Presumed Open Data

Data Science Challenge Description

Dr Stephen Haben

Data Scientist

Thursday 28th January 2021

Contents

1. Background

Distribution networks are expected to be under increased strain due to the growing uptake of low carbon technologies. Energy management systems, which control and shift power on the network will be increasingly common for handling headroom complications on the network. Further to this, it is likely that distributed renewable generation (solar and wind) may need to be utilised in areas local to where it is generated to help alleviate the effects further up the network. This also increases the opportunity to utilise more renewable energy resources and help the UK work towards its net zero carbon targets.

Unfortunately, energy generation from solar photovoltaics (PV) often occurs at periods when the demand is relatively low, whereas the peaks occur in the evening when there is little-to-no solar generation. In addition, if many local primary substations also have high penetrations of PV this can cause issues further up the distribution network.

This challenge focuses on the scenario of demand peak reduction for a primary distribution substation which is located near to a Solar PV farm. The aim is to optimally control a battery storage device connected to the primary substation and the solar farm to both reduce the evening peak period by utilising as much solar generation stored from earlier in the day. This type of control can help reduced strain on the network by reducing PV sent back up the network and reduce peak demand.

In other words, the aim of this challenge is to design an optimal schedule for a battery storage device by 'shifting' the solar energy generated during the daytime, to the evening period by charging during periods of high solar output and discharging in the evening during the high demand period.

The creation of an appropriate charging/discharging profile for the storage device is far from trivial, in particular because of the following:

- Demand, especially at the distribution level, is relatively volatile.
- Solar generation is also volatile and intermittent, and highly dependent on the stochastic coverage and level of cloud.
- The storage schedule must often be devised at least a day in advance. This increases the uncertainty of inputs to a scheduling algorithm.

The aim of this challenge is therefore to utilise available data, including data collected at the demand and generation sites, to optimise the control actions of the storage device. A detailed description of the data and the challenge will be given in the following sections.

2. The Data

There are several data sets which are to be utilised in this challenge.

2.1. Demand Data

The core data set is the demand data which comes from the Stentaway Primary substation near Plymouth, on the south coast of the UK. The approximate (latitude, longitude) co-ordinates are (50.364,-4.086). The data consists of half-hourly average Power values in Megawatts (MW) starting on the 3rd November 2017.

The data is relatively clean but there are some anomalous values, such as unusually large and small values from shortages and periods of constant demand, but these are relatively few. Entrants may wish to clean the data further to improve their final submission.

The demand data for the challenge period will not be available, and so a major part of the challenge is to create a discharge schedule for the data without knowing the future demand. It will have to be estimated based on historical data and other data inputs.

The time stamp is in UTC/GMT and the demand is given as average values (in MW) over the next 30 minutes from the time stamp. I.e. the demand at time stamp 15:00 will refer to the average power over the period 15:00 to 15:30.

2.2. Solar PV Generation Data

The solar PV generation data is from a 5MW solar Farm in Devon, UK. The site is located at (longitude, latitude) = (50.33,-4.034) and is not too far from Stentaway substation.

The data is in 30minute resolution with timestamps in UTC and also starts on the 3rd November 2017 (same as the demand data). The following variables are part of this data set:

- PV generation (in MW) is an average value over the next 30minute period. e.g. 11:00AM time stamps indicates an average value over 11:00 to 11:30 period.
- Solar irradiance data (in W m^{-2}) and are also given as an average value over the next 30minute period from the given time stamp (e.g. as with the PV generation values 11:00AM time stamps indicates an average value over 11:00 to 11:30 period).
- Finally, there is also a temperature value which is the instantaneous measurement, in degrees C, of the PV module temperature. Note this will often be different to the ambient temperatures.

The solar PV data will not be available for the challenge period this means the appropriate charging of the PV will be dependent on using the weather data, historical values of the generation, and the other available data.

There are some gaps in the data but not too many. It is left to the participants with how they may clean the data.

2.3. Reanalysis Weather Data

Weather is an important component of energy systems. However, selection or engineering features from multiple sites could be beneficial. For this reason, temperature and irradiance data has been extracted from several sites using the MERRA-2 reanalysis data[1].

