Running Calculation Methodology for information mining in datasets

Author, date and version of the document

Etienne Bayenet
12 Rue Geischleid
L-9184 Schrondweiler
http://bayenet.jimdo.com

Date : 2018-01-28

Version : 2.1

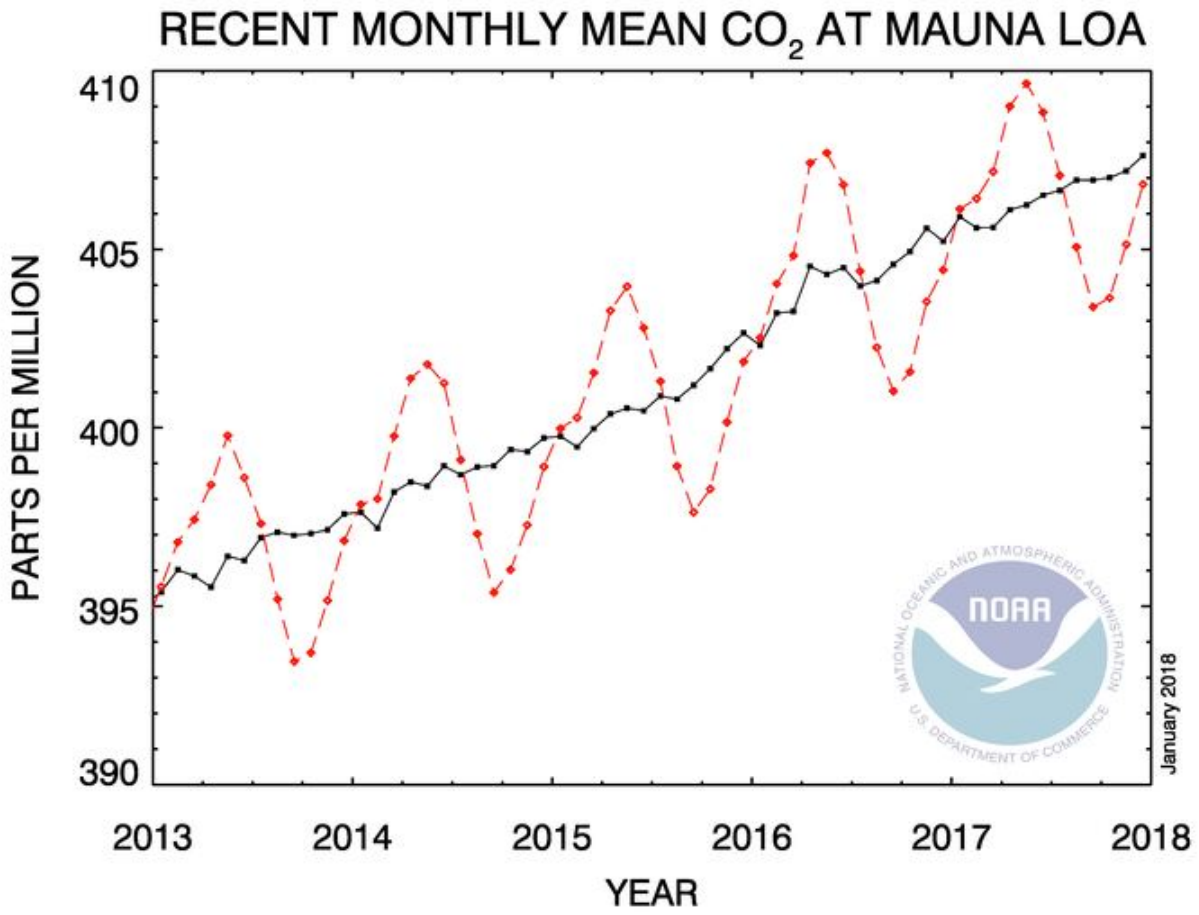
Table of contents

What is it about ?

The running calculation methodology is a method that has always been around in order to try to find trends in non linear data or to get some insight information out of fast changing values.

The given trend is not calculated, but the way data is presented on the graph should make it possible for the reader to understand what's happening.

A typical example is the CO2 concentration at Mauna Loa.



https://www.esrl.noaa.gov/gmd/ccgg/trends/

- Red dots show the current monthly average,
- Black dots show the average between the current and the 11 previous red dots.

The problem is that the interesting information is not always just an average of the values, sometimes it will be the daily maximum or minimum, a percentage between day and night, a sum if some parameters are true, the difference between actual and previous week...

All the different spreadsheet functions can be used : average, minimum, maximum, sum, if, sumif... even cosinus if it makes sense. Functions can also be mixed, for example a division of two sums in order to get a percentage.

Instead of calculating static values, the running calculation methodology takes advantage of the copy-paste function of the spreadsheet. Instead of calculating a daily average, 24

hours average can be calculated centered on each hour of the month.

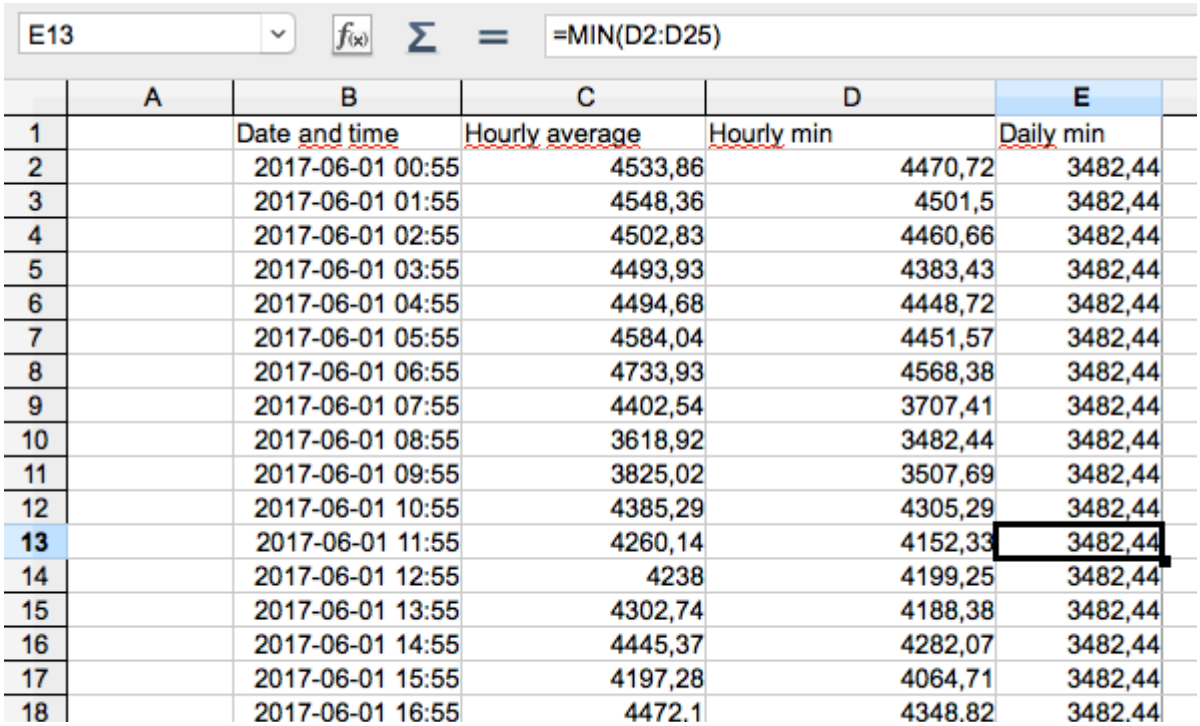
Any time range can be used, but the time range is a very important issue and will allow different information to come out of the graph. It is often interesting to show different time ranges on the same graph.

