Analyzing Correlated Solar Data (Case Study 5 - STAT 536)

Dean Sobczak and Jeremy Meyer

March 14, 2019

Abstract

This report contains the methods, assumptions, and results of statistical analysis performed on a solar power data set. Specifically, the authors develop an autoregressive lag 1 process model to help them answer certain research questions posed a power company. The questions include (1) how much the panels are degrading over time and (2) how much power the panels are projected to generate over the following year (2018). The results and conclusions of this study have the potential to help the power company better plan for customer power needs as more homeowners are likely to adopt solar technology in the coming decades. The results can also help homeowners predict the lifetimes of their solar panel systems.

1 Introduction

From films such as Al Gore's "An Inconvenient Truth" to political pacts like the 2015 Paris Agreement to reduce greenhouse gas emissions, advocates have raised global awareness of climate change. Most scientists today believe that human activities (especially those related to burning fossil fuels like coal and gasoline) have contributed to a rise in the global average temperature in the past several decades. Researchers posit that such an increase can lead to rising sea levels and extreme weather. Since virtually all forecasts for the health of the planet are grim by climate change proponents, policymakers have begun taking measures to reduce their countries' dependence on fossil fuels for energy. Some states in the U.S. even offer tax incentives for homeowners who adopt solar power, since solar panels on someone's roof do not emit the greenhouse gases linked to climate change.

However, solar power is not a panacea for climate change. Although they are more environmentally friendly than a coal power plant, solar panels provide fluctuating amounts of energy due to factors outside of a homeowner's control (e.g., season, panel angle, sun direction, cloud cover, etc.). In addition, current solar technology is not perfect because the panels presumably degrade and generate less power over time. Therefore, most people cannot rely fully on solar power for their needs, so they pay a utilities company to provide power through more traditional sources to make up the difference. Therefore, power companies are interested in better understanding the solar panel degradation process so that they can more accurately predict a homeowner's supplemental power needs (i.e., those that are beyond what their solar panels provide) over the next year. Homeowners are also interested in quantifying the degradation (if there is any) because then they can then estimate the lifetime of their expensive solar panel system.

1.1 Exploratory Data Analysis

The stakeholders to this study hope that a data set containing kilowatt hour (kWh) measurements from a single solar panel system mounted on a home will enable them to achieve their goals. The data set provided for this analysis contains 1,096 observations for each day during January 2015 to December 2017. The data set only has two variables, Date (the date of the measurement) and kwh (the amount of kilowatt hours generated by the solar panel system for a given date). Each observation is equally spaced apart in time. We will focus our exploration in this section on the univariate (one variable)

distribution of the kilowatt hour observations and how those observations possibly change over time. If the measurements do change over time, we will probably have to use a different model than the traditional linear regression model because that model assumes that observations are independent with respect to time.

To investigate the univariate distribution of the kilowatt hour observations, we produced the histogram in Figure 1 below.

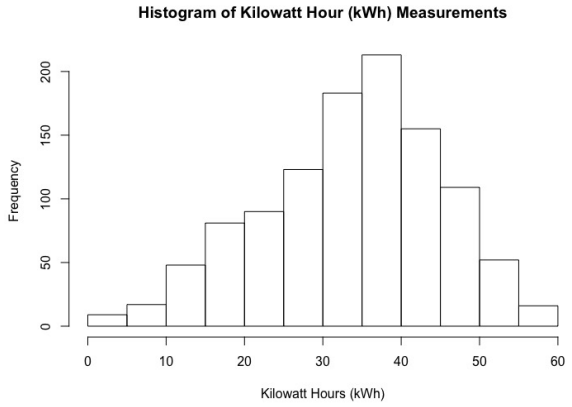

Figure 1: Histogram of 1,096 Kilowatt Hour (kWh) Measurements from 2015 to 2017

The kWh measurements look quite normally distributed. The histogram has the familiar bell-curve shape of the normal distribution, albeit somewhat left-skewed (i.e., it has a longer left tail). However, we are not very concerned about this slight skew, so we should be able to use a statistical model (such as the autoregressive lag 1 model) that assumes the response variable is normal.

We also explored the relationship between observation date and the kWh generated. In other words, we examined how the kWh produced by the solar panels changed over time. From the information given by stakeholders at the outset of this analysis, we expected to see seasonal fluctuations in kWh and a decrease in the average kWh from 2015 to 2017. Figure 2 below seems to confirm our guess.

