Exploring data using gnuplot

Johan Montelius

November 25, 2013

1 Introduction

In this assignment you will explore some data using the gnuplot tool. You will learn how to generate graphs that describe your findings and how these graphs can be included in a document in an automatic process.

This is not a gnuplot nor a LaTeX tutorial. You will have to find other sources describing exactly how these tools work. However, by following the structure of this tutorial you should get a good start.

1.1 the files

If you're reading this text from the file `energy.pdf` you probably have the tools that you need to generate diagrams and the final pdf. You probably also have read the README, but just in case you haven't done so, here is the information again. You should have the following src files:

- energy.txt : the latex src. This is the file that you can use as a template to generate your report.
- 2011.dat : the data file containing statistics of electricity production in Sweden 2011
- 2012.dat : the data file containing statistics of electricity production in Sweden 2012
- summary.dat : summary of the electricity production 2011
- wind.p : a gnuplot script to show how to generate a small diagram that can be included in your report

You will need the following software:

- pdflatex: to generate a pdf from latex src
- gnuplot: to generate png from the data set

- make: to automate the process

There are other ways to generate a pdf from latex but using pdflatex is the simplest. If you have already been using LaTeX then feel free to use any method you like.

1.2 the data

The data you will use for this exercise is statistics on electricity production and consumption in Sweden. We have statistics for both 2011 and 2012. The files contains hour-by-hour data as follows:

- wday (1): 1-7, representing Monday to Sunday
- day (2): the day of the year 1–365/366
- date (3): 1-31, representing the day of the month
- month (4): 1-12, representing January-December
- year (5): the year 2011 or 2012
- time (6): 0-23, representing the hour of the day for each date
- total (7): the total electricity consumption, i.e. production and import/export
- wind (8): wind power production
- hydro (9): hydro electric power production
- nucl (10): nuclear power production
- res (11): gas and diesel power plants used in reserve or to balance production
- combined (12): electricity produced by combined heat and power plants
- other (13): ... well, other
- sun (14): seriously, does the sun ever shine
- import (15): import or, if negative, export
- price (16): the price on the NordPool Spot market in SEK/MWh

All electricity figures are given in MWh and the price is in SEK per MWh. The data, except for the price information, is from Svenska Kraftnät (Swedish National Grid) [?], a state-owned public utility. It is publicly available from their web site but you need some editing to get into a cvs file.

If you browse the data you will quickly see that hydro electricity and nuclear power is what powers Sweden but there is a lot more to learn from this data.

source	2011	2012
hydro	67	79
nuclear	58	62
combined heat power	9.8	8,2
wind	6.2	7.3
total	141	157

Table 1: Total electricity production in 2011 in TWh

2 electricity production in Sweden

The aim of this assignment is to learn how to explore data using gnuplot. You need not know everything about electricity production to complete your report but if you know nothing at all it will of course be very hard to interpret the data. This is thus a very short introduction to the Swedish market when it comes to electricity. More information about Swedish electricity production can be found at Svensk Energi [**?**], a non-profit industry and special interest organization.

2.1 water and nuclear

Swedish electricity production is, as shown in the table **??**, completely dominated by hydro electric and nuclear power. Sweden is basically self-sufficient, but depending on how much water there is in the reservoirs or how cold the winter is, we need to import or can export electricity (during 2011 we exported about 7 TWh and in 2012 as much as 19 TWh).

The total amount of *installed* hydroelectric power is approximately 14 GW and the power plants operate at 50 percent of their capacity over a year. There are ten nuclear power plants with a total installed capacity of approximately 9 GW. During 2011, the availability was low (around 70 percent) due to several major upgrades, 2012 was a much better year and production was close to normal (almost 80 percent). As you will see, Swedish power consumption varies over the year and both hydroelectric and nuclear power plants operate at low capacity over the summer; hydroelectric power to save water and nuclear power to make yearly revision.

2.2 gas, heat and wind

There are few gas or diesel power plants; these are mainly used to balance the electricity in the grid or as a last resource during cold winter days. However, there are a number of combined heat and power plants. These are fueled with everything from garbage and forest products to diesel and are used to produce both heat and power. You might ask why it is necessary to produce heat and how it is sold, but most larger cities in Sweden have city-

wide central heating systems that will heat a large part of the city buildings during winter.

Wind power has increased during the last ten years from close to nothing to an installed base of almost 3.6 GW by the end of 2012. This was a record year for wind production with a total production of more than 7.3 TWh, which is comparable to one of the smaller nuclear power plants. A wind power turbine typically has an effect of 1-3 MW and the total number of turbines were at by the end of 2012 close to 2400. The yearly average availability is around 25 percent but the production can vary from a few percent to eighty percent of the installed base.