For example if you make an electricity consumption graph with a 30 days (a month) average and a 365 days (a year) average, you can compare short and long term evolution, see how you stand compared to the last year average... Monthly and yearly consumption can also be calculated and compared on a graph, but with the running calculation methodology, you will have a smooth line and it will be much less work to get the information. Furthermore you will not have to worry about 28, 30 and 31 months, and of how many week-ends are in the month. A small change in the formulas would even allow you to compare weekly and yearly values.

I first used the method in the context of facility management. Improved technology makes it possible to get much more datasets to control the facilities, but it is not always easy to get something meaningful out of a spreadsheet. I hope this document will provide you a good overview of the possibilities that this method offers and will help you to find the information you need.

Which time range should I use

A time range defines how much data will come in the formula. In the example bellow, there are hourly data and a daily minimum is searched, 24 hours will have to be considered in the formula.



E13 | =MIN(D2:D25)

	A	B	C	D	E
1		Date and time	Hourly average	Hourly min	Daily min
2		2017-06-01 00:55	4533,86	4470,72	3482,44
3		2017-06-01 01:55	4548,36	4501,5	3482,44
4		2017-06-01 02:55	4502,83	4460,66	3482,44
5		2017-06-01 03:55	4493,93	4383,43	3482,44
6		2017-06-01 04:55	4494,68	4448,72	3482,44
7		2017-06-01 05:55	4584,04	4451,57	3482,44
8		2017-06-01 06:55	4733,93	4568,38	3482,44
9		2017-06-01 07:55	4402,54	3707,41	3482,44
10		2017-06-01 08:55	3618,92	3482,44	3482,44
11		2017-06-01 09:55	3825,02	3507,69	3482,44
12		2017-06-01 10:55	4385,29	4305,29	3482,44
13		2017-06-01 11:55	4260,14	4152,33	3482,44
14		2017-06-01 12:55	4238	4199,25	3482,44
15		2017-06-01 13:55	4302,74	4188,38	3482,44
16		2017-06-01 14:55	4445,37	4282,07	3482,44
17		2017-06-01 15:55	4197,28	4064,71	3482,44
18		2017-06-01 16:55	4472.1	4348.82	3482.44

Cell E13 takes the minimum values between D2 and D25 (from 00:55 to 23:55), and cell E14 the minimum value between D3 and D26 (from 01:55 to 00:55 on the next day). A simple copy paste makes it possible to make a graph with daily minimum values on multiple years dataset.

Cells E2 to E12 have been filled with E13's value in order to have data on the border of the graph. This is sometimes interesting if you need a value for other calculations.

To define the size of the time range, it is important to know what is to be found.

The time range should match the real world context. Nature and finance have cycles (hours, morning, afternoon, half day, days, weeks, months, quarters, half years, years...) that have to be respected in order to get useful graphs :

- the time range can match commonly used cycles (hours, days, weeks...).
- the time range can be the result of a division of a commonly used cycle.

If the time range is not a commonly used cycle or a division of it, data could become inconsistent. For example a 5 days time range will be unstable because there will be dots with 5 working days and 0 weekend days considered (Monday to Friday) , and other dots with 3 working days and 2 weekend days (Friday to Tuesday). A better time range would be 7 days (a week) or one day.

Different time ranges can be presented on the same graph in order to compare short and long term behavior.

Full physical/financial cycle time ranges (an hour, a day, a week, a month a year...)

Hourly, Daily, weekly, monthly and yearly time ranges are very useful because they allow to compare the general evolution with a smooth line since a full cycle is used.

- Daily time range shows the evolution in weekly or monthly dataset. It makes it possible to analyze for example the impact of weather on electrical consumption.
- Weekly time ranges will show evolution in monthly or yearly dataset. It can be used for example to analyze water consumption. Weekend won't be seen on the graph.
- Monthly time ranges (30 or 31 days - the spreadsheet applications requires a constant number of days) makes it possible to see the evolution in yearly or multiple years dataset. If monthly values are used, daily averages have to be used otherwise February will always have the lowest result because it is the shortest month. The number of week-ends in the month might also create distortions in the results.
- Yearly value allow to see long term changes, and also make it possible to compare monthly values with general trends. In yearly values, seasonal changes don't appear.

Division of physical/financial cycle time ranges

The time range has to be shorter than a cycle when the aim is to detect behaviors or problems that happen inside a cycle. In that case, the graph will oscillate and what has to be analyzed is the difference betweens the ups and downs of each wave. This makes it possible to compare night and day, morning and afternoon, the different seasons...

Typical time range should be a division of the cycle (1/2, 1/3, 1/4, 1/5, 1/6...).

Here are some example of information that can be retrieved :

- to check a leak in the water pipes : a 2 to 4 hours consumption will make it possible to check night values (less AC, less cooking, less people in the bathrooms, less industrial production...). A higher than normal value means that there is a leak or an open tap. Small leaks will not be detected with this method.
- to check the electricity consumption and compare it to the activity in the facility. A 4 hours time range makes it possible to check activities and compare nights, mornings and afternoons.
  Night values makes it possible to check if systems are turned off correctly.
  Public holidays bring interesting data because many comfort systems (AC, air renewal, heating...) run (nobody thought at turning them off) and can be used as reference for an empty facility.

Alignment of lines when different time ranges are on the same graph

When different time ranges appear on the same graph, it has to be decided where the average data will be aligned with the individual data. Will the calculation be done :

- with the values before a datapoint,
- with values before and after the datapoint, the datapoint being in the middle of the range,
- with values after the datapoint.

For example, will a yearly average be placed on the 1st of January, the 1st of July or the 31st of December.