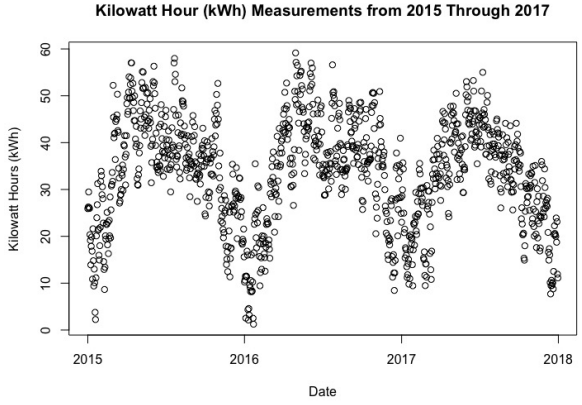

Figure 2: Scatterplot Between Kilowatt Hour (kWh) Measurements and Date

Clearly, we cannot justifiably use a regular linear regression model to analyze this data. The scatterplot shows an obvious pattern of increases and decreases in kWh over the three years recorded, so a single line would not adequately explain the relationship between kWh and Date. Furthermore, the pattern is curvilinear (i.e., resembling a curve), with each year's observations producing an arc or upside-down "U," so drawing a line through these points would be inappropriate. Both of these facts together suggest that solar panel output is seasonal, and the fact that the tops of the yearlong arcs appear to decrease year-over-year hints at the degradation of the solar panels over time.

We also created the same plot but connected all of the points to highlight the daily oscillations in kWh that are hard to see in Figure 2. See Figure 3 below.

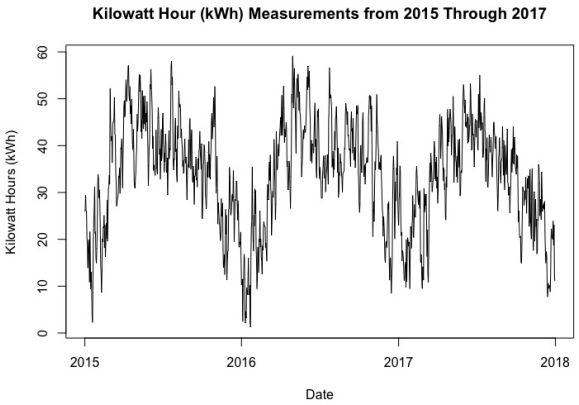

Figure 3: Line Plot Between Kilowatt Hour (kWh) Measurements and Date

1.2 Dealing with Correlated Data

From these plots, we believe that we will need to develop a sophisticated model that will account for both the macro and micro trends in kWh over time when we analyze this data to answer the posed research questions. Luckily, such relationships can be appropriately modeled with a linear model that accounts for the correlation between the kWh observations. Correlation is a measure of the nature and strength of the linear relationship between two variables. Correlation values can be between -1 and 1, where -1 indicates perfectly negative correlation (i.e., when X is "high," Y tends to be "low") and 1 indicates perfectly positive correlation (i.e., "high" values for variable X correspond with "high" values for variable Y). A correlation of 0 suggests no correlation, meaning the best line between the two variables on a scatterplot would be a horizontal one. In this data set, not much correlation seems to exist between Date and kWh. Nevertheless, each kWh observation seems correlated with observations that came before it, a phenomenon known as autocorrelation.

Correlated data appear in a variety of real-world contexts, especially if the data sets contain temporal and/or spatial information. Our intuition suggests that the because solar panels generate power by absorbing sunlight and converting it into electricity, a kWh value should be highly related to the daily weather (e.g., sunny, cloudy, rainy, snowy, foggy, hazy, etc.). Since weather patterns tend to persist for weeks or months at a time at a given location, we expect, for example, that a kWh measurement on June 25, 2016 should be similar to one from June 15, 2016. Thus, we should not treat our 1,096 kWh as independent observations because each measurement likely does not provide one unique "piece" of information (i.e., the kWh output for a particular day is in some ways a function of what the kWh output was for possibly the last few weeks).

Autocorrelation function plots are often helpful when exploring possible correlation among one variable's observations and seeing how many observations likely influence a particular observation's value. We have included such a plot for the kWh observations from 2015 to 2017 below in Figure 4.

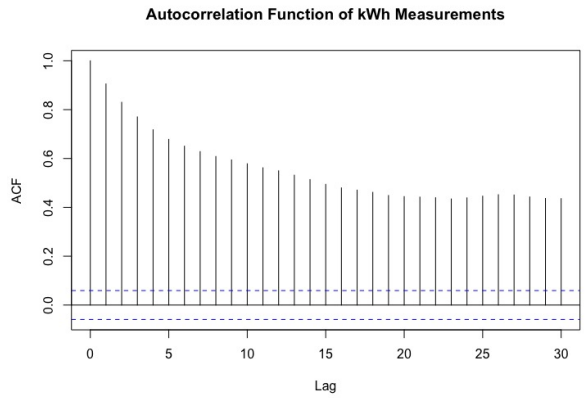

Figure 4: Autocorrelation Function Plot of the Kilowatt Hour (kWh) Measurements

Clearly, strong correlation exists among kWh values in our data set. This correlation lingers even after going back thirty days in time (denoted as lag 30). Thus, we would be naïve to use a standard linear regression model to analyze this data because such a model assumes that the autocorrelation for the kWh values is 0.

While we could still fit a linear regression model with no autocorrelation on this data and get unbiased estimates for our regression coefficients (i.e., our estimates would be correct on average), the standard errors would be too small for such estimates. This is because we would be assuming we had more pieces of information than we actually had in our data set. Stated differently, we would be using all N observations in the data set to estimate variances for regression coefficients instead of correctly using some number of observations less than N in our computations. Therefore, any confidence or prediction intervals that we calculated from such a model would be too small and not attain the expected confidence level. For this reason and based on the plots above, we developed an autoregressive lag 1 (AR(1)) model to address the correlation in the solar data and achieve the goals of the analysis.