2.3 the balance

Since there are limited resources to store electricity you always want to produce exactly what is needed so that the production is not wasted. Power consumption varies a lot on daily, weekly and (naturally) yearly bases, making the balance of the production a major problem. Different types of power plants have very different characteristics when it comes to adapting to changes in power consumption.

Nuclear power plants cannot change their production quickly (and the fuel does not cost very much) so you would like to turn them on and have them running as much as possible. Hydroelectric plants can alter their production quite rapidly and can cope with changes on a daily and even hourly basis. To cope with the quickest changes you need gas or diesel plants, which function as a reserve to balance the grid.

Wind farms are the least helpful in this process since production is fully dependent on the wind. If the wind blows you want to produce as much as possible even if demand is low. If the total wind production increases you have to increase the amount of gas or diesel power plants since the variability in the network will increase.

3 exploring the data

The first thing that you should do is explore the data using gnuplot in interactive mode. You could apply some statistical methods to the data and calculate, for instance, the mean, median and variance, but before you do this, you should get an understanding of what the data looks like and how the different parameters are linked to each other.

3.1 one parameter

Start gnuplot and type the following:

```
plot "2011.dat" using 7
```

This command will create a plot of all the entries in column 7 (total consumption). The 8760 entries represent the hours of the year. What do you see in this graph? What is the pattern due to?

If you do not see a plot in a window it's probably because the terminal is not set correctly. On a Linux system it should be for example `x11` *and on a Windows system it could be* `win`*. Set the terminal with the command* `set terminal`

The diagram does not contain any additional information, labels or titles. In this phase of your work you don't need to care too much about readability. You know what the graph represents and the information is for your eyes only.

You can choose to plot the consumption at a given hour every day using the *every* directive. The example below will plot two sets of data, one that starts the first element (0) and one that starts on the twelfth (11). Look at the graph. Why are the who data sets so different? What happens in July?

```
plot "2011.dat" every (24)::0 using 7, \
     "2011.dat" every (24)::(11) using 7
```

The backslash is there to break the line and is needed only if you write the command on two lines.

In gnuplot you can do simple things such as grouping entries. If we want to find out how the total energy consumption varies by month we can issue the following plot command:

```
set boxwidth 0.5
plot "2011.dat" u 4:($7/1000) smooth frequency with boxes
```

With this command we want to group the data based on the value of the forth column and say that the aggregated value in each group should be the sum of the values in the seventh column divided by 10^3 (the y-axis is thus in GWh). The $ notation is a way to access the value of the seventh column inside a mathematical expression.

The *smooth frequency* directive also sorts the groups based on their group value, i.e. 1-12. Hence, it does not matter in what order the entries are written, since the boxes will always be plotted starting with the box for *January* (that is 1). What does the power consumption look like per month? Do a similar plot for the days of the year (column 2), you might use a *lines* plot instead of *boxes* to more easily see the result (using a line is however not correct since we have consumption per day and there obviously is no consumption between two data points).

```
plot "2011.dat" u 2:($7/1000) smooth frequency with lines
```

The consumption clearly has a frequency component of a handful of days, why is this? Can you plot a box diagram that clearly shows this?

You have the tools to quickly find the pattern of energy consumption during a day. It's quite interesting to see the pattern for different sources, and in what way and for what reasons they differ. Which sources can adapt to the hourly demand? Make graphs for wind, hydro and nuclear and see how well they adapt to the changes in demand.

3.2 distribution

To better understand questions such as how much consumption varies and what the most common values are, you can try the following command:

```
plot "2011.dat" using 7:(10*rand(0)-5)
```

This will plot the same values but now the consumption is on the x-axis, whereas the y-axis is random. You will quickly see that during most hours the consumption is between 11 and 15 GWh but that we have plenty of hours where the consumption is up to 25 GWh. Although this plot gives very little details, you quickly get a feeling for the data.

If we want to look at variation and learn for how many hours we have had different consumption values we can make a frequency plot.

```
plot "2011.dat" using 7:(1) smooth frequency w boxes
```

The *smooth frequency* directive will sort the entries in increasing order and add the values per entry. The directive `using 7:(1)` says "use column 7 and give each entry a value of one, regardless of its value". The result is thus a plot where we see for how many hours a we have had a specified consumption.