[1] Extracted based on the following code: https://github.com/emilylaiken/merradownload

Reanalysis weather data are estimates of weather variables at the numerical weather prediction grid points based on assimilation of historical weather data. In this challenge hourly irradiance (in Watt per square meter, or Wm^{-2}) and surface temperature (in Celsius) data from 1st January 2015 have been extracted from the following sites[2]:

- Location 1: latitude 50.5, longitude: -4.375
- Location 2: latitude 50.5, longitude: -3.75
- Location 3: latitude 51, longitude: -3.75
- Location 4: latitude 51.5, longitude: -2.5
- Location 5: latitude 50, longitude: -4.375
- Location 6: latitude 50, longitude: -3.75

The locations correspond to grid points on the numerical weather prediction grid. Locations 1, 2, 5 and 6 surround the substation location but location 5 and 6 are within the English Channel.

The irradiance data is time-averaged over the next hour period (e.g. 10:00 AM time stamps means the period 10:00 to 11:00 AM), but the temperature is an instantaneous value at each hour of the day. As with the other data sets the time stamps refer to UTC. Unlike the demand and PV data, the reanalysis data will be treated as a weather forecast and will be available throughout the challenge periods.

[2] There are many different versions of temperature and irradiance values which MERRA-2 makes available. In this challenge the temperature field is 'T2M' from the 'inst1_2d_asm_Nx' collection or irradiance the field is SWGDN field in the collection "tavg1_2d_rad_Nx". The description of these fields can be found here: https://gmao.gsfc.nasa.gov/pubs/docs/Bosilovich785.pdf

3. Detailed Description and Problem Formulation

3.1. Mathematical Description

The aim of this challenge is to maximise the percentage evening peak reduction for the demand on a primary distribution feeder for each day over a week period, whilst utilising as much solar PV generation as possible to do so.

This is done by finding an appropriate charging profile for the storage device so that it charges at the correct rate during daytime periods when there is high solar generation, and discharges at the correct rate during the evening period, defined as the period from 3.30PM to 9PM (UTC time).

More precisely suppose the actual future demand for the day d of the test week ($d = 1, \dots, 7$) is given by

$$\boldsymbol{L}_d = (L_{d,1}, L_{d,2}, \dots, L_{d,48})^T$$

Where $L_{d,k}$ is the average power (in MW) over the k^{th} half hour of day d, of the week so $(d,k) = (1,1)$ represents the period from midnight to 00:30AM on the first day, $(d,k) = (7,2)$ is the period from 00:30AM to 01:00 AM on the seventh day etc. The aim is to determine the best average charge or discharge (in MW) for the storage device over each day of the week, given by

$$\boldsymbol{B}_d = (B_{d,1}, B_{d,2}, \dots, B_{d,48})^T$$

(where $B_{d,k}$ is the average power (in MW) over the k^{th} half hour of day d) to minimise the peak demand over the evening period (the half hours $k = 32\ to\ 42$) given by

$$\min_{\boldsymbol{B}_d \in \mathcal{B}_d} \left\{ \max_{k \in \{32,\dots,42\}} (L_{d,k} + B_{d,k}) \right\}$$

$\mathcal{B}_d$ is the set of all feasible profiles for the battery storage device and are defined by various constraints. First, the battery storage device is limited by the maximum import and minimum export rate of charging and discharging respectively,

$$B_{min} \leq B_{d,k} \leq B_{max}$$

Where in this example the maximum charge rate is $B_{max} = 2.5$MW and the maximum discharge rate is $B_{min} = -2.5$MW. Secondly, the battery cannot charge beyond its capacity. Let $C_{d,k}$ represent the total charge (in MWh) in the battery on day d and half hour k, and so

$$0 \leq C_{d,k} \leq C_{max}$$

Where the maximum capacity for this challenge is $C_{max} = 6$MWh. The change in the total charge in the battery from one step to the next is related by

$$C_{d,k+1} = C_{d,k} + 0.5 B_{d,k}$$

In other words, the total charge is changed by the average rate of charging. The 0.5 is to convert power (MW) into energy (MWh). Notice there is assumed to be no losses from charging/discharging the battery. Finally, to simplify the calculation, the battery is only allowed to discharge during the evening period, i.e.

$$B_{d,k} \leq 0 \qquad \text{For } k = 32, \dots, 42$$

And can only charge prior to this

$$B_{d,k} \geq 0 \qquad \text{For } k = 1, \dots, 31$$

The final constraint is that the battery must start with zero charge on the first half hour of each day of the week. I.e. $C_{d,1} = 0$ for $d = 1, \dots, 7$.