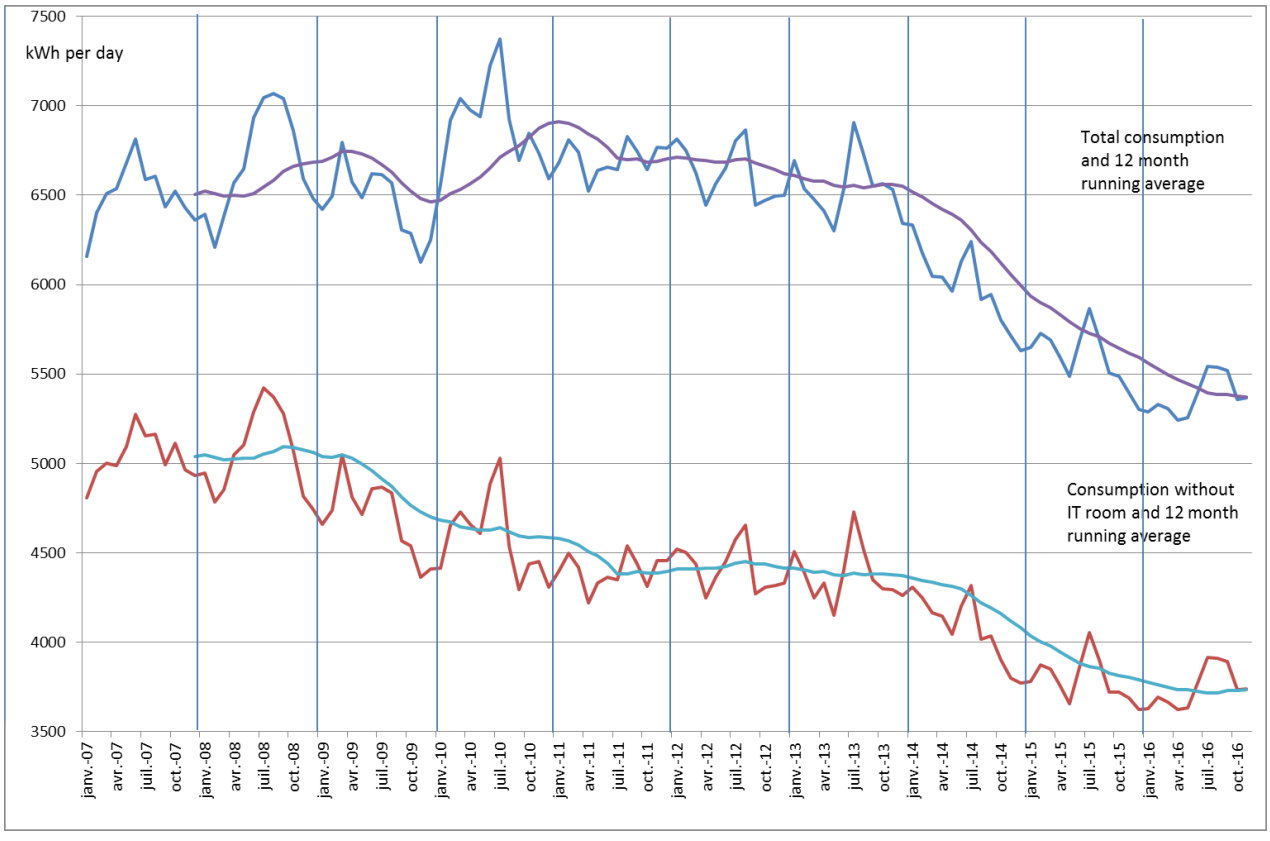
Two different contexts can to be considered :

1. The main goal is to compare the actual values with average values. In that case, calculation range will be before the datapoint.
2. The aim is to analyze how a system behaved in the past. In that case, having the datapoint in the middle of the calculation range will make it easier to compare the different lines on the graph.

I have no example of alignment where the calculation range would be after the datapoint, but this doesn't mean that it never would make sense.

Example of graphs with different time ranges

Graph with monthly and yearly average electricity consumption for an office building with IT Room. Datapoint at the end of the range of the average calculation.

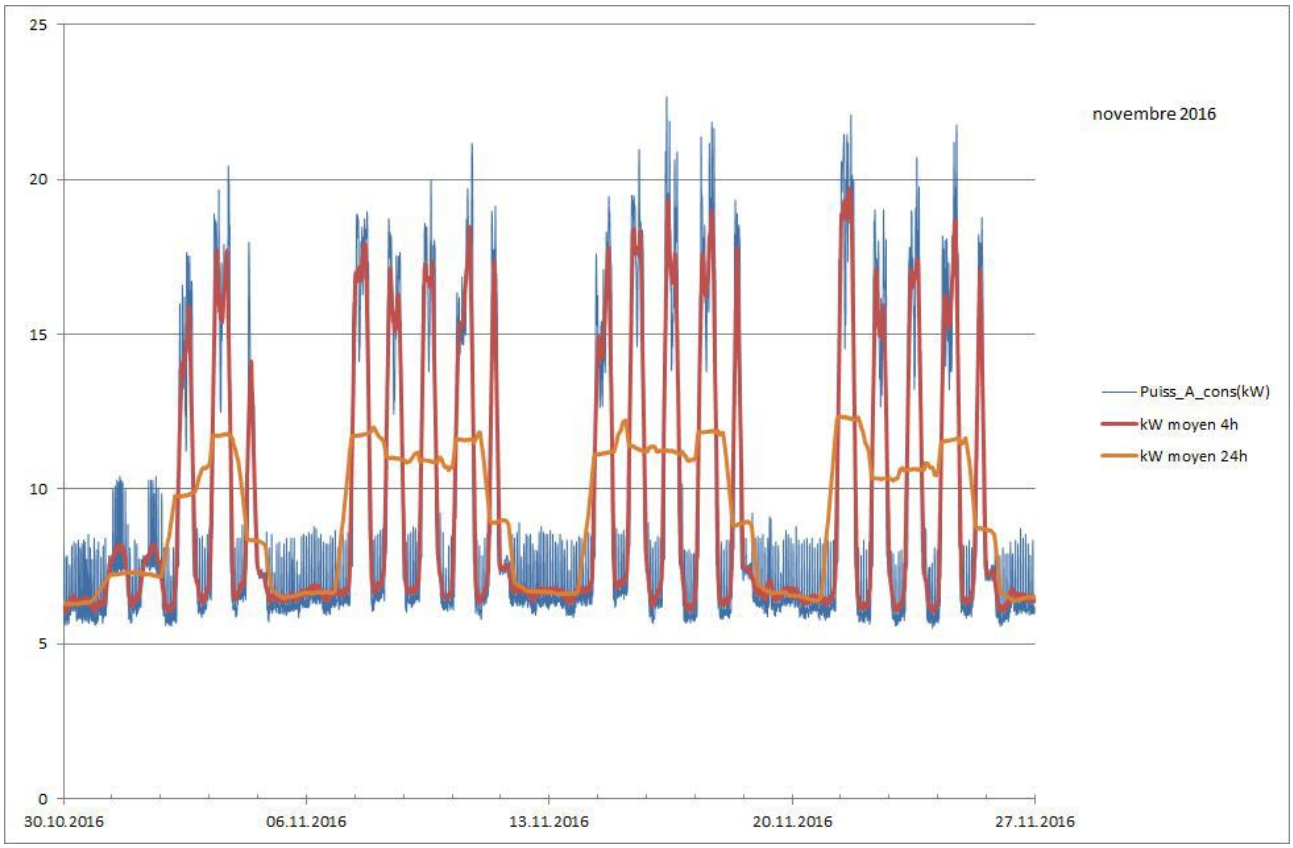


In this graph, yearly average represents the average of the last 12 months. This makes it possible to see if the current consumption is under or above average. Summer consumption is usually above average because of the AC.

In this graph, yearly average starts with one year delay. This is because you need a full year in order to have the first data point.

Yearly average could have been centered (2007 value presented on July 1st), but this would have made it more complicated to compare current value with average value.

Different time ranges averages electricity consumption on a building, datapoint in the middle of the range used for the average calculation.



- Blue line, 15 minutes power data, gives the information about minimum and maximum
- Red line, 4 hours power average, makes it possible to see if peaks are following the load, shows that the technical systems are down at night and that Friday afternoon has a lower load. The two first days were public holidays.
- Orange line, daily power average, provide information regarding the consumption of the day.

In this graphs, 4h average and daily average are calculated with the data point in the middle of the calculation range. This makes the comparison of the different lines easier. This graph doesn't aim to show how the building stands now, but how it stood in November 2016.