Since we have 8760 entries, this frequency plot is not very clear. Therefore, it is better to group the entries into larger bins, for instance of the size 200 MWh. This can be done with the following statements:

```
bin(x,s) = (floor(x/s)*s) + (s/2)

set boxwidth 200

plot "2011.dat" using (bin($7,200)):(1) smooth frequency with boxes
```

The *bin function* will take a value and group it into a value of, in this case, multiples of 200. We will thus get a diagram indicating for how many hours the consumption was from 0 to 200 MWh, 200 to 400 MWh, and so on. The `with boxes` statement gives us a histogram where the y-axis is number of hours that the consumption was in an interval of 200 MWh. The height of

the boxes (and since all boxes have the same width, also area) is the number of hours, but an hour where the consumption is 20GWh is of course twice as important as an hour where 10GWh is consumed. We can therefor do the similar plot:

```
plot "2011.dat" using (bin($7,200)):($7) smooth frequency with boxes
```

Now the boxes resembles how much energy that was actually consumed at different levels of usage. We see a slight change were the weight is moved towards the high end of the x-axis.

Take a look at the different energy sources. How do they differ? Which sources is the most valuable? What on earth is going on with nuclear power?

If we would like to have graph showing the distribution of how much value in Swedish crowns the wind farms produce we could do this with the following command:

```
set boxwidth 100000
plot "2011.dat" using (bin(($8*$16),100000)):(1) \
                 smooth frequency w boxes
```

This graph shows that for for approximately 1500 hours the farms produce a combined value of less than 100,000 SEK an hour. During roughly half the year the production is less than 300,000 SEK an hour. Do a similar plot for hydro electric production. What does the distribution look like?

3.3 discover correlation

Naturally, correlations are one thing we want to explore and scatter plots are a very good tool in understanding relationships between parameters. Try the following:

```
plot "2011.dat" u 7:16
```

Now we have a scatter plot of the total consumption and the price. As you see, the price goes up as demand increases. It's definitely not a direct relationship but there is clearly a relationship. You see groupings of data and you start to wonder what drives the price up to 700 SEK if the demand is still low. Or, why do we have prices that are close to zero if there is still a demand for 12 GWh? You could calculate a statistical correlation of the two variables but looking at the scatter plot gives you more insight in the relationship.

There are some forms of the *smooth* directive that can be used to reveal dependencies that are not obvious but your first approach should be to look at the raw data and get a basic understanding. If smoothing functions are used directly you might miss clusters of outliers that are significant in the understanding of the data.

3.4 the right data

You will notice that gnuplot is a great tool to generate graphs from data but it's not the best tool to transform data. For instance, it is not obvious how to get hold of the min and max values of a column or to calculate something as simple as the sum of all values. There are ways of doing this but we will not deal with such methods here.

One alternative strategy is to do the calculations when the data is generated. Since you are all programmers and will most likely look at data that you have produced yourselves, this is often a simple solution.

In the file *summary.dat* you will find statistics such as the mean, median, and first and third quartiles. Notice that we have the production types as the rows, while the columns give the various metrics. We do this in order to use the built-in clustered histogram.

```
set style histogram clustered gap 4
set boxwidth 0.5
plot "summary.dat" u 6:xtic(1) title columnhead w histogram , \
                "" u 6 title columnhead w histogram
```

Notice the use of "" to indicate that we use the same source and want to have the next column clustered with the first. You can easily add more of the columns to the plot.

Using the histogram plot might not be the best way to show this data as you will see later in this tutorial; it is only an example of a different technique that you should explore further.

4 Presenting your findings

Once you have explored the data it is time to decide what information you want to present and in what way this is best done. You need to decide if the point is most effectively made in a graph or as a table and, if a graph is chosen, how to design the graph.

4.1 tables or graphs

As an example, look once more at the data file *summary.dat* with statistical information on our current data. How would you like to present this information? should we use a table as in table **??**?

In this table, hydro electric power production is given with four or five significant figures, whereas wind power with three or four. This makes the table hard to read. One should try to be consistent regarding the number of significant figures given, and in this example it might be sufficient to give the numbers using two significant figures as shown in table **??**.

source	min	1'st	3'rd	max
hydro	2546	5808	9679	12839
nuclear	3713	5284	6787	9162
combined heat power	146	400	1789	2751
wind	15.0	350.8	992	2287

Table 2: Hourly electricity production in MWh by type

source	min	1'st	3'rd	max
hydro	2500	5800	9700	13000
nuclear	3700	530	6800	9200
combined heat power	140	400	1800	2800
wind	15	350	990	2300

Table 3: Hourly electricity production in MWh by type

This is much better and it becomes much easier to get a feeling of how the energy sources relate to each other. When you present numbers you should always ask yourself why you are doing so, and what you want your readers to understand. It is not always the case that numbers in a table give the best picture of your data. If you want to compare values, or show how two or more parameters are linked, it is often better to use a graph. Try, for instance

```
plot [0:9] "summary.dat" u 2:3:4:5:6:xtic(1) w candlesticks
```

and compare the readability of the graph and the tables above.