2 Model Description

From our exploratory data analysis, we determined that a linear model that incorporated a non-zero, kilowatt hour correlation component would be suitable for this analysis. While a variety of models exist for accounting for autocorrelation among response variable values, we chose to focus on an AR(1) model. We will justify this model in more depth in the following section.

The AR(1) model is a special case of an autoregressive process, which is actually a class of Gaussian processes. Gaussian (or normal) processes are widely used in time series modeling and spatial statistics applications because of their nice theoretical properties. At a high level, the Gaussian process data model assumes that all data that one might be interested in analyzing follows a multivariate normal distribution. That is, all response variable values are jointly distributed as normal. In the time-correlated observations context, this means that all data past, present, and future is believed to be a finite sample of arbitrary length from a normal population with a certain mean and variance/covariance structure. Under this data model, it is not difficult to generate predictions for future observations because we can simply use the conditional distribution of the future observations given the past ones.

To explore this in more mathematical detail, consider the following equation:

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

Equation 1 is the general data model for a Gaussian process. In this equation, $\boldsymbol{Y}$ is a vector of the observed (or past) T observations of kWh, and $\boldsymbol{Y^*}$ is a vector of the unobserved (or future) K observations of kWh. These two sets of kilowatt hour values together, $\tilde{\boldsymbol{Y}}$, jointly follow a normal distribution with a partitioned mean vector of ($\boldsymbol{X\beta}$,$\boldsymbol{X^*\beta}$)'. $\boldsymbol{X}$ represents the design matrix, which contains a column of 1's for the intercept and columns for time in days (an index variable from 0 to the number of observations in the data set, N, minus 1), the proportion of the year that has passed, and the square of that proportion (we will explain these explanatory variables more in section 3.1). $\boldsymbol{\beta}$ represents the vector of coefficients, or effects, of each explanatory variable on kilowatt hour production (e.g., as time increases by 1 day, we expect that the number of kilowatt hours produced will increase on average by β_{time}).

The variance/covariance matrix of $\tilde{\boldsymbol{Y}}$, $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}$, is partitioned into four parts. $\boldsymbol{R_Y}$ is the correlation matrix of the observed kilowatt hour observations ($\boldsymbol{Y}$); $\boldsymbol{R_{Y^*}}$ is the correlation matrix of the unobserved kilowatt hour observations ($\boldsymbol{Y^*}$); and $\boldsymbol{R_{Y,Y^*}}$ is the correlation matrix between the observed and unobserved kilowatt hour values. These sub-matrices are multiplied by the scalar, σ^2, to produce the variance/covariance matrix needed to identify this normal distribution. σ^2 is the variance of the residuals, or deviations of kilowatt hours from their mean (the mean is either $\boldsymbol{X\beta}$ or $\boldsymbol{X^*\beta}$ depending on whether the observations are observed or not).

One of the remarkable properties of the multivariate normal distribution is that each conditional distribution derived from it is also normal. This makes the multivariate normal distribution one of the most tractable distributions to deal with analytically. Thus, the predicted values for the unobserved kilowatt hours in this Gaussian process can be obtained with the following equation:

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

In Equation 2, all Greek and Latin letters have the same meaning as in Equation 1. This equation basically states that the expected value of the unobserved kilowatt hour measurements given the observed measurements is the mean of the future measurements plus some adjustment based on the residuals of the past measurements, which are correlated with the future measurements.

Although the variance/covariance matrix of the response variable, $\boldsymbol{\Sigma}$, identifies the normal distribution that we assume the kilowatt hour measurements follow, statisticians often tend to work with the correlation matrix, $\boldsymbol{R}$, instead. Each element of $\boldsymbol{R}$ comes from applying a correlation function so that the ij^{th} element of $\boldsymbol{R} = \rho(t_i, t_j) = Corr(y_{t_i}, y_{t_i})$, or the correlation between two kilowatt hour observations at times i and j.

As mentioned above, the AR(1) model we chose for this analysis is a special case of an autoregressive process (which is a type of Gaussian process). Autoregressive (AR) processes differ from other Gaussian processes such as moving average (MA) processes because every observation is correlated with every previous and future observation (even if that correlation is small).[1] In general, an AR(p) process results in each observation being correlated with its neighboring p observations (which is called lag p), with a potentially different correlation coefficient attached to each lag. In the AR(1) model, $Corr(y_{t_i}, y_{t_i}) = \phi^{|t_i - t_j|}$ is the correlation function, so we only need to estimate one correlation parameter, ϕ, to model the correlation between any two observations. Researchers in this area generally model the correlation in an AR(1) model in terms of the residuals, not the raw response variable values. This modeling is illustrated in Equation 3 below.

$$\epsilon_t|\epsilon_{t-1} \sim N(\phi\epsilon_{t-1}, \sigma^2) \qquad \text{or} \qquad \epsilon_t = \phi\epsilon_{t-1} + \omega_t, \omega_t \overset{iid}{\sim} N(0, \sigma^2) \tag{3}$$