The other aim of the storage device is to maximise the amount of generation from PV solar sources. In other words, when $B_{d,k} \geq 0$, (i.e. when importing) then the charge can be written

$$B_{d,k} = P_{d,k} + G_{d,k}$$

Where $P_{d,k}$ is (average) power stored in the battery from the solar generation and $G_{d,k}$ is energy stored from the grid at half hour k. Let $P_{d,k}^{Total}$ be the total energy generated by the solar PV (in average MW) over half hour k over day d. Whenever the battery is charged, if energy is generated from the solar photovoltaics (PV) then this is used to charge the battery first. If more charge is requested than what is available from the solar PV then the remainder is taken from the grid. In other words, there are the following two scenarios when charging the battery:

1. If $B_{d,k} \geq 0$ and $B_{d,k} \leq P_{d,k}^{Total}$, then at time step k, the battery is charged with energy solely from the solar PV, i.e. $P_{d,k} = B_{d,k}$, and $G_{d,k} = 0$.
2. If $B_{d,k} \geq 0$ and $B_{d,k} \geq P_{d,k}^{Total}$ then at time step k, then the battery is charged with $B_{d,k} = P_{d,k} + G_{d,k}$, where $P_{d,k} = P_{d,k}^{Total}$ and $G_{d,k} = B_{d,k} - P_{d,k}^{Total}$. In other words, the extra energy required is taken from the grid.

The final proportion of energy in the storage device from the solar PV during day d is therefore given by

$$p_{d,1} = \frac{\sum_{k=1}^{31} P_{d,k}}{\sum_{k=1}^{31} B_{d,k}}$$

Thus, the second objective is to maximise $p_{d,1}$ for each day d in the week. The two elements (percentage peak reduction and proportion of PV energy stored) will be combined to produce the final assessment score, see next section.

3.2. Scoring Function

For each day $d = 1, \dots, 7$, of the current trial period, a score will be assigned to each participant based on the following cost function:

$$S_d = R_{d,peak}(p_{d,1}C_1 + p_{d,2}C_2)$$

Where

- $R_{d,peak}$ is the peak reduction (as a percentage) on day d defined by $100\left(\frac{Peak_{old} - Peak_{new}}{Peak_{old}}\right)$
- $p_{d,1}$, $p_{d,2}$ are the proportion of energy stored in the battery from solar energy and from the grid respectively on day d. Notice that $p_{d,1} + p_{d,2} = 1$. The formula for $p_{d,1}$is given in the previous section.
- $C_1 = 3$, $C_2 = 1$ are weights for the solar and grid energy, respectively. These weights are based on the relative lifetime GHG emissions intensity of solar and electricity from the grid, see appendix for justification.

The scores will be calculated for each day of the week and the score for that task-week will be the average over the seven values, i.e.

$$S_{final} = \frac{\sum_{d=1}^{7} S_d}{7}.$$

3.3. Benchmark

A simple benchmark will also be included with the other scores to help teams assess their models. This will simply assume the demand and generation in the test week is the same as the previous week and the charging and discharging profile will be based on this. This a particularly naïve method but will perform well if the conditions are very similar from one week to the next.

4. Submission Guidelines and Rules

4.1. Registration

Only registered members are eligible to take part in the challenge.

To sign up to the challenge participants must register here: https://www.westernpower.co.uk/pod-data-science-challenge

For the first submission please cc all members of your team so we know who is part of whose team (team members need to stay fixed throughout the challenge).

The registration will stay open until one week after the kick-off (4th February 2021).

4.2. Task Weeks

There will be four assessed tasks each consisting of a week period.

- **Task 1**: For the week period from 2018-10-16 to 2018-10-22 (inclusive).
- **Task 2**: For the week period from 2019-03-10 to 2019-03-16 (inclusive).
- **Task 3**: For the week period from 2019-12-18 to 2019-12-24 (inclusive).
- **Task 4**: For the week period from 2020-07-03 to 2020-07-09 (inclusive).

