressive 1 correlation structure.

One of the shortcomings in our analysis was extrapolation for our predictions. The data we had for 2015-2017 wasn't able to show the random increase in sunlight for 2018, which led to our predictions being lower than the actual values. Another issue with the analysis is the use of only the dates to predict kWh when other variables could be useful in prediction.

Going forward, it would be interesting to examine the effects of other variables like weather, temperature, etc on kWh. It would also be interesting to look at multiple individuals' solar panel data so we could compare trends between different people.

6 Teamwork Statement

Spencer primarily worked on the Introduction, Results, and Conclusion sections. He also helped with the other sections. Gabe primarily worked on the Model Fit and Justification and Performance sections. He also helped a bit with the Results section.