These equations make it clear that the residual at time t, ϵ_t, is a function of the previous residual in time, ϵ_{t-1}. While the two statements in Equation 3 are equivalent, we will focus on the statement on the right for interpretation. The previous residual in time acts like an explanatory variable with weight ϕ, and each ω_t adds some random normal variability to each residual. Notice that the variance of these ω_t's is σ^2—the same σ^2 that appears in Equation 1. ϕ can thus be interpreted as the correlation between two

[1] To be fair, exponentially correlated processes also have this same property, but an exponential correlation structure is the same as an AR(1) one if all the time intervals are equally spaced in the data set.

residuals (or the lag 1 correlation). Putting all this together, we obtain the distribution of all kilowatt hour residuals in the AR(1) model in Equation 4 below.

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_T \end{bmatrix} \sim N\left(\mathbf{0}, \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & ... & \phi^{T-1} \\ \phi & 1 & \phi & ... & \phi^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi^{T-1} & \phi^{T-2} & ... & ... & 1 \end{bmatrix}\right) \tag{4}$$

Note that the matrix scaled by σ^2 is the correlation matrix, $\boldsymbol{R}$. Thus, because ϕ is a correlation between -1 and 1, all off diagonal elements will have a magnitude less than or equal to 1. In other words, every residual and thus every kilowatt hour observation will be correlated with every other observation, but observations very far away in time will be nominally correlated. Combining Equations 1 and 4, we obtain our AR(1) model for this analysis.

$$\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & ... & \phi^{T-1} \\ \phi & 1 & \phi & ... & \phi^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi^{T-1} & \phi^{T-2} & ... & ... & 1 \end{bmatrix}\right) \tag{5}$$

In summary, we assume that the solar data provided in this study was generated by a Gaussian process, specifically an autoregressive lag 1 process, with mean $\boldsymbol{X\beta}$ and variance/covariance matrix $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}$, as shown in Equation 5. The model assumes linearity (i.e., all explanatory variables have a linear relationship with kilowatt hour production), normality (i.e., the residuals that fall above and below the fitted line $\boldsymbol{X\beta}$ follow a normal distribution with mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{R}$), and equal variance (i.e., the spread of the residuals along the entire fitted line is the same). We also assume that our correlation structure will account for all correlation among kWh values. So, if we were to "decorrelate" our observations via a transformation, then the resulting new vector of kWh measurements should appear independent.

We believe this AR(1) model will enable us to (1) help power companies better under the solar panel degradation process and (2) accurately predict a homeowner's supplemental power needs (i.e., those that are beyond what their solar panels provide) over the next year. We will achieve (1) with the time explanatory variable included (the resulting regression coefficient for that variable will be like an overall slope for the power generated by the solar panels over the three-year period). We will achieve (2) by producing predictions from the conditional distribution of the future kWh measurements based on the past measurements (like in Equation 2, except we will use estimates for σ^2, $\boldsymbol{\beta}$, and $\boldsymbol{R}$). Thus, we should be able to answer the overarching questions of the analysis and enable stakeholders to better plan for homeowner power needs with this model.

We also believe the proposed AR(1) model accounts for the nuances of the data set explored in the previous section. The relationship between Date and kWh is clearly not linear—the solar energy production appears to fluctuate daily and throughout the year. Therefore, a standard linear regression model would not be appropriate to analyze this data. By adding time, the proportion of the year that has passed, and the square of the proportion of year that has passed (a quadratic term) as explanatory variables to this AR(1) model (all of these are more or less functions of the original Date variable in the data set), we should be able to capture the overall degradation trend of the solar panels (if there is one) and approximate the yearly fluctuations and curvature with the two proportion covariates. The next section verifies that our proposed model is indeed suitable to analyze this solar data.

3 Model Justification

In this section, we briefly discuss why we chose which variables to include in the model, the model's assumptions, and then investigate how well it performs. We then describe in detail why we ultimately

selected this specific model over other potential models. The reasons included its ability to capture lengthened correlations, lower predicted mean squared error, and model simplicity.

3.1 Variable Justification

Recall that earlier we mentioned the explanatory variables were a time term, the proportion of the year that has passed, and a quadratic term of the year that has passed. We included the time term to capture the degradation of the solar panels as they age. From Figure 3, it can be seen that the data is periodic or seasonal. We need to capture this effect smoothly since the kWh measurements are continuous. Since this seasonal effect seems to repeat year to year, we included a Year term on a numeric scale from 0 to 1 that describes how much of the current year has passed. This was calculated by taking the number of days that have passed since January 1st of that year and dividing by the total number of days in that year. The rescaling was done for consistency since 2016 was a leap year and had an extra day. We also added a quadratic term (the proportion squared) to capture the concave-like shape that happened every year. Thus, the model we used to fit the data was:

$$\mathbf{Y} \sim \beta_0 + \beta_1(\mathbf{time}) + \beta_2(\mathbf{Year}) + \beta_3(\mathbf{Year}^2) + \boldsymbol{\epsilon}, \quad \text{where} \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{R}) \tag{6}$$

Each symbol means the same as previously mentioned, and the explanatory variables are boldfaced because they are treated as vectors. $\mathbf{Year}^2$ denotes taking the Year vector and squaring every element.

