

Data preparation

There were two cvs data files used in this project. The first one contained parameters from the PV – installation from six locations and the second one contained weather data from five locations in UK. From both files only Forest Road data were used.

PV data

PV data for Forest Road location called dataFR, primary contained 7709 rows and 63 columns. This data were collected from 01.10.2013 to 03.10.2014 with 1 hour frequency. This number of days multiplied for 24 hours should give 8808 records instead of 7709. The table below shows which months were lacking of observation days and how much:

Months	Records/ hours in month
January	0.98
February	0.29
March	0.85
April	0.90
May	0.97
June	1.00
July	1.00
August	1.00
September	0.97
October	0.99
November	0.74
December	0.75

Below table shows columns descriptions and the action that was taken for the column. Columns with “Filtrated” phase in name had corelation factor 1 with corresponding to them not filtered data, so they were removed. The last six columns, related to V_MAX_Rise_vs_... contain combination of data that were included in different columns so also were not taken not account. As it comes out of the original report, to each substations there were many customers connected. According to the percentage share of PV installations related to the total number of customers connected to the substation the mutual influence of substation and PV installation can be strongly disturb by the others connected customers. Besides that, substations data are lacking 20 % or 43 % of data, so all this columns were removed. For columns lacking the data from 01 to 31.10.2013 nothing was done because later this table was merged with weather data table which were available from 09.12.2013. The new column ‘Time’ was created to unify the time in dataFR and weatherFR data frames.

Number	Column name	Non-null count	Dtype	Taken action
0	SerialNo	7709	object	Removed
1	Substation	7709	object	
2	datetime	7709	object	
3	t_date	7709	object	Removed
4	t_time	7709	object	Removed
5	d_y	7709	int64	Removed

6	d_m	7709	int64	
7	d_d	7709	int64	Removed
8	d_w	7709	int64	Removed
9	t_h	7709	int64	Removed
10	t_m	7709	int64	Removed
11	V_MIN	7709	float64	
12	V_MIN_Filtered	7709	float64	Removed
13	V_MAX	7709	float64	
14	V_MAX_Filtered	7709	float64	Removed
15	I_GEN_MIN	7709	float64	
16	I_GEN_MIN_Filtered	7709	float64	Removed
17	I_GEN_MAX	7709	float64	
18	I_GEN_MAX_Filtered	7709	float64	Removed
19	I_IMPORT_MIN	7709	float64	
20	I_IMPORT_MIN_Filtered	7709	float64	Removed
21	I_IMPORT_MAX	7709	float64	
22	I_IMPORT_MAX_Filtered	7709	float64	Removed
23	P_GEN_MIN	7709	float64	
24	P_GEN_MAX	7709	float64	
25	P_IMPORT_MIN	7709	float64	
26	P_IMPORT_MAX	7709	float64	
27	Q_GEN_MIN	7709	float64	
28	Q_GEN_MAX	7709	float64	
29	Q_IMPORT_MIN	7709	float64	
30	Q_IMPORT_MAX	7709	float64	
31	S_GEN_MIN	7709	float64	
32	S_GEN_MAX	7709	float64	
33	S_IMPORT_MIN	7709	float64	
34	S_IMPORT_MAX	7709	float64	
35	thdV_MIN	6988	float64	Lacking the data from 01-31.10.2013
36	thdV_MIN_Filtered	6988	float64	Removed
37	thdV_MAX	6988	float64	Lacking the data from 01-31.10.2013
38	thdV_MAX_Filtered	6988	float64	Removed
39	thdI_GEN_MIN	6988	float64	Lacking the data from 01-31.10.2013
40	thdI_GEN_MAX	6988	float64	
41	thdI_IMPORT_MIN	6988	float64	
42	thdI_IMPORT_MAX	6988	float64	
43	f_min	6984	float64	
44	f_max	6988	float64	
45	Substation_VA_MIN_Filtered	6104	float64	Removed - lacking 20% of data
46	Substation_VA_MAX_Filtered	6106	float64	
47	Substation_VB_MIN_Filtered	6103	float64	
48	Substation_VB_MAX_Filtered	6105	float64	
49	Substation_VC_MIN_Filtered	6102	float64	
50	Substation_VC_MAX_Filtered	6104	float64	