There will also be a practice task, Task 0, which will be the period 2018-07-23 to 2018-07-29 (inclusive) which will take place prior to the assessed tasks and will serve to allow participants to get used to the data and the submission process.

Notice that the tasks cover a range of seasons and periods. This is to ensure that the scores reflect the average score across the variety of situations which the control would need to be applied.

The final score for the whole challenge will be the average score over all four tasks as outlined in Section 3.2.

The day after the submission of the current task, extra data will be released for the next task. Only the weather data will cover the period of the task week, all other data will be historical up to midnight of the day prior to the Task week.

4.3. Timelines

The following are the important dates for the challenge:

7th December 2020: Registration Opens.

28th January 2021: Data Science Challenge kick-off event and release of initial data set and practice challenge.

4th February 2021: Registration Ends and teams must be finalised.

11th February 2021: Submission deadline for trial/practice challenge.

12th February 2021: Start of Task 1 and release of data.

25th February 2021: Submission deadline for Task 1.

26th February 2021: Start of Task 2 and release of data.

4th March 2021: Submission deadline for Task 2.

5th March 2021: Start of Task 3 and release of data.

11th March 2021: Submission deadline for Task 3.

12th March 2021: Start of final Task 4 and release of data

18th March 2021: Submission deadline for Task 4 and for the outline report for methods.

End-March 2021: Winner(s) announced.

TBA: Presentation for challenge winners.

4.4. Submission Guidelines

The data for each of the tasks will be shared on WPD's Open Data Hub: https://www.westernpower.co.uk/innovation/pod/ on the day when the new task starts (See the **Error! Reference source not found.** Section above for dates). Participants will need to register to t he hub for access to the datasets whose folders will be labelled "pod_ds_taskX" where X is the current task number. In there will be the datasets to be used for the current challenge as well as the submission template.

The given submission template must use this for all entrants submissions. It will consist of a timestamp column and an empty charging/discharging column which must be filled in giving the team's estimate for the optimal charging/discharging schedule for the battery during the task week. Remember discharging should be given as negative values.

The name of the submission template must be adapted to be written as "teamname_setX" where teamname must be replaced with your chosen team name and K is the task number. Hence K=0 for the practice challenge, K=1 for the first assessed task etc. If this convention is not used the score will not be included in the final assessment.

For the first submission please cc all members of your team so we know who is part of whose team (team members need to stay fixed throughout the challenge).

The submission must be sent to podchallenge@es.catapult.org.uk before midnight of the submission deadline for each task (See the Section 4.3 above for dates) to qualify.

4.5. Main Rules

- In order to qualify for a prize, the entrant(s) must make a submission prior to the deadline for all four tasks. If a task is missed then the entrant is no longer eligible for the final prize, although their average scores can still be included in the leader board.
- Only one submission is allowed per team, so make sure it counts!
- The teams are responsible for their charging profiles obeying the constraints. A check will be done by the organisers when processing their submission. If any constraints are violated then that submission will be void. So please check that the battery doesn't go beyond the capacity and the charging/discharging rate is also within limits! In addition, check that charging and discharging is only within the correct time periods (see the Section 3 for more details).
- Teams must use the submission template provided, no other formats or files will be accepted.
- Only registered participants are eligible to enter and all members of a team must be registered. Team members are to notify organisers of the members of their team and team name (see sections 4.1 and 4.4 above)

- Teams sizes are limited to a maximum of five people.
- To be eligible for the journal article prize the entrant(s) must first submit a short outline (2-3 pages) of their method and approach with the deadline for the final challenge.

4.6. Prizes

There are two different prizes available. Up to two of the highest scoring academic teams will be offered a free option to publish their method and results in the "Forecasting and Management Systems for Smart Grid Applications" Special Issue for the Energies journal[3]. If there are winning entrants from commercial companies wishing to publish then this is negotiable.

Other top entrants are invited to pitch their work and solutions to senior representatives from Western Power Distribution, Centre for Sustainable Energy and Energy Systems Catapult as well as other major stakeholders during a special workshop for the event.

The prizes are subject to the Rules outlined in the above section.