3.2 Model Assumptions Discussion

Since we have assumed a Gaussian process for the data, it is important to check if the response variable (kWh) is actually multivariate normal. That is:

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{R}) \tag{7}$$

where $\mathbf{Y}$ is a vector of kWh measurements, $\mathbf{X}$ is a matrix containing the predictor variables (time, Year, Year^2), $\sigma^2\mathbf{R}$ is the covariance matrix where $\mathbf{R}$ is the AR(1) correlation structure matrix specified earlier.

We first check for linearity between the explanatory and response variables. Since linearity is not affected by the correlations, we can simply look at added-variable plots on the AR(1) model. These plots (see Figure 5) show the marginal relationships between each predictor and the response while accounting for all other predictor variables. Due to the fairly linear relationships in the following three plots, we will conclude no significant problems with linearity.

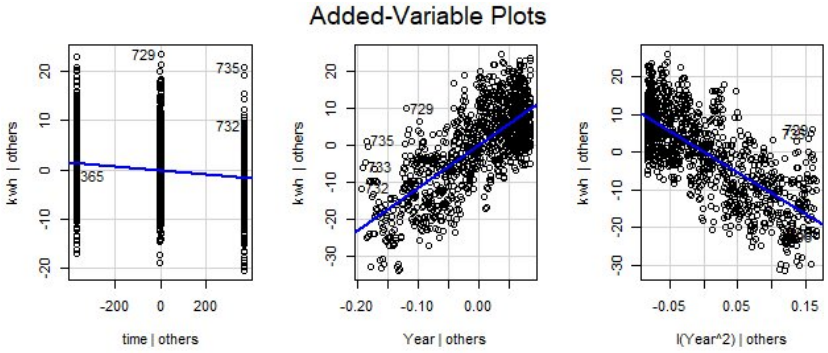

Figure 5: Added-Variable plots for each term in the AR(1) model. These all look very linear.

The next assumption that we will check is independence of residuals. However, since the model has a covariance structure associated with it, the residuals are not independent. Instead, we will decorrelate our regression model by pre-multiplying $\mathbf{Y}$ in Equation 7 by the inverse lower Cholesky decomposition,

which we will call L^{-1}, of $\mathbf{R}$. It can be shown that this multiplication will change our fitted model to a model with a simpler covariance matrix, as given in the following equation.

$$L^{-1}\mathbf{Y} \sim N(L^{-1}\mathbf{X}\beta, \sigma^2\mathbf{I}) \tag{8}$$

Since the covariance is now a diagonal matrix, the residuals should be independent. After taking out the correlation structure, there shouldn't be any dependence between the residuals left. An autocorrelation plot of the decorrelated residuals is shown in Figure 6. The plot reveals little to no correlation between the decorrelated residuals. Thus we can assume the residuals, after accounting for our defined correlation structure, are independent.

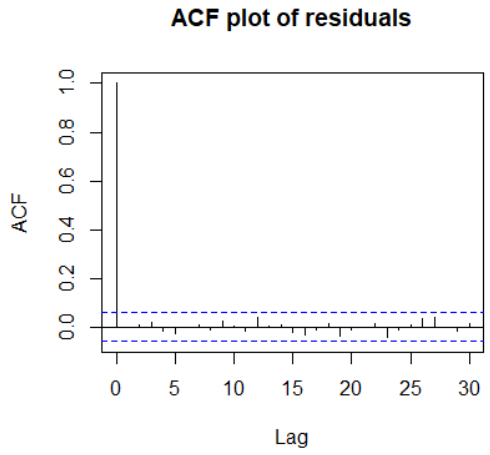

Figure 6: Autocorrelation for the decorrelated residuals

The last two assumptions, normality and homoscedasticity of residuals, are shown in Figure 7. These were checked from the decorrelated model in Equation 8, but since that model is a linear combination of Equation 7, these conditions should also hold for our raw data. The variance seems to remain consistent across different fitted values, and the standardized residuals line up nicely with a normal distribution.

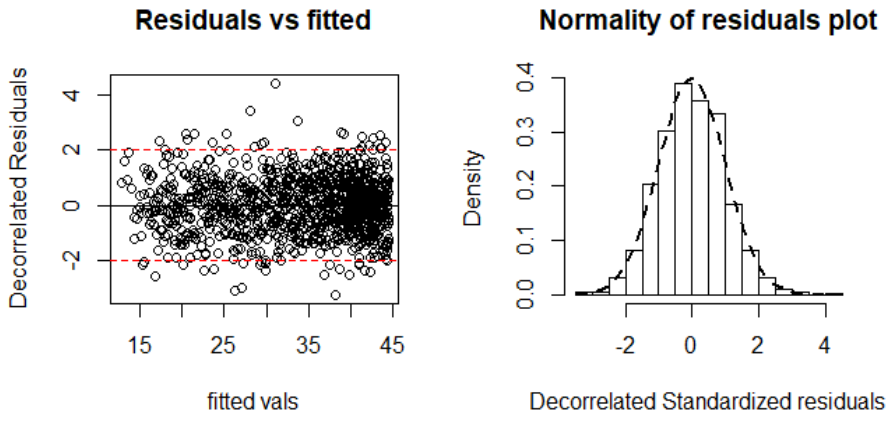

Figure 7: Plots checking for homoscedasticity and normality of residuals. The dashed lines on the left plot indicate 2 standard deviations.

