DELFT UNIVERSITY OF TECHNOLOGY

IN4400 - Data analysis project

Rogier Slag - 1507761

October 24, 2014

1 INTRODUCTION

For this research I decided to combine several things I wanted to do for some time: combining several sources of demographic information together with the new tools Microsoft released a while back for Excel 2013 (on Windows).

The research was intended to be relatively simple. Instead it aimed to give additional insights by plotting the data in a natural and logical way. Using this approach the data makes sense to anyone and is easy and natural to interpret. One could therefore state the goal was to vouch ones intuition with statistics.

2 DATA ACQUISITION

Several sources of data were used, all from the Dutch CBS institute (www.cbs.nl). A lot of their data is publicly available from the StatLine website (statline.cbs.nl). The data is generally sorted per topic, and one can select which (parts of a) data set one is interested in.

Not all data is available as the author would have preferred. Due to the goal of plotting the data geographically, the input data needed to be sorted per city. Preferably one would have used postal codes, but these are not always available for free (buying the data sets for several thousand euros was not an option for this research). Hence the call was made to work on a city level, which was clear enough to gain general insights in the trends of the last few years.

To limit the number of years, it was decided to only include the years 2005 up to 2011. This is long enough to spot some trends, but has the nice addition we can see the effect of the worldwide financial crisis on several aspects of demographic behavior.

However over a set of years, some changes have happened in the Netherlands as well; these required some cleaning of input data. First some cities have been grouped together on a governmental level. Since the CBS uses the same mapping, some modifications needed to be done (deleting the cities in years they did not yet exists and delete old cities once they disappeared). However, since the newer cities are usually larger, they will show up more profound compared to the several smaller cities they originally consisted of.

Another factor which complicated the original acquisition of data was how the CBS introduced a new method for classification in 2009. A correction had to be done for this. Otherwise the values in the entire range would suddenly skew in the middle of 2009.

3 DATA ANALYIS

To get a clear indication of the data, we started with simple data on population count over the years of 2005 to 2011. This data was preprocessed, as described in the previous paragraph, to delete cities which were not present in certain years.

It has to be made clear only the population which had an income during the entire year is included in this analysis. From this data it is evident there has been a tendency for many people to move into the larger city boundaries, whereas people are moving away from the countryside. Meanwhile, in the countryside, we see that many smaller municipalities are grouped together in larger ones, which is why is appears to be growing a bit as well. Apart from that, the expected regional features can be seen. There is a huge spike where Amsterdam and Rotterdam are located, also the conglomerate between Rotterdam and The Hague is evident. On the countryside we can also clearly see the cities which function as a regional hub, such as Groningen, Enschede, or Maastricht.
Another clarification to the data is the bars show the relative number of people with relation to each other. Hence a conclusion can be drawn that more people are having an income in the cities compared to the countryside when we get closer to the year 2011.

When we intuitively think about this, we can see the data shows what people are seeing and telling each other. Since the Netherlands is becoming more of a service-oriented country (compared to a manufacturing country), more and more people are moving into office buildings. These buildings are often grouped together on relatively large office parks, which in turn are located in regional hubs. This explains the evident rise of the regional hubs in the countryside.
This is a bit different for larger cities: due to their larger population and better infrastructure, they are more likely to attract main or branch offices of larger companies. Since these companies require highly trained personnel, which move to such locations in order to be close to their work.
Once the economic recession hits, we see that (contrary to popular belief), the trend is still that more people are having jobs. However, the earlier growth is clearly reduced. For some

municipalities the growth even completely stagnates, but only for a few it goes down. One has to note this concerns the people with an income during the entire year, and therefore does not actually model various possibilities, such as people who only had a contract for six months when their contract got terminated. Also it is possible the people still had a job, but less income. This is what will be looked at next.