3 https://www.mdpi.com/journal/energies/special_issues/forecasting_management_systems_smart_grid_applications

5. Questions and Answers

Q. Rather than charging and discharging the battery, can I also redirect energy from the solar to the network, especially when the peak period overlaps with when solar is generating?

A. Although power electronics could divert energy to the grid, for the sake of simplifying the challenge the energy can only be transferred to the grid by charging and discharging the storage device at specific periods of the day.

Q. Can I charge and discharge the storage device at the same time?

A. No, the devise can only be in charging model (positive average power) or discharging (negative). Hence there is only one column for the submitted charging profile for participants in the submission template.

Q. Can we utilise additional external data sources.

A. No, only data provided by the organisers can be utilised to keep the challenge fair and self-contained.

Q. Is the data already cleaned?

A. The data has only been pre-processed slightly. There is still some gaps and erroneous data which the teams will have to decide what to deal with. The aim is to have the challenge include some of the typical problems which are present when working with real data. However, a dataset has been chosen where there is a limited number of errors so entrants can focus on the challenge. Some details on the data are included in their corresponding heading in the data section 2 above.

Q. Do I have to enter as an individual or can I join as a team?

A. Teams of up to five members can join. All members must be registered before 4th February.

Q. Can teams confer?

A. No.

Q. Can I submit multiple entries.

A. No, the first submission is taken as the only submission for each week's task although exceptional circumstances may be considered.

Q. How do you become valid for a prize?

A. The prize-winning teams are those that enter all assessed parts of the competition (Tasks 1 to 4) and also at the end submit an outline of the method and approach. This will not be shared beyond the organisers without permission from the participants. To be eligible for the publication the journal have asked that there is academic authors on the paper but this is negotiable.

Q. Where do I send my solutions?

A. Submissions should be sent to podchallenge@es.catapult.org.uk

Q. Should we use the full battery capacity, won't this degrade the battery?

A. The limits given here can be viewed as effective capacity. Battery degradation is not a focus of this challenge and hence has been ignored.

Q. Will you account for rounding errors in the constraints for the solutions submitted?

The participants must check that their solutions don't violate the constraints before they are submitted. There may be some tolerance included in the scoring functions for rounding errors, depending on the outcomes from the practice challenge.

6. Appendix: Justification for cost function weights

The weights are based on relative average lifecycle green-house-gas emissions intensity of average electricity and solar electricity[4]. This gives the Lifecycle GHG emission intensity for Solar PV is 85 tonnes C02e/GWh. We estimate the average Lifecycle GHG emissions of the UK by doing a weighted average of the Lifecycle GHG emission intensity for the other technologies where the weights are the percentage contribution to the total electricity generation in 2019[5]

Hence an average whole system lifecycle GHG emission intensity would be

$$\frac{1}{100}(888 \times 2.1 + 733 \times 0.3 + 499 \times 43.3 + 85 \times 3.9 + 45 \times 11.4 + 29 \times 17.3 + 26 \times 1.8 + 26 \times 19.9) = 256.02$$

In tonnes CO2e/GWh.

This is assuming the following breakdown of the fuel: Nuclear (17.3%), biomass (11.4%), hydro (1.8%), wind (19.9%), solar (3.9%), other (2.4%), gas (40.8%), coal (2.1%), oil (0.3%). Other was rolled into gas and since the percentages do not add to 100 then another 0.1 was also added, giving gas an updated percentage of 43.3%.

Since the Lifecycle GHG emission intensity of solar PV is 85 tonnes CO2e/GWh. This means the average grid (using 2019 breakdown) is $\frac{256}{85} \approx 3$ times the lifecycle GHG emission intensity of solar PV.

[4] http://www.world-nuclear.org/uploadedFiles/org/WNA/Publications/Working_Group_Reports/comparison_of_lifecycle.pdf

[5] https://www.carbonbrief.org/analysis-uk-low-carbon-electricity-generation-stalls-in-2019

Energy Systems Catapult supports innovators in unleashing opportunities from the transition to a clean, intelligent energy system.

Energy Systems Catapult
7th Floor, Cannon House
18 Priory Queensway
Birmingham
B4 6BS

es.catapult.org.uk
info@es.catapult.org.uk
+44 (0)121 203 3700