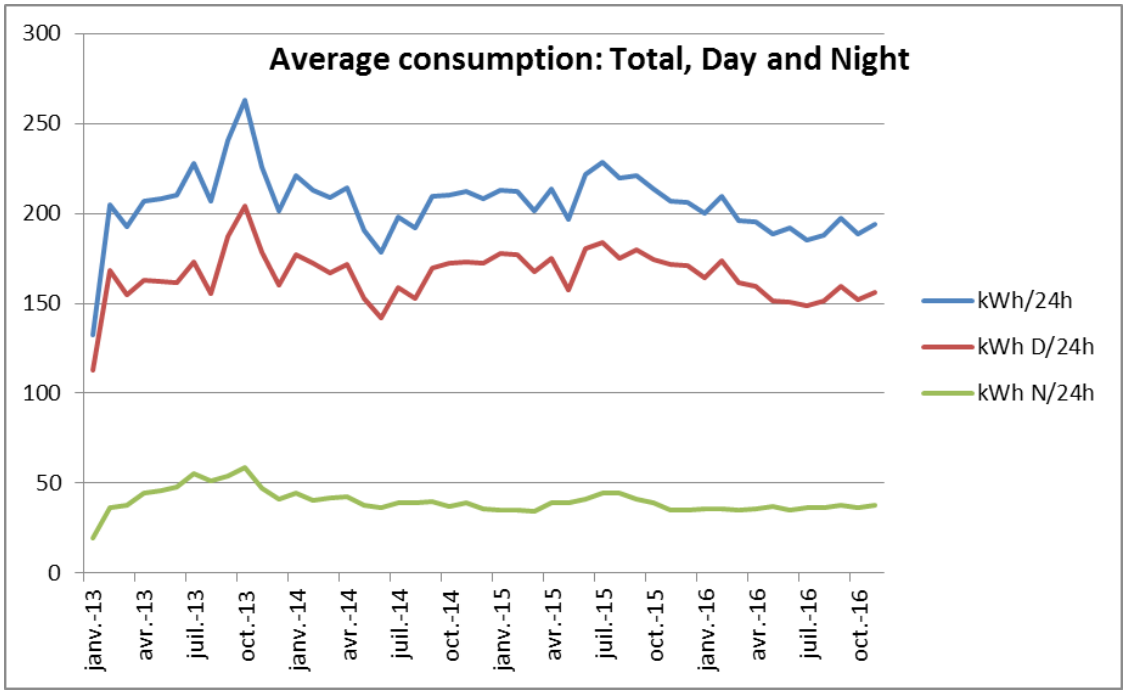
Which function should I use ? Some examples.

Any function available in the worksheet application can be used. Here are some examples.

Average

Average function is the most used. Here are some examples.

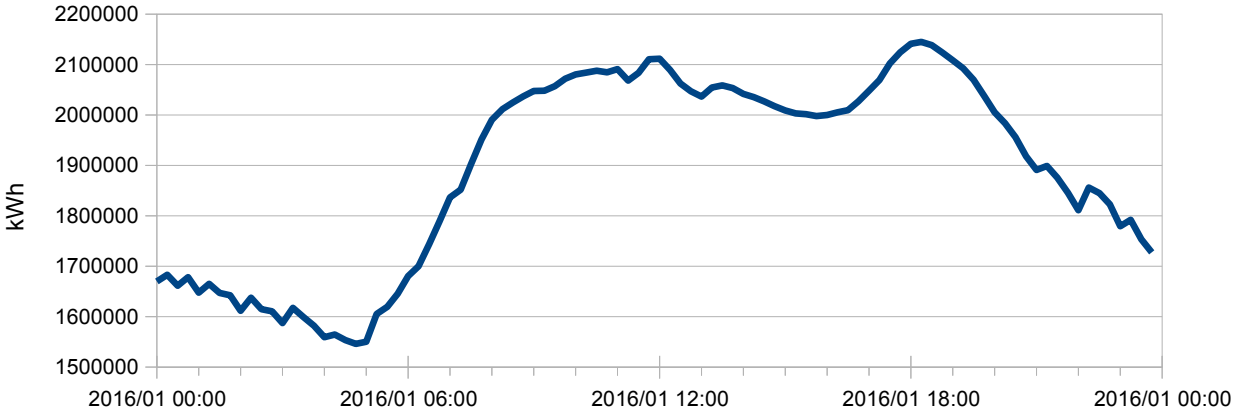
Monthly average consumption of offices



The average function just equalize the data. The wider is the time range, the smother is the curve.

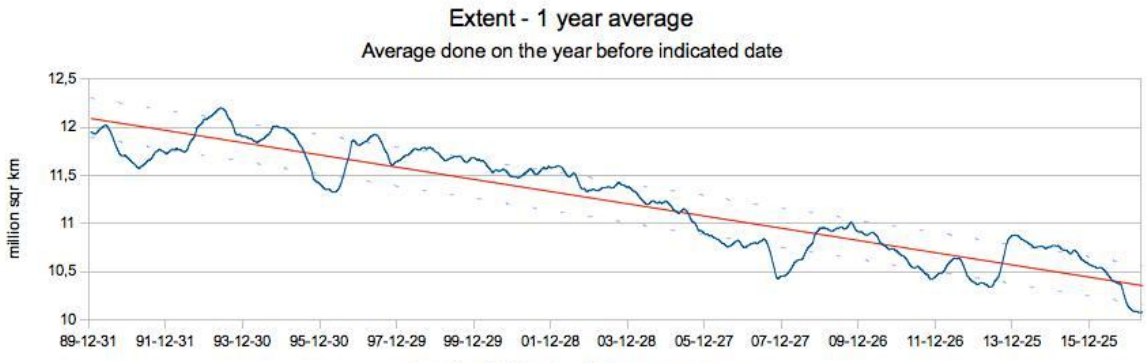
This graph was directly done with monthly data from the electricity invoice, and the spreadsheet calculation was to transform monthly data in daily average kWh to avoid lower value for shorter months (February, April...).

Hourly average on 15 days data



This graph has been realized with Swiss grid 15 minutes data and shows an average load curve based on 15 days. The evening peak occurs only in the winter (data are from January 1st to January 15th).

Yearly average of the sea ice extent

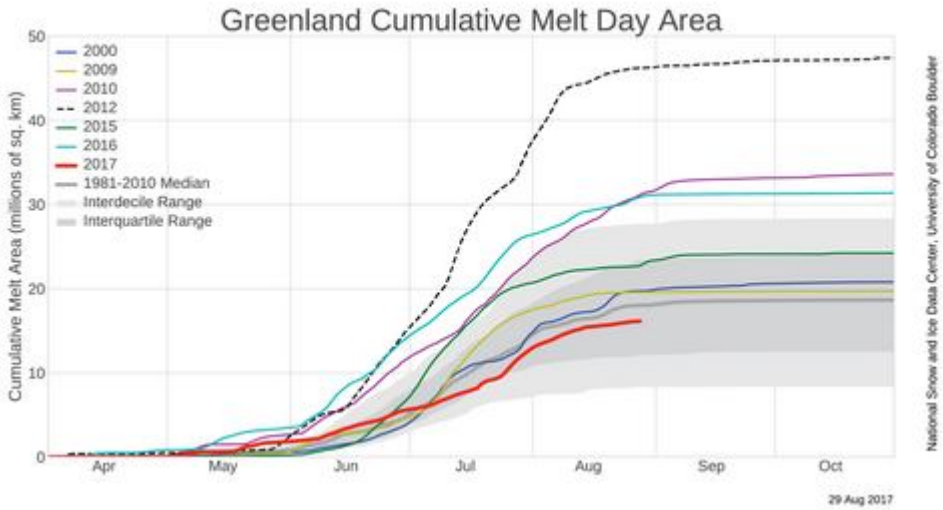


Yearly averages are interesting in this case because it includes minimal and maximal values, so it really shows how the sea-ice has behaved during that year, there is no winter or summer shift in the data. In this graph, data for the 95-12-31 (31st December 1995) is the average sea ice extent for 1995.

