**Data 670 Data Analytics**

James Taylor

**Professor Hany Saleeb**

**Assignment 2**

**9/14/2021**

**Executive Summary**

Solar generated electricity has become a viable source to meet power demands in recent years with better technology, investment and society’s increased interest in renewable energy. Unlike other contributors of electrical generation (coal, natural gas, nuclear) in the US (1), solar has little influence over supply of the electrical-generating source (sunlight in this case). The uncertainty caused by solar’s variable electrical output can produce inefficiencies in the electrical grid, both concerning resources and pollutants.

      Building a model that accurately predicts a solar plant’s output in advance would help limit these potential inefficiencies by limiting excess power generation. It will help power grid administrators plan for how much solar energy could be created, whether it be available from the solar facility or an electrical storage facility that stores solar-generated power for use at peak times. Knowing this optimizes how supplemental power sources are used and would likely reduce pollution since those sources are mostly non-renewable. A model would be useful for planning when those supplemental power sources are needed to deliver power to the grid. Plants sometimes take hours to prepare themselves for electrical generation and inefficient timing means either unnecessary run-times or inability to meet the grid’s power demands.

      A variety of models will be made to help power grid administrators predict how much electricity a solar plant will produce. These models (or modeling techniques) could be used at other power grids to plan and optimize their operations.

# Project Scope

Problem Description

The problem is solar is an inconsistent source of electricity. Its availability is affected by the climatic conditions and time of day. That means solar’s ability to provide electricity to the grid is inconsistent. To meet the demand for electricity, grid administrators need to plan and have available sources of electricity to provide for the grid. A model predicting solar’s electrical production depending on climate factors will help grid administrators plan to meet electrical demands and maximize efficiency.

         The problem is important for two reasons, optimization and pollution. A grid administrator’s ability to know in advance what solar’s input would be allows them to plan what sources/plants need to be producing electricity (coal, nuclear, hydropower, etc.). However, these plants can’t be turned on/off instantly and mostly take hours to prepare to effectively produce energy. That means administrators have to predict when to have a plant start its “warm-up” procedures so that it can be producing electricity at a time when the administrators need it.

Because solar produces an inconsistent amount of power, administrators have to be conservative when using solar and have other supplemental electricity sources available in case solar fluctuates and doesn’t meet the electricity demand of the grid. In other words, administrators can’t rely on solar like they can natural gas or coal, so they have to have other sources readily available to funnel into the grid if solar’s electrical production drops. Having to keep those other resources available is inefficient and costly. It also produces unnecessary pollution because supplemental electricity sources skew nonrenewable.

The location of where electricity is generated is crucial as well. Electricity is not free to get from point A to B. There’s all sorts of costs like building and maintaining the lines, and places have varying ways of deciding who gets the bill, but another cost is electricity loss during transmission. For example, there are very productive solar plants in Nevada but that is far from Los Angeles. Every mile of distance is more electricity sent that will not get to its destination.

Business Understanding

         In the US, how energy is bought and sold to be delivered to end-users becomes complicated because it depends on region, state and even city level differences. For most of the US there are three systems on how energy is bought/sold to then be delivered to end-users, which is regulated vertical monopolies, deregulated customer choice, and deregulated wholesale market.

         In every market system, someone is responsible for guaranteeing electrical supply to customers. In the traditional regulated markets, it’s the natural monopoly utility. In the deregulated markets the utility provider has met their customer’s demand by purchasing power from plants. The inefficient costs are in having plants prepared to produce electricity that was never needed. There are also pollution costs since those plants skew nonrenewable.

Organization

There are ten Regional Transmission Organizations (RTO) that are independent and not-for-profit that manage the bids in wholesale electricity markets, which make up about 60% of US electrical supply (2). They are the ‘middle-man’ between electricity generators and the rest of the grid. They are responsible for ensuring the grid has enough power plus the reserve margin. They will need to know how much power solar plants can produce, or the solar plant can tell them the RTO that figure.