Apart from people actually having a job, it is interesting to see how much money they make on their job. CBS data only models the money people make from getting paid by an employer, and therefore does not give a complete picture. Factors such as an own company (which may hold money as well) or profits on the stock market are not modeled.
One of the firsts things we can see is how incomes in the Randstad area are relatively higher than in the remainder of the country. However, this does no further increase in specific large cities such as Rotterdam or Amsterdam. Hence it looks like the specific city has little influence compared to the general area. Obviously there are some noticeable locations where the average is higher: cities as Wassenaar, Bloemendaal, Laren and Haren clearly top their surroundings. Strangely Rozendaal is also among the top cities, with an average income of around 50000 €.
Over time, we see the average net income for most cities is still climbing, however the highest incomes take the recession the hardest. They are among the only ones which significantly see their average income decrease. The increase of average net income for everyone does slow down starting in 2008. This of course correlates nicely with the moment the economic crisis hit the Netherlands.
Mostly the incomes of people are quite closely together, strengthening the idea that the Netherlands has a highly equally distribution of wealth (in average!). However, the net income says little about the amount of money one is free to spend; for some regions, prices may be a bit higher.

Hence we decide to look at the *free income* which is defined as the income once on has paid for the required expenses of living and any surcharges have been incorporated. Once we start to look at these values, we see an even more flattened area. Hence, once we correct for regional factors, people earn on average about the same.
In this case it might be even more interesting to take a look at the free income. We see here that, although the highest incomes still decrease the most, everyone in general loses once we pass the 2008 milestone. Before that, incomes proved to be growing a bit each year, but from 2008 onwards the costs of living increased harder than the net incomes of households. Intuitively this is true, salaries and pensions have been indexed only a little (or not at all), medical insurance has become a lot more expensive and many surcharges for households were scrapped. Since the data for this research only goes from 2005 to 2011, it does not yet include factors such as the reduced mortgage deduction. It will be interesting to load the number up to 2017 in the spreadsheet once they are available.
The decrease of the most incomes can be explained by the explanation above. However, this does not explain why the highest incomes took the crisis much harder. One of the possible explanations is the "crisis tax" introduced by the Dutch government in which incomes above 150000 € payed an additional tax of 16% over the amount earned over that threshold.

Finally we combine the income and population data to verify where the *economic centres* in the Netherlands are. These are defined as location where there the multiplication of *persons with an income during the year* with *free income* is a lot higher than in surrounding areas. Once we plot this on the map of the Netherlands, we immediately see the locations we expect: Amsterdam, Rotterdam, and The Hague. However we also the the influence of the income, Rotterdam is scores almost 20% less than The Hague in this metric (due to lower incomes on average). Furthermore we see there are little of such centres in the north, Zeeland, or Limburg. On the other hand, Noord-Brabant shows quite a scala of cities which have significant economic impact (such as Eindhoven, Tilburg, and Breda).

Next item to look at is the difference between "native" Dutch people and immigrants (the term immigrants here is used in the same fashion the CBS uses it). For this we look at several parameters: education level (specifically, the number of graduates per level per city) and the ratio of where people live.
Starting with where immigrants actually live: we can that relatively the largest number lives in the four big cities of the Netherlands. A strange spot is the tip of Limburg, where another hotspot is located. Although by many people not really considered immigrants, the relatively large German and Belgian population here makes the ratio skyrocket to around 35%.
Over time we see only a few slight movements in the general picture. In the center of Limburg, the concentration seems to grow, the same goes for the Randstad area where the "open spaces" between the large cities are gradually being filled.
Next a look is taken how the different generations (first and second generation are considered) are located throughout the country. For the largest cities it seems there are much more first generation immigrants than second generation, this seems strange. No explanation for this could be found. For smaller cities this does not hold. Over time, we see most second generation immigrants are moving out of the large cities, whereas first generation immigrants seem to be moving out. However once the economic recession reaches the Netherlands, people tend to move back to those cities and the population of both grows.

Concluding a look at education level is given. Apparent from the bar graph, once can deduce their does not seem to be any correlation between education and average income. However, this is misleading. The education levels are sorted on the city one obtained his/her education, people tend to move to other cities later on. This also explains why some cities score very high on the number of WO masters; these are simply all the cities housing universities.
To get a first indication of these numbers, one can plot which degrees were obtained in which places. From this, we see more people tend to get a MBO degree over time. This is somewhat counter-intuitive, since one could expect that due the economic situation many people tend to continue their study until the job market would be better. A direct check to the CBS institute about this data was done, but this data should be accurate.
Another view is to select natives and immigrants separately and correct for the number of inhabitants. This can give an indication which group is more likely to go for a higher education. It turns out this is not easily answered (due to an issue in Powermap, see below). However, one can make an educated guess there is no significant difference (this holds for the large

cities). Smaller cities tend so skew the results quickly; Powermap is unable to correct this.