In this case, a linear trend was also calculated. Calculated trends makes sense here because ice is a physical element on which physical rules apply (ice won't melt less because it is Sunday, but electricity consumption could be lower because it is Sunday).

This graph was done with NSIDC data http://nsidc.org/arcticseaicenews/charctic-interactive-sea-ice-graph/ and was published in the Artic Sea Ice Forum https://forum.arctic-sea-ice.net/.

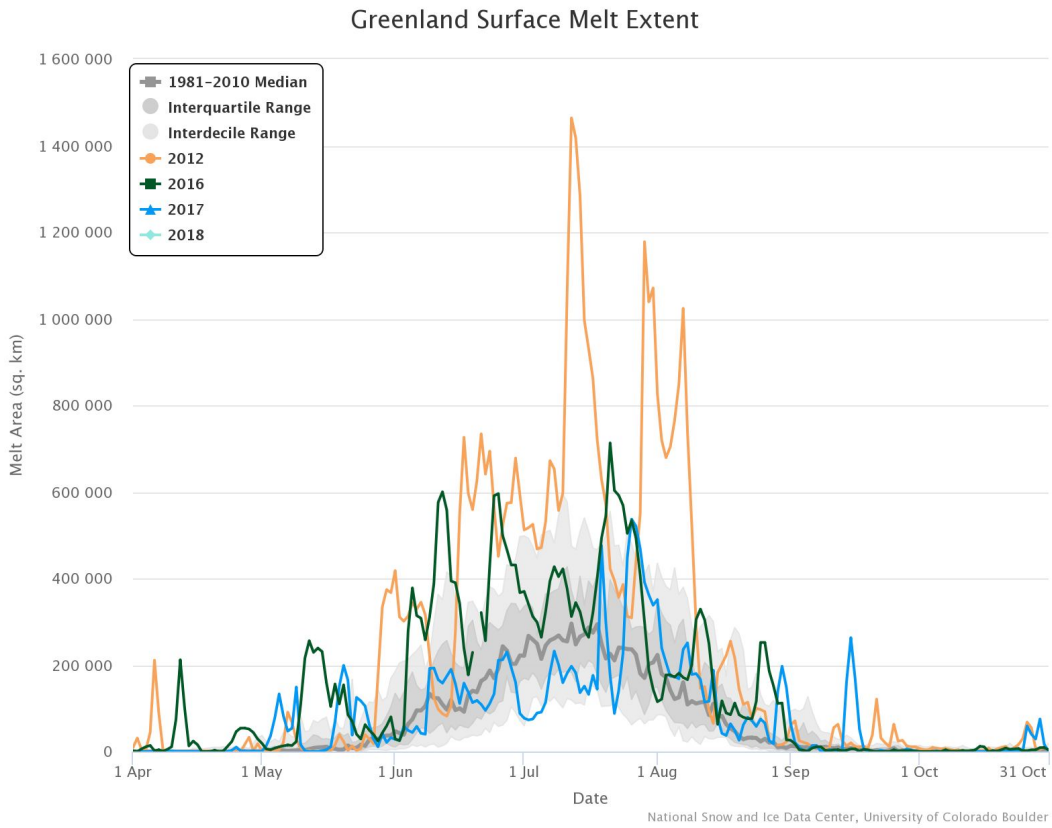
Mathematical integration



Greenland Melt Day Area is a daily calculation of how many sq. km of ice in Greenland have the upper surface melting. It helps for example to define how weather is impacting the Greenlandic ice sheet.

Mathematical integration on a spreadsheet is a formula where actual value is added to the sum of the previous values. It is mostly used to compare annual data like the water consumption of the AC.

Greenland Melt Day Area raw data looks like this :

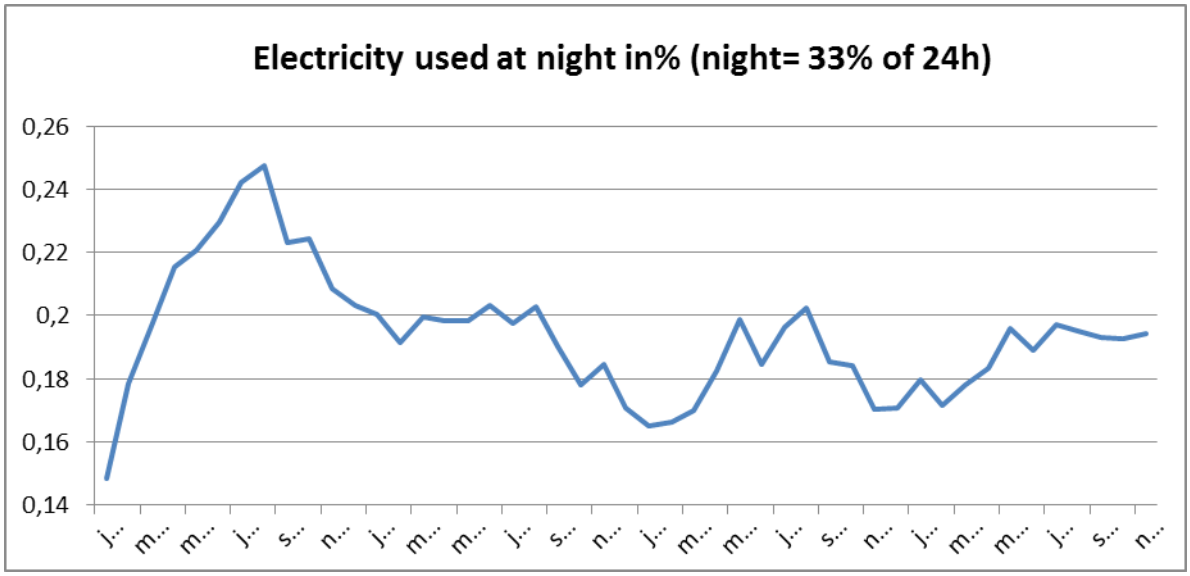


http://nsidc.org/greenland-today/

Percentage

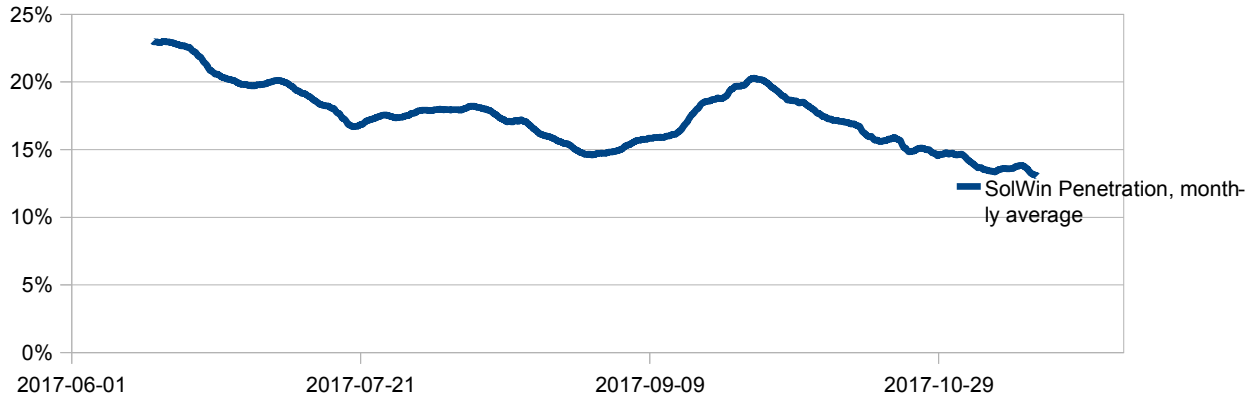
Percentage can be very efficient method to compare variable datas. It is a division between two sums.