Stakeholders

There are a few groups that would be interested in predicting solar’s future output. RTO who manage the grid want to know so they can plan which plants need to be running and prepare to run to ensure they meet the grid's electricity demand. Customer’s and retail utilities will be interested as well because accurate predictions will make solar a more reliable electricity source, meaning the more expensive supplementary electricity sources won’t have to be relied on as heavily, potentially resulting in lower costs. Older plants that take longer to fire up may have prolonged lives as planning when they need to start their warm-up process can be made easier with a better idea of the electrical supply’s future capacity.

Define Business Area

The business area is the energy utilities sector. More specifically it will make solar a more effective source, which currently contributes 1-2% of the total US electricity depending on how it’s measured. Electricity in the US is heavily regulated by the government but for-profit companies still operate much of it. Either way a model could return cost savings.

Business Objectives

The model will help reduce costs and pollution by reducing uncertainty for solar electrical generation. This is good for consumers as this likely means cheaper energy and less pollution to generate it. Solar plants would become a more reliable energy source, minimizing the need for supplemental electricity sources.

         Having a model that would predict solar’s output for the next day, or the next hour would remove some variability. This could lead to more trust in solar to power the grid, potentially spurring more investment in solar.

Business Success Criteria

         A successful model for the business application is an accurate model. There are two probable ways a model can help. One is to make it’s best prediction given the independent variables. This would give the user’s the most likely output. The user, like an RTO, would then have to decide how much to trust this number because the model likely didn’t make a perfect prediction and the actual will be higher or lower than it.

         Another successful model could be one that can give the user a degree of likelihood the solar plant can meet this threshold. For example, a model like this can output the largest kilowatt threshold that it believes will be achieved with 95% confidence. This would give the user an idea of the highest electrical generation level that is very likely to occur.

Background Research

         There is a decent amount of research in the solar output prediction area. It has real monetary value and it’s also an interesting exercise. There are websites (<https://www.renewables.ninja/)> where you can get predictions on a solar system anywhere in the world. One research paper was quite accurate in predicting when solar output would drop by having a camera pointed at the sky, taking in information on clouds (3). There are numerous papers on model solar output with varying techniques.

         The dataset used was publicly available, so there is a significant amount of open source analysis and modeling on it. I have not looked at much of the models done by others, but it appears many different techniques were used.

Gaps in this Problem Resolution

         There are many data sources for aggregate electricity. These would be electrical usage or production across countries or regions. There are less sources of data at the solar plant level, where an individual plant publishes their electrical generation. This may be because individual plants are more likely to be private entities, meaning they probably don’t have to release certain information for reasons like competitive advantage and company privacy.

         Solar plants likely have some predictive models for their equipment. These simpler models don’t have interchangeability between solar plants (a model trained at one plant won’t be good at predicting at another plant). Number of solar panels, their sizes, quality, orientation, and location vary between solar plants.

         There is a decent amount of research on optimizing electrical generation from solar. Not all this information will be able to be utilized however, because there is limited information in our data, like the different types of radiation (direct and diffuse).

Proposed Project

         I’m doing this project for the potential benefit of solar generated electricity being utilized efficiently and I enjoy making future predictions. I’ll use historical data to train a model meant to predict our data’s solar plant’s electrical generation, mostly weather data.

Key Performance Indicators

         The performance indicators will be about the accuracy of the model. The first would be the RSS, or Residual Sum of Squares. The lower the standardized residuals sum the better our model is fitting our data. This technique would be used if the model is producing a point estimate as opposed to a binned prediction.

         If binned predictions are chosen, the accuracy metrics would become focused on accuracy (percent predicted correctly). These are slightly different because it is a classification problem.

         If a model has a confidence level, say a model is 95% confident the generation will be around this point, we will be able to look at the test data, compare it to the actual and see if it’s confidence levels were right. For example, if the actuals from the test are 95% confident the actual will be within +/- 10 kW, but all of the actuals are outside that band then we are misrepresenting how accurate our model is.

Project Insights of your Data Analysis

      I expect radiation and temperature to be highly correlated with electrical output. These are synonymous with direct sunlight which bodes well for solar panels. Sunlight can be blocked by clouds, which decreases both temperatures and irradiation.  The higher intensity on the panels the more electricity will be generated.