51	Substation_thdVA_MIN_Filtered	4401	float64	Removed – lacking 43 % of data
52	Substation_thdVA_MAX_Filtered	4401	float64	
53	Substation_thdVB_MIN_Filtered	4401	float64	
54	Substation_thdVB_MAX_Filtered	4401	float64	
55	Substation_thdVC_MIN_Filtered	4401	float64	
56	Substation_thdVC_MAX_Filtered	4401	float64	
57	V_MAX_Rise_vs_VA_MIN_Filtered	6104	float64	Removed - contain combination of data that were included in different columns
58	V_MAX_Rise_vs_VA_MAX_Filtered	6106	float64	
59	V_MAX_Rise_vs_VB_MIN_Filtered	6103	float64	
60	V_MAX_Rise_vs_VB_MAX_Filtered	6105	float64	
61	V_MAX_Rise_vs_VC_MIN_Filtered	6102	float64	
62	V_MAX_Rise_vs_VC_MAX_Filtered	6104	float64	

Tables below include descriptions of data and were taken directly from PV Tool - Dataset Notes - PLEASE READ - Updated 2014-12-02 - 2.docx available to download from: <https://data.london.gov.uk/dataset/photovoltaic--pv--solar-panel-energy-generation-data>

Explanation of Field Names

For all rows:

SerialNo	Serial Number of the LvaDevice
Substation	Substation Name
Type2	"Feeder" (At the substation), or "Endpoint"
To	Feeder name, or for Endpoints this will say "Customer" or "Network". (Network means the ones we were previously calling service cables, I'm now calling them Network Endpoints because service cable implies they are only looking at one customer, and some of them are looking at multiple phases and multiple customers.)
datetime	Date & Time
t_date	Date
t_time	Time
d_y	Year
d_m	Month
d_d	Day
d_w	Day of Week
t_h	Hour
t_m	Minute
VA,VB,VC	A/B/C Phase Voltage (V)
IA,IB,IC,IN	A/B/C Phase Current (A)
PA,PB,PC	A/B/C Phase Real Power (kW)
QA,QB,QC	A/B/C Phase Reactive Power (kvar)
SA,SB,SC	A/B/C Phase Apparent Power (kVA)
thdVA,thdVB,thdVC	A/B/C Phase Voltage Total Harmonic Distortion (%)
thdIA,thdIB,thdIC	A/B/C Phase Current Total Harmonic Distortion (%)
f	A Phase Frequency (Hz)
Ignore_VoltageNotConnected	Flags records that come from LVAdedices with no voltage connected (you can ignore this field)
Ignore_CurrentNotConnected	Flags records that come from LVAdedices with no current connected (you can ignore this field)
_Filtered	Voltage Columns = Voltages <200V removed

	Current Columns = Currents >500A removed thdV Columns = thdV measurements removed whenever the corresponding voltage is <200V
--	--

For All Endpoints:

Substaton_	Indicates measurements taken at the substation
Rise fields	Difference between endpoint voltage and substation voltage. For customer endpoints, the customer's (single-phase) voltage is compared to all three phases at the substation. For network endpoints, phase A at the endpoint is compared to phase A at the substation, etc. <u>NB I have not confirmed that this relationship is correct, phase A at the endpoint may in fact correspond to phase B or C at the substation.</u>

For Customer Endpoints:

V	Single Phase Voltage
I_GEN	Generator Current (always +ve regardless of direction)
I_IMPORT	Import Current (always +ve regardless of direction)
P_GEN, Q_GEN, S_GEN	Generator Real Power (kW), Reactive Power (kvar), Apparent Power (kVA). +ve = Generating Should never be -ve
P_IMPORT, Q_IMPORT, S_IMPORT	Imported Real Power (kW), Reactive Power (kvar), Apparent Power (kVA). +ve = Import -ve = Export

Customer Endpoint Apparent vs Registered sizes

Substation	Apparent PV Size	Size in G83 Register
Bancroft Close	3.50	1.89
Forest Road	3.00	3.29
Suffolk Road	0.50	1.52
Alverston Close	3.00	3.29
YMCA	0.45	0.60
Maple Drive East	4.00	3.83

Weather data

Weather data for Forest Road location called weatherFR, primary contained 14625 rows and 38 columns. This data were collected from 09.12.2013 to 17.11.2014 with 30 min frequency. This number of days multiplied for 48 hours should give 16444 records instead of 14625. The table below shows which months were lacking of observation days and how much:

Months	Records/ half_hours in month
January	1.00
February	0.47

March	0.27
April	1.00
May	1.00
June	1.00
July	1.00
August	1.00
September	1.00
October	1.00
November	1.00
December	1.00