Night/day consumption comparison for offices



At the beginning, printers and coffee machines were on all the time.

Solar and wind generated electricity penetration on the CAISO network



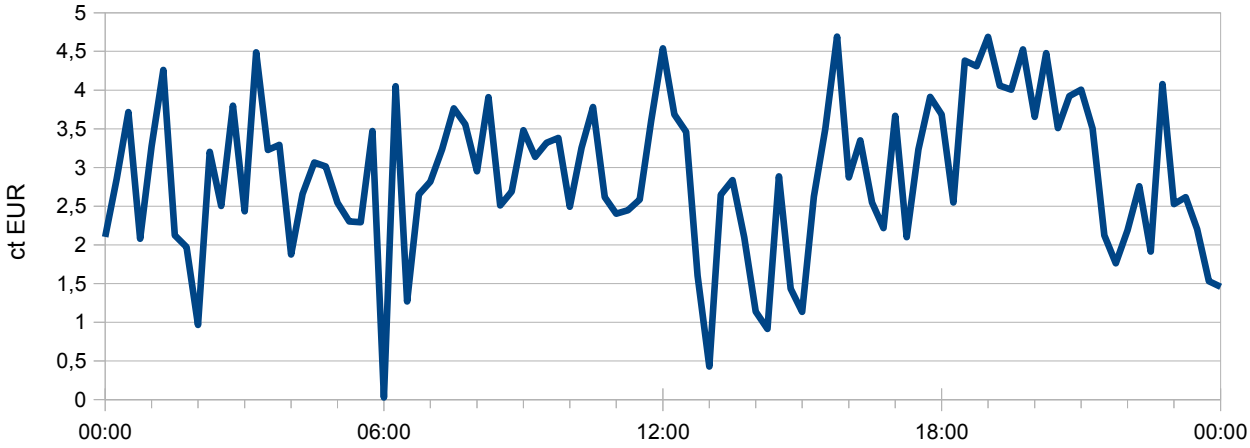
Penetration goes down in the winter because there is less sun and less wind. Data come from http://www.caiso.com/informed/Pages/ManagingOversupply.aspx.

Minimum and maximum values

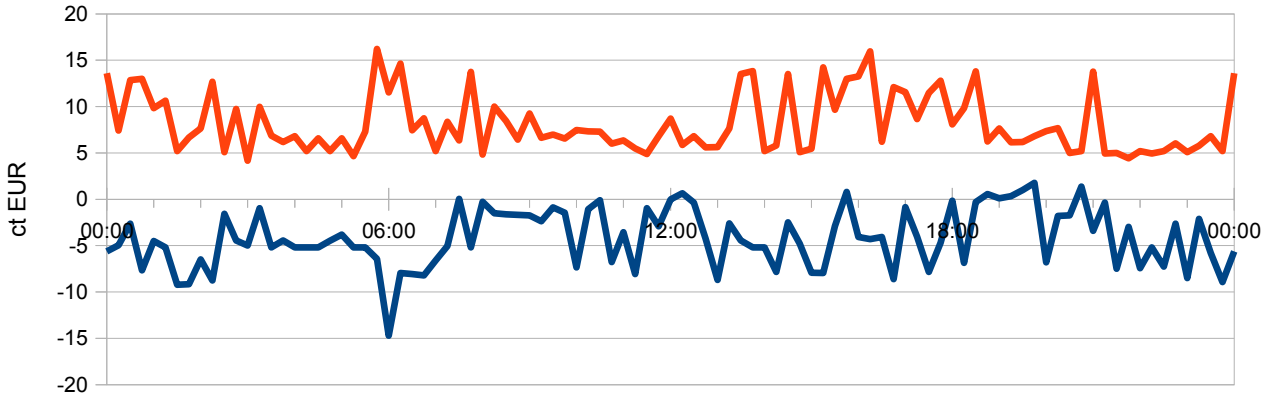
Balance electricity cost per kWh June 5th to 18th 2016, Luxembourg

These graphs represent the cost of the electricity used to balance grid between the load and the generation.

Average

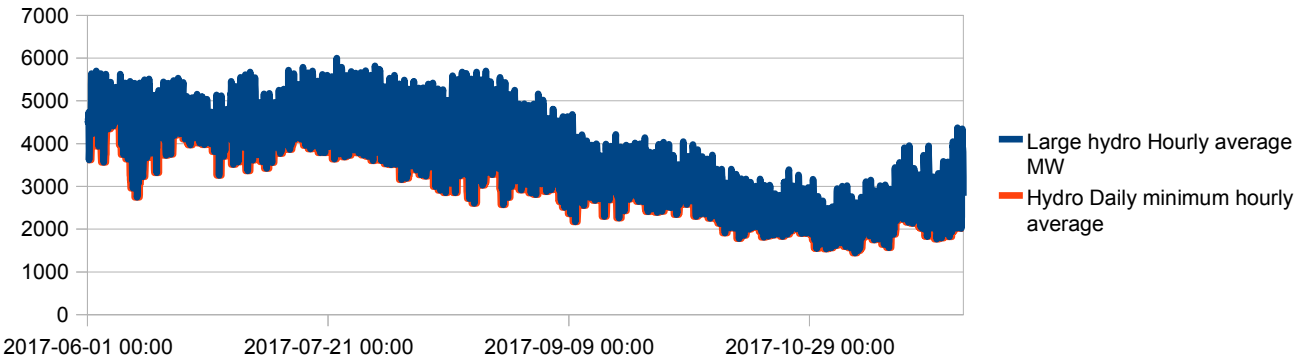


Minimum and maximum



06:00 seems to be the most complicated moment keep in balance.

CAISO, large hydro generation (hourly average and daily minimum)



The daily minimum generation value is interesting because it is a value that is probably achieved without pumping, which would mean that it is truly CO2 free. It makes it possible to calculate a hypothetical base load for a CO2 free large hydro generation.

Sum

The sum can be used, but it is not always easy to handle because displayed value is not in the same range than measured value. It can be used for example if we have hourly values for water consumption, but the final graph should be in daily values. In such a context, displayed value would be 24 time bigger than the original data.