         Another likely discovery is that the most productive times will be during mid-day. That is when sunlight coming from its highest point and less light is reflected off the atmosphere, delivering the most intense light to the ground. Obviously at night there isn’t much electricity generated.

         The previous timestamp’s electrical generation may correlate well with the next one’s. The weather and sunlight conditions are likely similar since little time has passed to allow the conditions to change much. A successful model may take the previous timestamps output to help it make a prediction.

Data Set Description

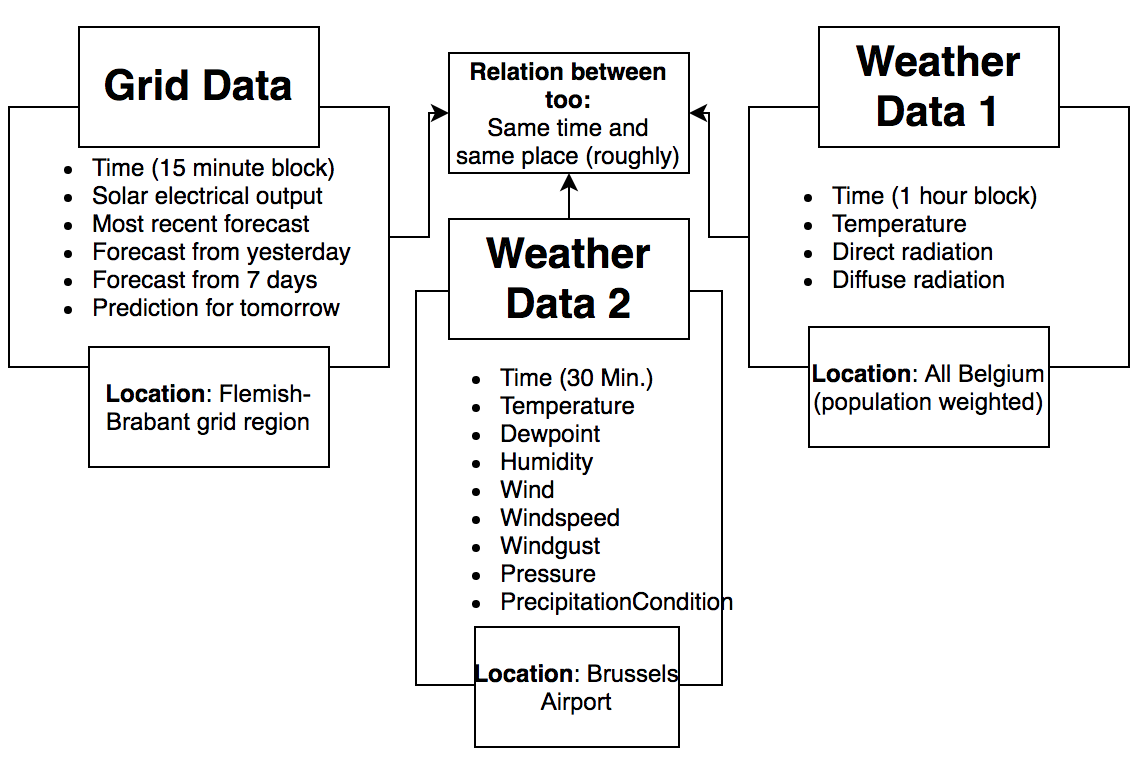
         There are three datasets being used in this dataset. The first is the data from Elia Group, which is an electrical transmission company that the FEBEG (Belgian Federation of Electricity and Gas Enterprises) started an initiative to make more data public. One of those publications is solar power electrical generation by region. The data used was from the Flemish-Brabant region, which surrounds the city of Brussels. It has megawatts produced from solar power in that region every 15 minutes, resulting in about 35,000 observations. It has a timestamp, the solar generation in MW (Megawatts), the most recent prediction for this time, the MW prediction for this time from the day before, the MW prediction for this time from seven days before and the prediction for this time but tomorrow's electrical output. All are in MW’s.

         Another dataset is sourced from Open Power Systems, which gets their data from renewables.ninja which used the NASA MERRA-2 reanalysis. It produces temperature, and both direct and diffuse radiation for every hour in 2019, which is 8,760 observations. These values are country-aggregated, so the values represent the whole of Belgium, but the data is population-weighted. That is beneficial because of the top five populous cities, the top two are from the region we have solar data for, and the fifth one is the region that is surrounded by our region (Brussels has its own region within Flemish-Brabant region).