Below table shows columns descriptions and the action that was taken for the column. This data were taken every half an hour so to merged them with dataFR table only full hour records were used. The new shape was 7313 rows and 38 columns. Column THSWIndex had only '---' so was removed. In some columns records with index 4007 (according to the indexing after reshaping the data frame) contained '---'. This value was replaced with the mean of neighbouring indexes values. In case of 'WindDir' and 'HiDir' there were 475 records '---'. In these cases it probably meant no wind so was treated as separate category. All columns types besides Site, Date, Time, WindDir and HiDir were changed into float64. * Based on this categorical columns (WindDir and HiDir) WindDir_new and HiDir_new columns were created containing only int32 data according to this key: {'WSW': 0, 'NW': 1, 'W': 2, 'WNW': 3, 'SW': 4, '---': 5, 'S': 6, 'SSE': 7, 'NNW': 8, 'E': 9, 'ESE': 10, 'ENE': 11, 'N': 12, 'SE': 13, 'NNE': 14, 'NE': 15, 'SSW': 16}.

Number	Column	Non-null count	Dtype	Action taken
0	Site	14625	object	
1	Date	14625	object	Changed into datetime
2	Month	14625	int64	
3	Time	14625	object	
4	Hour	14625	object	Removed
5	TempOut	14625	object	
6	HiTemp	14625	object	
7	LowTemp	14625	object	
8	OutHum	14625	object	
9	DewPt	14625	object	
10	WindSpeed	14625	int64	
11	WindDir	14625	object	*
12	WindRun	14625	float64	
13	HiSpeed	14625	int64	
14	HiDir	14625	object	*
15	WindChill	14625	object	
16	HeatIndex	14625	object	
17	THWIndex	14625	object	
18	THSWIndex	14625	object	Removed
19	Bar	14625	float64	
20	Rain	14625	float64	
21	RainRate	14625	float64	
22	SolarRad	14625	object	
23	SolarEnergy	14625	object	

24	HiSolarRad	14625	object	
25	HeatD-D	14625	object	
26	CoolD-D	14625	object	
27	InTemp	14625	float64	
28	InHum	14625	int64	
29	InDew	14625	float64	
30	InHeat	14625	float64	
31	InEMC	14625	float64	
32	InAirDensity	14625	float64	
33	ET	14625	float64	
34	WindSamp	14625	int64	
35	WindTx	14625	int64	
36	ISSRecept	14625	float64	
37	Arclnt	14625	int64	

The Table below includes descriptions of weather data and was taken directly from Weather Data 2014-11-30.xlsx available to download from: <https://data.london.gov.uk/dataset/photovoltaic--pv--solar-panel-energy-generation-data>.

Date		
Time		
TempOut	Outdoor Temp	
HiTemp	Outdoor Hi Temp	
LowTemp	Outdoor Low Temp	
OutHum	Outdoor Humidity	
DewPt	Dew Point	
WindSpeed	Wind Speed	The WeatherLink samples the wind speed reading from the station a number of times during the interval (the actual number depends on your archive interval). Those readings are averaged to determine at the average wind speed for the interval.
WindDir	Wind Direction	The WeatherLink samples the wind direction reading a number of times during the interval (the actual number depends on your archive interval). If wind speed is greater than 0 when the WeatherLink samples wind direction, it places a "marker" into one of sixteen "bins" that correspond to the sixteen compass points. At the time of the archive, the WeatherLink determines which bin contains the most markers and writes the corresponding wind direction to the archive memory as the dominant wind direction.
WindRun	Wind Run	Wind run is a measurement of the "amount" of wind passing the station during a given period of time, expressed in either "miles of wind" or "kilometers of wind". WeatherLink calculates wind run by multiplying the average wind speed for each archive record by the archive interval.
HiSpeed	Hi Wind Speed	
HiDir	Hi Wind Direction	
WindChill	Wind Chill	
HeatIndex	Heat Index	<p>The Heat Index uses the temperature and the relative humidity to determine how hot the air actually "feels." When humidity is low, the apparent temperature will be lower than the air temperature, since perspiration evaporates rapidly to cool the body. However, when humidity is high (i.e., the air is saturated with water vapor) the apparent temperature "feels" higher than the actual air temperature, because perspiration evaporates more slowly.</p> <p>Note: WeatherLink uses the Steadman (1979 & 1998) formula to calculate Heat Index, which is more accurate than the method used by the Vantage Pro/Vantage Pro2 consoles and is calculated for all temperatures.</p>
THWIndex	Temperature Humidity Wind (THW) Index	The THW Index uses humidity, temperature and wind to calculate an apparent temperature that incorporates the cooling effects of wind on our perception of temperature.
THSWIndex	Temperature Humidity Sun Wind (THSW) Index	The THWS Index uses humidity, temperature, the cooling effects of wind and the heating effects of direct solar radiation to calculate an apparent temperature.