4 Data possibilities

Using the data it would also be possible to check for correlation over time between variables as *income*, *city*, and *whether one is consider to be a native*. However, due to some problems with the data (changing definitions, non-public data, and unclear methodology) this was determined to be outside the scope of the project.

The same was true for a more in-depth approach which also modeled variables such as *age*. Although this data was readily available, the enormity of the information set became to much for Powermap and Powerpivot to handle (somewhere north of 400 000 rows). For analyses with this amount of data, one should seriously consider whether Excel is still the right tool to use; the authors feeling is that the size of the current data set is already dangerously close to the limit. An example of this is the performance of drawing the graphs in on the map, when natives vs immigrants are considered, combined with different education levels, all over time.

5 Problems down the road

Using new features of Excel 2013 surely is nice to do, but the software therefore still has some bugs. As the author uses OSX, Parallels Desktop was used to virtualize Windows 7, on which Excel 2013 could be installed (unfortunately Office for Mac is not close to being on par with the Windows versions).

However using the new *powerX* features of Microsoft office (Powerpivot and Powermap most noticable) are not entirely stable yet. One problem turned out to be quite hard to debug: graphs in Powermap do not always show up when the bar graph in Powermap is used. This was the case on the retina display and office external displays, but not on TUDelft external displays. When moving Excel between the screen the visualization would simply disappear. After a lot of digging this turned out to be due to the scaling Parallels uses on high-resolution displays: it sets some scaling (so on a retina display not everything is too small in Windows). However, some Powermap features only work when the scaling is exactly 100%. Once found, the scaling options were disabled and work could continue.

Another issue is Powermap using Bing to find where to plot the data, however in some fashion this is non-deterministic. Although we use exactly the same set of municipalities, it differs to which extend they are scattered across the globe. Sometimes Powermap will locate a city in Belgium, but for the next plot that city might well be in Germany or the USA. In each of these cases, it is wrong (it should place it in the Netherlands). Hence, for a given worksheet, Powermap should be deterministic. Additionally it may use context to see what is plotted. If 96% of the cities is surely in the Netherlands, it might give precedence to the city Barneveld in the Netherlands too instead of picking the one in the USA.

Another drawback to Powermap is how it is currently impossible to model stuff according to their size, thereby making the bar graph a surface graph over the Netherlands. A city as Amsterdam generates a bar of a size x and height h, but the much smaller city next to it will also have a size x bar. In this case we would like the ground area of the bar x^2 to scale with an optional parameter, in this case the number of habitants.
This is a much larger issue for heatmaps, in which a value of 1% for a village and 40% for the big city next to it makes the area appear as 20%. Hence, zooming out means that a lot of information is lost. This visualization can therefore be misleading when not given proper thought.

Finally one can only use Powermap efficiently when combined with Powerpivot. Powerpivot is fairly strict about data types; if one of these is not correct it will give no warning, but Powermap might refuse to adequately plot the data. In that case one needs to verify for themselves where the error did possibly occur. In this fashion Excel is more of a programming tool than an office program.

6 Conclusion

In the end, it seems intuition is not wrong or stupid: there is a correct correlation between the actual data and the initial guesses made before the research. However, as it turns out there are some cases in which one assumes something based on some observations which cannot be extrapolated to a much larger scale.

Additionally the new features of Excel have proven to be worthwhile additions to the set of tools a statistician may use. The graphing uses of Powermap are very powerful and give everyone easy access to create visually appealing models of information, for which regular bar charts would be confusing or way too large. Additionally the heatmap feature is also very useful.

A downside is the same Powermap. Although powerful, it is not yet ready for the greater public to use. The downsides of Powerpivot and Powermap, combined with the occasional bugs, make the currect version not yet good enough to recommend to a friend, unless you might be looking for a way to assist them at every step in the progress.