3.3 Model Performance Evaluation

After demonstrating that the model is valid, we now feel comfortable moving forward and fitting the AR(1) model to our data. In order to evaluate prediction performance, we built a training model only using the first two years of the data (2015-2016). We then extrapolated for the remaining year (2017) and calculated root predicted mean squared error (RPMSE), average prediction interval width, and prediction interval coverage by comparing the predicted values to the true ones. RPMSE can be thought of as the distance our AR(1) model's predictions are off by on average. Thus, the smaller the RPMSE, the more precise our model's predictions are.

The model had a RPMSE of about 7.01 on the 2017 data, which is a measure of the standard deviation of the error around the fitted line. The average prediction interval width was 33.68, with 97.5% coverage at a 95% level. Overall, the prediction width is very high (covers about 58% of the range of the data), and the coverage is larger than what it should be. This is shown in Figure 8. From the graph, it seems that the upper prediction bound is too high, which may be because the degradation between the first two years is considerably less than the third year. Having more data might help the model produce more accurate prediction bounds.

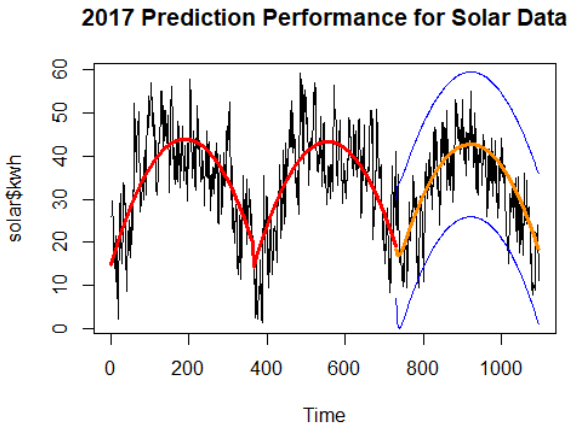

Figure 8: Model fit for 2015-2016 (red) and model predicted fit (orange) with prediction intervals (blue).

In order to evaluate model fit, we calculated an R^2 value. However, the data is correlated, so in order to preserve the interpretation of R^2, we must compute it from the decorrelated data. We fitted $L^{-1}\mathbf{Y}$ against $L^{-1}\mathbf{X}$ on the entire data set as in Equation 8. Thus, we obtained an R^2 of 0.7024, which means that after accounting for correlation, the model explains about 70% of the variability around its mean. We would conclude that this is a reasonable fit, especially after considering the amount of noise in the data. In summary, we are confident that the proposed model predicts and fits the data adequately well to achieve the goals of this analysis.

3.4 Comparison with Other Correlation Structures

In addition to the AR(1) model described above, we also explored different correlation structures to see which one fit the solar data the best. We fit models using the MA(1) and ARMA(1,1) correlation structures (see Equations 9 and 11 below for how we model the residuals in each model, respectively). As a side note, we also explored the exponential covariance structure. However, because all the time intervals are equally spaced in this data set, the exponential and the AR(1) produce the same estimates.

The formula for the moving average process (MA(1)) residuals is similar to the one for AR(1), but it has a few key differences.

$$\epsilon_t = \theta\omega_{t-1} + \omega_t, \omega_t \overset{iid}{\sim} N(0, \sigma^2) \tag{9}$$

The residual at time t is a function of the normal random component of the residual from the previous time, ω_{t-1}, which is weighted by the correlation, θ, and the normal random component for current time, ω_t. A residual in the AR(1) model, though, is a function of the previous residual, not the previous residual's random component.

The correlation function in the MA(1) model also has a different tail behavior than the AR(1) model does because it dies off after 1 lag. See Equation 10 below.

$$Corr(\epsilon_{t_1}, \epsilon_{t_2}) = \begin{cases} \frac{\theta}{1+\theta^2} & \text{if } |t_1 - t_2| = 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

We found that this model had a RPMSE of 7.04, an average prediction interval width of 27.84, and an estimated 95% prediction coverage of 94.2%. These statistics are admittedly similar to those of the AR(1) model. Nevertheless, we needed to rank these models by some metric, and since the AR(1) model had slightly better estimated RPMSE than the MA(1) model, we chose it as the better model.

Besides these metrics, though, AR(1) seemed to fit the raw data correlation better that we found with the autocorrelation function plot in Figure 4. Correlation between kWh observations undoubtedly persists for longer than one day (which is the assumption of the MA(1) model), so for that reason alone we believe that AR(1) should be favored over MA(1) when analyzing this data.