Bar	Barometric Pressure	<p>The weight of the air that makes up our atmosphere exerts a pressure on the surface of the earth. This pressure is known as atmospheric pressure. Generally, the more air above an area, the higher the atmospheric pressure. This, in turn, means that atmospheric pressure changes with altitude. For example, atmospheric pressure is greater at sea-level than on a mountaintop. To compensate for this difference in pressure at different elevations, and to facilitate comparison between locations with different altitudes, meteorologists adjust atmospheric pressure so that it reflects what the pressure would be if measured at sea-level. This adjusted pressure is known as barometric pressure.</p> <p>Barometric pressure changes with local weather conditions, making barometric pressure an important and useful weather forecasting tool. High pressure zones are generally associated with fair weather, while low pressure zones are generally associated with poor weather. For forecasting purposes, the absolute barometric pressure value is generally less important than the change in barometric pressure. In general, rising pressure indicates improving weather conditions, while falling pressure indicates deteriorating weather conditions.</p>
Rain	Rainfall	In calculating both daily and yearly rainfall totals for the software, the WeatherLink checks the total rain register on the station. The WeatherLink compares the current total rain value to the previous total rain entry in the archive memory to determine the amount of rainfall that occurred during the interval.
RainRate	High Rain Rate	Available for Vantage Vue, Vantage Pro and Vantage Pro2 stations only, the rain rate is calculated by measuring the time interval between each rainfall increment. When there is rainfall within the archive period, the highest measured value is reported. When no rainfall occurs, the rain rate will slowly decay based on the elapse time since the last measured rainfall.
SolarRad	Solar Radiation	What we call "current solar radiation" is technically known as Global Solar Radiation, a measure of the intensity of the sun's radiation reaching a horizontal surface. This irradiance includes both the direct component from the sun and the reflected component from the rest of the sky. The solar radiation reading gives a measure of the amount of solar radiation hitting the solar radiation sensor at any given time, expressed in Watts per square meter (W/m ²). The value logged by WeatherLink is the average solar radiation measured over the archive interval.
SolarEnergy	Solar Energy	The amount of accumulated solar radiation energy over a period of time is measured in Langleys. 1 Langley = 11.622 Watt-hours per square meter = 3.687 BTUs per square foot = 41.84 kilojoules per square meter
HiSolarRad	Hi Solar Radiation	
HeatD-D	Heating Degree-Days	One heating degree-day is the amount of heat required to keep a structure at 65°F when the outside temperature remains one degree below the 65°F threshold for 24 hours. One heating degree-day is also the amount of heat required to keep that structure at 65°F when the temperature remains 24°F below that 65° threshold for 1 hour.

CoolD-D	Cooling Degree-Days	Likewise, one cooling degree-day is the amount of cooling required to keep a structure at 65°F when the outside temperature remains one degree above the 65°F threshold for 24 hours. One cooling degree-day is also the amount of cooling required to keep that structure at 65°F when the temperature remains 24°F above that 65°F threshold for 1 hour.
InTemp	Indoor Temperature	
InHum	Indoor Humidity	
InDew	Indoor Dew Point	
InHeat	Indoor Heat Index	
InEMC	Equilibrium moisture content	Moisture content in wood affects both the size and strength of lumber. If you know the EMC of the storage or manufacturing area (which is derived using temperature and humidity readings), you can also determine the moisture content of the wood stored there.
InAirDensity	Air Density	Air Density (the weight of 1 cubic foot or 1 cubic meter of air) is a valuable tool for racing enthusiasts, because it helps determine the optimal jetting under current weather conditions.
ET		
WindSamp		
WindTx		
ISSRecept	ISS Reception Rate	The ISS Reception rate shows the percentage of wind data packets that have been successfully received by the Vantage Vue, Vantage Pro or Vantage Pro2 console.
ArcInt		