         The third dataset is scrapped data from the Weather Underground website. It is from the Brussels airport weather station. There was an option to pay for the historic data, but instead I wrote a python script that looped through the dates of 2019 and pulled the data from a table in the HTML. This data was per-30 minutes.

         The datasets will have to be combined in a creative way since one is per-hour, one is per- half hour, and the other is per-15 minutes.

High-Level Data Diagram



The above diagram shows a basic view of the datasets and how they can be related to each other. They are at the same time and the same region (generally).  They do have separate timescales, but creative solutions can overcome this.

Data Definition/Data Profile

         The weather dataset has complicated underpinnings from NASA, but the outputs are simple enough. One important distinction is that ‘radiation’ is probably not the right word for the measurements. Instead, irradiance would be a better word to use. That is the measure of light-energy that is received on a square meter flat surface. Photons carry energy from their source and deliver it to whatever they hit. This is a very important measure for solar generation.

         Another important topic is the direct and diffuse irradiance. It's best to demonstrate the difference. During sunny days, your person projects a shadow on the other side of the light source. Those photons would’ve hit the ground but your body was in the way. That means no light is coming directly from the light source to that spot, which is the direct irradiance. Direct irradiance is energy directly from the source. However, if no light is coming from the light source and to the ground (because the body is in the way), shouldn’t the shadow be completely black? The reason it isn’t and you can still see things in a shadow is because of diffuse light waves.  Those are photons that bounce around our atmosphere and come from random directions. Those are the photons that land inside your shadow, preventing it from being pitch-black.

         The balance between direct/diffuse is telling. On clear blue sky’s at noon the irradiance is overwhelmingly direct. There is still some diffuse, but the atmosphere is minimally redirecting waves. On very cloudy days the irradiance can essentially be 100% since clouds are blocking all light directly from the sun.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Temperature | Integer | variable for BE in degrees C |
| Radiation Direct Horizontal | Float | weather variable for BE in W/m2 |
| Radiation Diffuse Horizontal | Float | weather variable for BE in W/m2 |

         The solar electrical generation dataset has more variables on it. Clearly the grid administrators use some predictive model to attempt to guess at the potential output. More analysis will be done to see their accuracy, and potentially use it as a baseline to compare to ours. The main variable we’re interested in is Corrected Upscale Measurement. This is the Megawatts generated by the solar facilities in the Flemish-Brabant region. This is data from the grid operators, not from the solar plants. The plants certainly have data on weather and their electrical generation but it’s likely not available for previously stated reasons. It is still pertinent to the goal of this paper—helping grid administrators predict solar output to better optimize meeting the grids demand.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Most Recent Prediction | Float | Most recent prediction of MW generated for this time block |
| Day Ahead Prediction | Float | Yesterday’s prediction of MW generated for this time block |
| Week Ahead Prediction | Float | 7 days ago prediction of MW generated for this time block |
| Corrected Upscale Measurement | Float | Megawatts of energy generated during that time block from the solar panels |
| Day Ahead Prediction (11 Hours) | Float | 11 hours prior’s prediction of MW generated for this time block |

         The Weather Underground scrapped data has the following variables. It is historical data from the Brussels airport where climatic conditions were collected every 30 minutes. This data had to be scrapped from HTML tables using Python. There were 17,498 observations. About 20 data points were missing.

|  |  |  |
| --- | --- | --- |
| Variable | Format | Description |
| Time | String | Date and Time |
| Temperature | Integer | Temperature in F |
| Dew point | Integer | Dew point |
| Humidity | Decimal | In percentage |
| Wind | Integer | Wind in mph |
| Wind Speed | Integer | Speed of wind in mph |
| Wind Direction | Categorical | Compass direction of wind |
| Wind Gust | Integer | Strongest wind gust |
| Pressure | Float | Pressure in inhg |
| Precipitation | Integer | Precipitation in inches |
| Condition | Categorical | Overall weather condition |