We also looked at the ARMA(1,1) process. It is a combination of the AR(1) and MA(1) processes, as illustrated in Equation 11.

$$\epsilon_t = \phi\epsilon_{t-1} + \theta\omega_{t-1} + \omega_t, \omega_t \overset{iid}{\sim} N(0, \sigma^2) \tag{11}$$

This model had a RPMSE of 7.02, an average prediction interval width of 33.63, and an estimated 95% prediction coverage of 97.5%. Therefore, the ARMA(1,1) model performed similarly to the AR(1) model but had slightly worse RPMSE. Therefore, we elected to use the AR(1) model for our analysis.

In addition to the above comparisons, we also compared these three types of correlation structures under a different explanatory variable set. In the model described in Section 2, we utilize time, the proportion of the year that has passed (Year), and the square of the proportion of the year that has passed (Year^2) as explanatory variables for kWh. In our competing design matrix, we included time, added 12 month factor variables, and removed the proportion covariates. Across all three correlation structures, the corresponding RPMSE errors were larger (7.42 for AR(1), 7.59 for MA(1), and 7.44 for ARMA(1,1)), the corresponding prediction interval widths were narrower (30.48 for AR(1), 25.13 for MA(1), and 30.39 for ARMA(1,1)), and the corresponding coverages were smaller (95.9% for AR(1), 89.6% for MA(1), and 95.9% for ARMA(1,1)). Again, we prioritized RPMSE when considering all the metrics together presented no clear model winner. Because the AR(1) quadratic term model had the smallest RPMSE out of any of these, we deemed it as the best model for this analysis because the stakeholders value prediction capacity so highly.

Therefore, we will proceed forward and report the results of our AR(1) model with the time, proportion of the year that has passed, and the square of the proportion of the year that has passed explanatory variables in the next section.

4 Results

Before presenting our results, we think it is worthwhile to mention how we found our parameter estimates in our fitted model. For example, the estimates for $\boldsymbol{\beta}$ are obtained through the generalized least squares (GLS) method, given in Equation 12 below.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \implies \min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{12}$$

From this equation, the process for obtaining $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals (i.e., $(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$) is clearly the same for standard multiple linear regression and Gaussian process modeling. The only difference is an added term in the GLS formula, $\boldsymbol{\Sigma}^{-1}$, which is the general notation for the inverse of the variance/covariance matrix of the $\boldsymbol{Y}$.

Finding the $\hat{\boldsymbol{\beta}}$ that minimizes the generalized least squares is equivalent to finding the $\hat{\boldsymbol{\beta}}$ that maximizes the likelihood of the data. Therefore, some straightforward calculus yields the following estimated regression coefficients

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'}\hat{\boldsymbol{R}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{R}}^{-1}\boldsymbol{Y}, \tag{13}$$

where $\hat{\boldsymbol{R}}$ is the estimated correlation matrix of some kilowatt hour measures, $\boldsymbol{Y}$.

The ϕ lag 1 correlation parameter in Equation 5 is estimated through an iterative maximum likelihood procedure. By replacing each ϕ in $\boldsymbol{R}$ with $\hat{\phi}$, we obtain the estimated correlation matrix for our fitted AR(1) model, $\hat{\boldsymbol{R}}$.

Lastly, σ^2, in Equation 5 is estimated analytically through maximum likelihood like $\boldsymbol{\beta}$ (see Equation 14).

$$\hat{\sigma}^2 = \frac{1}{N}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{R}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{14}$$

Thus, the parameter estimates for our model in Equation 6 are:

	β_0 (Intercept)	β_1 (time)	β_2 (Year)	β_3 (Year^2)	ϕ	σ^2
Estimate	15.6076	-0.0042	113.8876	-109.6147	0.7961	58.0687
2.5%	11.4236	-0.0086	95.7428	-127.2525	0.7573	48.6686
97.5%	19.7914	0.0001	132.0322	-91.9769	0.8292	69.2843

Table 1: Model parameter estimates and 95% confidence intervals

To address the first research question, we simply have to interpret the β_1 coefficient in Table 1. Holding all other variables constant, after 1 day, we would expect a change of about -0.0042 kWh with a 95% confidence interval of [-0.0086, 0.0001]. Zero is included in the interval, so it isn't statistically significant; however, with only 3 years of data, we may not have enough power to detect a slight but steady degradation. Looking at a yearly degregation rate, we would expect on average a change of -1.54 kWh/year (with 95% CI of [-3.145, 0.058]) after accounting for seasonal changes. In the model, the highest expected output occurs on July 6th. The expected outputs for various future years on this date are in Table 2, though the change from year to year is linear. If this trend continues, by 2043, the peak output will be 0 kWh.