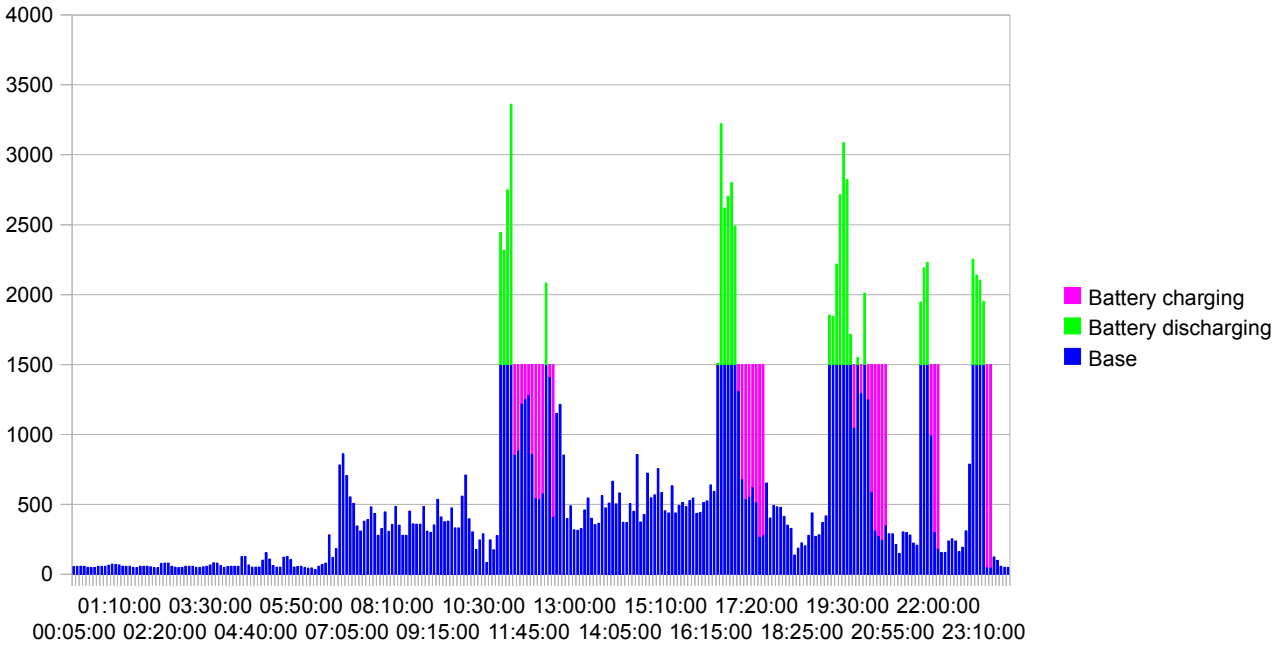
If Then Else

Conditional and boolean functions can be used if specific situations are to be analyzed. In the displayed example, the aim is to check how much power would be needed to do the peak-shaving of a single family house's load.

The graph on the next page shows 5 minutes data with a peak shaving at 1,5 kW.

The used function to create such a graph is quite simple :

- if (power>1,5 kW) then (add to battery requirement the consumption above 1,5 kW).
- if (power<1,5 kW and battery requirement >0) then (remove from battery requirement the power available under 1,5 kW)



Once all the dataset was handled, a search of the maximum battery requirement provided a value around 1 kWh of battery, which is compatible with a Vehicle to Grid concept.

The real function was a little bit more complicated because the efficiency of the battery loading was set at 80%.

This graph has been used to design a V2G T-shirt
https://schrondweiler.teemill.com/product/v2g/

Managing the dataset

Managing the dataset doesn't have good and bad solutions. Here are some typical problems with some easy solutions.

The solutions that are presented here are more for one shot analysis by people with normal knowledge of spreadsheet. If that type of actions is to be repeated regularly, there are more efficient methods to do it, but they require more competences from the user.

If you are yourself a spreadsheet specialist, you will not learn much in these pages.

Filling missing data points

Data points might be missing because of measurements problems or because null values are not provided by the data manager.

In order to be able to calculate with datas from different origins, it could be required to have the same number of data points in each dataset.

Here is an easy way to add or find the missing lines.

1. **Importing the data in a "help worksheet"**

   It is very important to keep the original dataset unchanged. So the first step will always be to copy the data in another worksheet.

2. **Adding a column "Status"**

   A status column is added to the help worksheet. All existing data are labelled something like "Original".

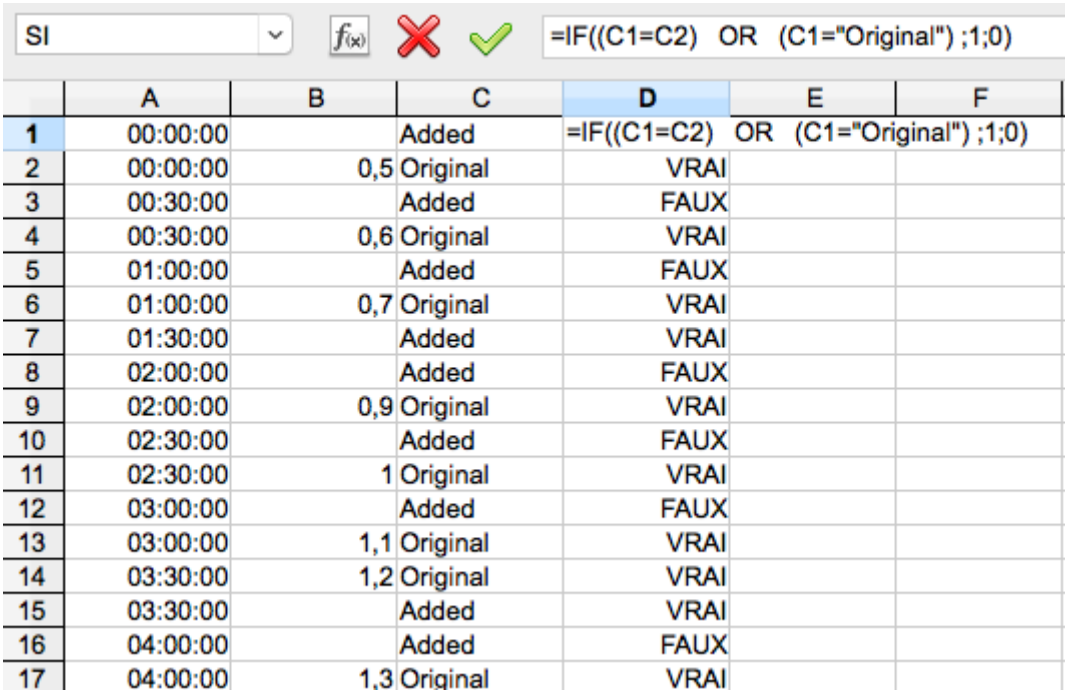
3. **Adding data in the date and time column**

   In the date and time column, an entry has to be added for each required data point. Their status value will be "Added"

4. **A one step sorting the data first according to the date and time column than according to the status column ("Added" values will be before "Original" values).**

5. **Adding a column "Data to be removed"**

   This column will contain an if then else function. If (("status of current line" = "status of next line") OR ("status of current line" = "Original") then "line is to be kept" else "line is to be erased").



SI =IF((C1=C2) OR (C1="Original") ;1;0)

	A	B	C	D	E	F
1	00:00:00		Added	=IF((C1=C2) OR (C1="Original") ;1;0)		
2	00:00:00	0,5	Original	VRAI		
3	00:30:00		Added	FAUX		
4	00:30:00	0,6	Original	VRAI		
5	01:00:00		Added	FAUX		
6	01:00:00	0,7	Original	VRAI		
7	01:30:00		Added	VRAI		
8	02:00:00		Added	FAUX		
9	02:00:00	0,9	Original	VRAI		
10	02:30:00		Added	FAUX		
11	02:30:00	1	Original	VRAI		
12	03:00:00		Added	FAUX		
13	03:00:00	1,1	Original	VRAI		
14	03:30:00	1,2	Original	VRAI		
15	03:30:00		Added	VRAI		
16	04:00:00		Added	FAUX		
17	04:00:00	1,3	Original	VRAI		

I have a French version, so VRAI = TRUE and FAUX = FALSE

6. **Sorting the data according to the "Data to be removed" column**
7. **Erasing the lines to be removed, and the status column.**

With these 7 steps, date and time consistency of the data can be insured.