4.2 automating the process

Once you know what you want to produce, it's better to use a small script that can be run automatically. If you look in the file *wind.p* you will find the gnuplot script used to produce the graph in figure **??**. In this file you will also see how we have created better labels for the axes etc. There is much more that you can do to make the data more easy to read. To use the script you either run it from the gnuplot shell using the command,

```
load 'wind.p'
```

or you run it from the regular shell giving it as an argument to gnuplot.

```
gnuplot wind.p
```

The graph together with its caption should be self-contained. One should not need to read the article to understand what the graph is all about. The

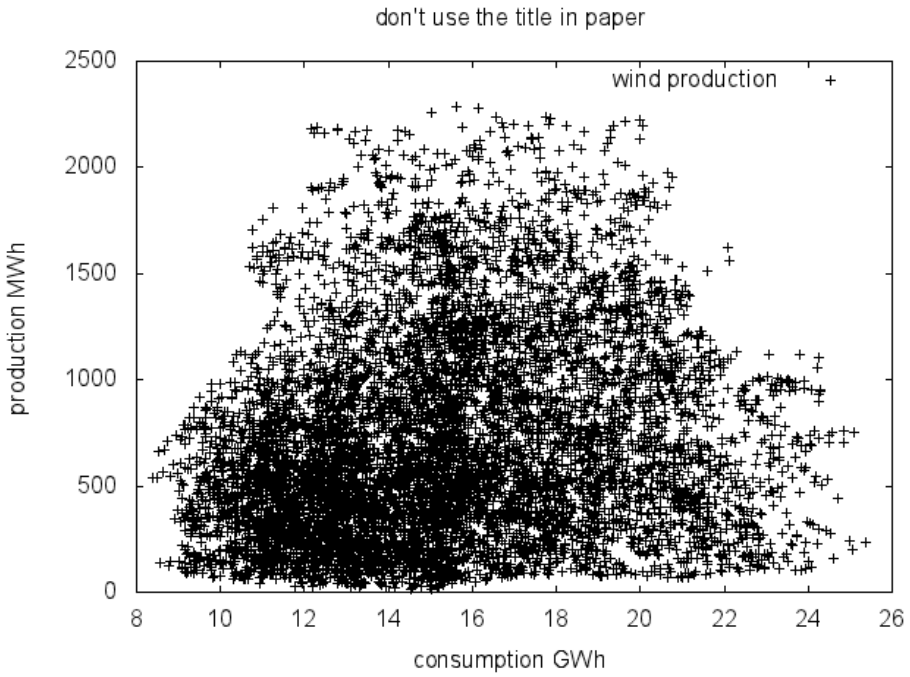

Figure 1: The correlation between total consumption and wind power production per hour.

gnuplot tool allows you to add a title in the graph but this is best avoided. It's better to include the information in the caption.

The image in this example is produced as a png file. You can produce other images. The LaTeX source shows you how to include the image in the document.

5 Your task

Your task is to explore the data on electricity production in Sweden and produce a report using this latex file, some gnuplot scripts and the make file as templates. The final pdf should be reproducible by running make and any changes in the data or scripts should be automatically taken into account.

The report should consist of two parts:

- The first part should show some graphs used to explore the data. You should include the gnuplot command used to generate the graph, a discussion of why you generated it, and your observations. The graphs need not be polished, but should be the rough graphs you produced when exploring the data.

- The second part should present a finding. In this part you should describe the data used and how it supports your finding. You should include one or two graphs that describe a situation. These graphs should contain titles, labels, grids, and any other items that make the graph easier to interpret.

The report should not be too long (approximately four pages), nor does your finding have to be very scientific. The idea is that you should learn how to explore a data set using gnuplot and how to include graphs automatically in a report. If you learn something about electricity production in Sweden, that is a bonus.

References

[1] *LaTeX http://tug.ctan.org*

[2] *Svensk Energi http://www.svenskenergi.se*

[3] *Svenska Kraftnät http://www.svk.se/*