Year	2018	2019	2020	2022	2025	2027	2030	2035
kwh	38.204	36.654	35.109	32.019	27.380	24.286	19.651	11.951

Table 2: Predicted peak KWH output (July 6th)

The other β estimates aren't as interpretable since Year is a value between 0 and 1 and it has a quadratic term, which means the slope is not constant. However, we think it is important to note that both of these terms have a significant relationship with kWh production because neither of their 95% confidence intervals include 0. Thus, we can conclude that the proportion of the year that has passed has a positive relationship with kWh production and that proportion squared has a negative relationship with kWh production at the 0.05 level. This second result makes sense because the kWh measurements seem to arc downwards over the course of the year, as noted in Figure 2. The value for ϕ, 0.7961, represents the correlation between residuals of 1 consecutive day, with 0.7961^n representing the correlation between days that are n spaces apart. 58.067 represents the variance or noise around the estimated line, which means the errors have a standard deviation of about 7.62.

Because we assume that the observed and future kWh observations are jointly distributed as normal, then obtaining prediction intervals for kWh output for 2018 is relatively easy. Thus, to answer the second research question, we can simply find the 0.025 and 0.975 quantiles of the conditional distribution of $\boldsymbol{Y}^*|\boldsymbol{Y}$. The results are shown in Figure 9. If we were to add the predicted values for each day in 2018 (i.e., the values along the orange line), we would estimate a total of 11,156.32 kWh (with a 95% confidence interval of [10,344.99, 12,030.36]) for the whole year.

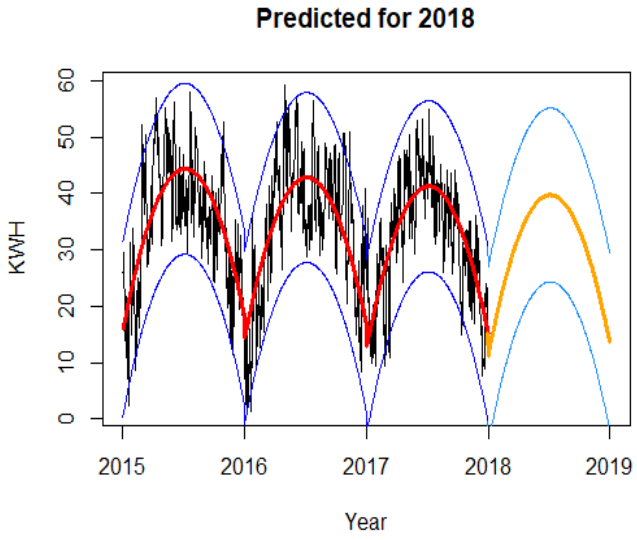

Figure 9: Model fit for 2015-2017 (red) and model predicted fit (orange) with prediction intervals (blue). Note the linear downward trend.

In summary, we found that the proportion of the year passed and the square of that proportion have a statistically significant relationship with kWh generated at the 0.05 level. We also found that the solar panel degradation rate (i.e., the time variable in our model), was not statistically significant at the 0.05 level. However, with more data (or a more generous significance threshold), we would likely see a significant result because our intuition says that there is an actual solar panel degradation process. In addition, we were able to use our AR(1) model to predict the kWh output for the following year (11,156.32 kWh with a 95% CI of [10,344.99, 12,030.36]). Thus, the power company should be able to use our results to more accurately forecast this particular homeowner's supplemental power needs for 2018.

5 Conclusion

In conclusion, we were able to successfully answer the questions about the degradation rate of the solar panels and the predicted power output for 2018. We accomplished this through the AR(1) model, which had the ideal correlation function behavior and lower mean squared error. By fitting time and a quadratic seasonal effect, we were able to generate a mostly smooth curve through the data. By interpreting the estimated coefficients in the model, the solar panels degrade linearly about 1.54 kWh/year, so by 2030, the peak output will be about half of what it started with. This effect wasn't quite statistically significant because zero is contained in the interval (-0.008, 0.0001), but with more years of data, we would likely have enough power to detect significance. We also answered the second research question by extrapolation, and we estimated a total output for 2018 of 11,156.3 kWh (95% CI: [10345.0, 12030.4]).

One major limitation in the model is that it doesn't account for environmental changes. Solar power generation can be very sensitive to the weather and if there is an abnormally cloudy year, the model won't pick up on those changes. This model also only works for solar panels in one geographical location.

The kWh output is likely very different in any other city. Although our AR(1) model is mostly smooth, it is discontinuous at each year mark, which may not be realistic. This can be resolved through splines, but that solution drastically decreases model interpretability. With only 3 years of data, we also assumed the solar panels decay at a linear rate, which may not be the case as the panels age.

We think a valuable next step in this analysis would be to gather data from more than one solar panel system (preferably in other geographical locations), and for longer periods of time than three years. It might also be helpful to study the effects that different climates have on solar output. By doing so, we believe our results would be more generalized. It would also give us more information about how the panels degrade over time, enabling us to make more accurate predictions for the future.

6 Team Work

This report was a collaborative effort. As far as the coding, Dean worked on the MA(1) and ARMA(1,1) models, and Jeremy worked on AR(1) and exponential ones. With regards to the report writing, Dean was primarily responsible for the Introduction and Model sections, while Jeremy was primarily responsible for the Model Justification, Results, and Conclusion sections. Dean also contributed to model comparison subsection of the Model Justification section. We both learned a lot from each other and had a fun time!