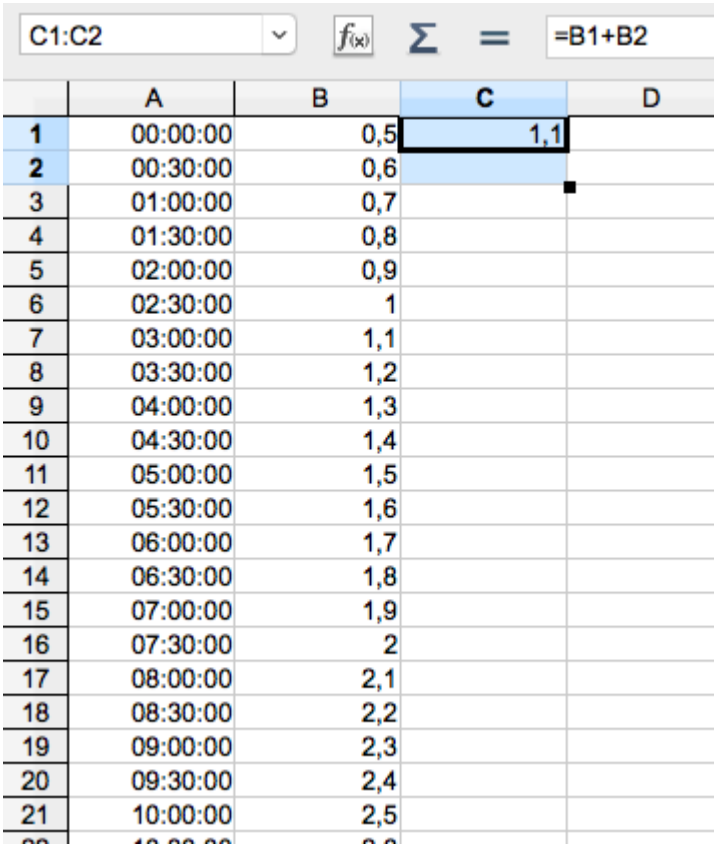
The filling of the empty data cell can be done using a similar method but has to be coherent with reality.

Increasing the time covered by a data point

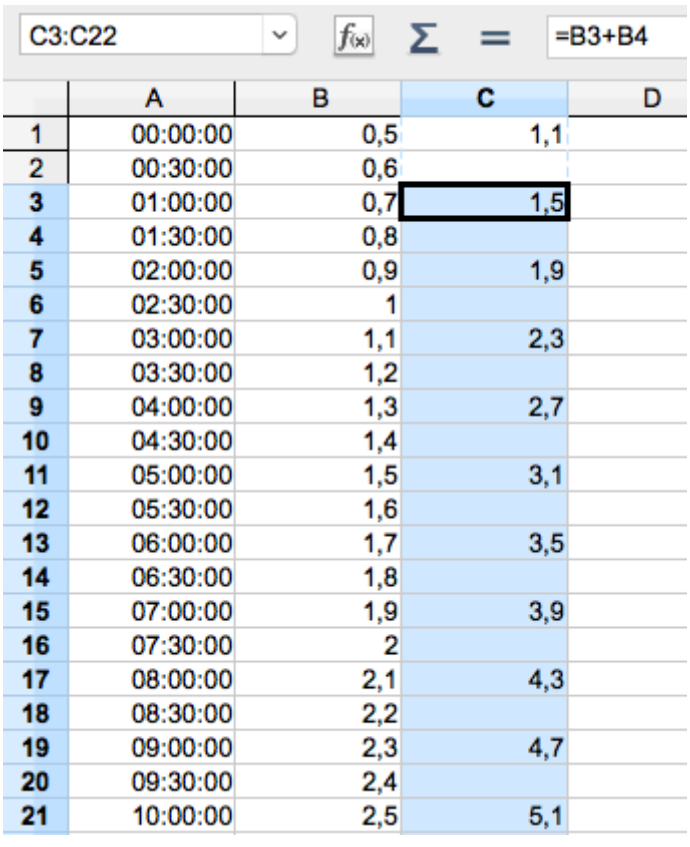
Sometimes there are just too many data points to be able to work easily with it. For electrical consumption, 5 minutes data points could be transformed in hourly data points.

Spreadsheets have a very interesting feature which is that when there is too much space selected, pasted cells are repeated until the selected area is full.

Here is an example where half hour data are transformed in hourly data.



	A	B	C	D
1	00:00:00	0,5	1,1	
2	00:30:00	0,6		
3	01:00:00	0,7		
4	01:30:00	0,8		
5	02:00:00	0,9		
6	02:30:00	1		
7	03:00:00	1,1		
8	03:30:00	1,2		
9	04:00:00	1,3		
10	04:30:00	1,4		
11	05:00:00	1,5		
12	05:30:00	1,6		
13	06:00:00	1,7		
14	06:30:00	1,8		
15	07:00:00	1,9		
16	07:30:00	2		
17	08:00:00	2,1		
18	08:30:00	2,2		
19	09:00:00	2,3		
20	09:30:00	2,4		
21	10:00:00	2,5		



	A	B	C	D
1	00:00:00	0,5	1,1	
2	00:30:00	0,6		
3	01:00:00	0,7	1,5	
4	01:30:00	0,8		
5	02:00:00	0,9	1,9	
6	02:30:00	1		
7	03:00:00	1,1	2,3	
8	03:30:00	1,2		
9	04:00:00	1,3	2,7	
10	04:30:00	1,4		
11	05:00:00	1,5	3,1	
12	05:30:00	1,6		
13	06:00:00	1,7	3,5	
14	06:30:00	1,8		
15	07:00:00	1,9	3,9	
16	07:30:00	2		
17	08:00:00	2,1	4,3	
18	08:30:00	2,2		
19	09:00:00	2,3	4,7	
20	09:30:00	2,4		
21	10:00:00	2,5	5,1	

In this example, when copying the cells C1:C2 in a larger area, the C1:C2 formulas are copied until the space is filled (C3:C22 in this example).

To check if the copy paste worked properly, it has to be controlled if the formula is always on the full hour line.

So transforming half hour data in hourly data was pretty easy. The last steps to do are :

1. Doing a paste special of the C column (hourly values) in order to save the values
2. Sorting the dataset according to the column containing the pasted hourly data.
3. Removing the lines where there is no hourly data.
4. Removing the columns that are not needed anymore.
5. Sorting the dataset according to the date and time column.

Calculating with a non constant but repeated number of data points

The idea here is for example to compare day and night electricity consumption. In most countries, the day is from 6:00 to 22:00, and the night from 22:00 to 6:00.

In this example, day data are summed at 10:00, and night data at 22:00.

Functions like =sumif() and =sumifs() can be used, but they are a little bit more complicated than the description below. It would be faster if the action has to be repeated on many datasets or if the dataset will grow as time goes. =sumif and =sumifs function

also have the advantage that they allow to work with datasets where some datas are missing.

1. Two columns are defined near the data set, one for the day values, and one for the night values.
2. At 00:00 of the first day, in the day values column is done the sum of the consumption between 6:00 and 22:00
3. At 00:00 of the first day, in the night values column is done the sum of the consumption between 00:00 to 05:59 and 22:00 to 23:59.
4. A copy paste of the formulas in the day values and night values is done on all the dataset. Be careful and copy all the empty cells between 00:01 and 23:59.
5. A paste special of the day and night values columns is done in order to have values and not formulas.
6. The dataset is sorted according to the day or night values column.
7. The lines without values in the day and night columns can be removed.
8. Original dataset other than the date column can be removed.
9. Sorting the dataset according to the date column will provide a dataset with day and night consumption for each day of the dataset.