Assignment 1

D3 Information Visualization

Joris Reijrink
Student number: 0847198
`j.reijrink@student.tue.nl`

Chris Hoedemakers
Student number: 0661115
`c.g.j.j.hoedemakers@student.tue.nl`

December 2, 2014

1 Introduction

In this report we analyze the data set `cities-data.txt`, containing a wide range of information about the different municipalities of The Netherlands. The goal of this project is to properly visualize part of this data in order to answer a couple of visualization tasks, that we will define in section 2. Once the visualization tasks are formulated, we argue for suitable visualization techniques for each of these tasks in section 3. In section 4 we finally describe our observations after implementing the techniques that were in chosen in section 3.

2 Tasks

In this section we define the visualization tasks that we would like to perform on the given data set. That is, we describe what information we would like to retrieve from the given data. However, before we define these tasks, we provide a framework (i.e., design space) for formulating visualization tasks.

2.1 Framework for Defining Tasks

In order to formulate a nice and clear visualization task, we provide a framework for doing so as presented in [7]. Here we distinguish between five different dimensions for each task, namely the *goal* (why is the task pursued?), *means* (how is the task carried out?), *characteristics* (what does the task seek?), *target* (on which part of the data is the task carried out?) and *cardinality* (on how many instances is the task carried out?). We shortly explain these five aspects in some more detail.

For the goal of the task we distinguish between three different types of analyses. We can have an exploratory analysis (or undirected search) that aims at deriving an hypothesis from an unknown data set, a confirmatory analysis (directed search) that aims to verify a found or assumed hypothesis, and finally a presentation that simple exhibits analysis results.

For describing how the task is being carried out we distinguish between three main ways of doing so. These are navigation, (re-)organization and relation. Navigation subsumes all means that change the extent or the granularity of the shown data, but that do not reorganize the data itself. (Re-)organization includes all means that actually adjust the data to be shown by either reducing or enriching it, and relation encompasses all means that put data in context.

Characteristics of the data can be subdivided in low-level data characteristics and high-level ones. Examples of low-level characteristics are data values of a particular data object or the data objects corresponding to a particular data value. Examples of high-level characteristics include more complex data patterns such as trends, outliers, clusters, frequency, distribution, correlation, etc.

The target of the visualization task describes on which part of the data the task is being carried out. This often boils down to a certain (sub)set of the different data attributes.

The cardinality finally specifies which part of the data instances we consider when carrying out the task. For instance just a single one, a certain subset, or just all of them.

2.2 Chosen Tasks

With this common framework for formulating visualization tasks defined, we move on to defining the actual tasks we want to perform on our data set about the Dutch municipalities. We mainly distinguish between two different tasks, of which the first one will be divided into two subtasks.

2.2.1 Task 1

The first task that aims at discovering interesting properties of the given data set, as well as confirming expected properties. This means that this task is rather broad, and numerous specific questions can be formulated from it. However, we will define two basic (sub)tasks for this task of analyzing the entire data set. These two subtasks are defined as *a)'analyze the relations between the different attributes in the set'* , and *b) 'compare the attribute values of the different municipalities with each other'*. The subset of attributes on which we will perform both of these tasks will be the same. This subset will consist of the attributes `OAD`, `STED`, `AANT_INW`, `BEV_DICHTH`, `AANTAL_HH`, `P_EENP_HH`, and `OPP_TOT`. For convenience we will call this set of attributes S_{attr}.

When formally defining task 1a, we observe that we adopt an *exploratory* approach, as we aim at *searching* interesting *relations* between different data attributes. From this we can (partly) derive the first three dimensions for our 5-tuple task description. The forth dimensions, the target, has also been specified already, namely by S_{attr}. This leaves us with the cardinality. Because we want to discover general relations, we want to look at all data instances (municipalities). However, some relations between attributes may be very obvious. We will therefore also be looking for *confirmation* of these obvious relations, and want to detect possible abnormalities or inconsistencies in them, such as *outliers.* The resulting formal task description then looks like this:

(exploratory|confirmatory, relation-seeking, relations|outliers, S_{attr}, all)

For task 1b we can define a somewhat similar task. The main difference is now that we will not be looking at the relations between the different data attributes, but are *comparing* the data values of the different data objects. In this way we again hope to find interesting characteristics such as *outliers* or *clusters.* As with task 1a, we take S_{attr} as our target and all instances as cardinality.

(exploratory|confirmatory, comparison, outliers|clusters, S_{attr}, all)

Specific questions that we could pose for the first of these two subtasks is for instance 'is there a positive relation between the attributes `AANT_INW` and `BEV_DICHTH`?'. If we think about this for a minute in advance, we may predict that this will indeed be the case. This would hence be a typical question where we try to confirm our hypothesis, and look for data objects that do not adhere to this. A much more general and exploratorive question would be 'which attributes have a typical positive or negative relation with each other?'. Again we may define some expected relations, such as a negative relation between `OAD` and `STED`. However, this question really aims at an elaborate (though high level) analysis of the data.

For task 1b we can also define a question of which we may already predict the answer. We could for instance try to confirm that the four largest cities in The Netherlands score highest on the `AANT_INW` attribute. We may even expect these values to be rather significantly larger than those of the other municipalities. Finally we can again pose a general, explorative question in the form of 'how are the attribute values of the municipalities distributed for a certain attribute, and can we explain this distribution?'.

2.2.2 Task 2

The second task we define concerns the topic of a high degree of aging population ('*vergrijzing*' in Dutch). The task is somewhat twofold, because we want to confirm that there are some municipalities in The Netherlands that suffer from a high degree of aging, as well as find out which municipalities these are. The degree of aging is defined by the percentage of inhabitants that are older than 65 years, compared to the percentage of working inhabitants. Because we do not have the exact percentage of working inhabitants of each population, we assume here that this percentage is somewhat similar to that of the percentage inhabitants aged between 20 and 65 years.

When formally defining this task we observe that we both want confirm an hypothesis as well as explore the data (explained above). Hence we will be searching for, and localizing the municipalities that suffer the most from a high degree of aging. The set of attributes that we will consider for this task consist of `P_00_14_JR`, `P_15_19_JR`, `P_20_24_JR`, `P_25_44_JR`, `P_45_64_JR` and `P_65_EO_JR`. This set will from here on be denoted by T_{attr}. Similar as for task 1, we again want to consider all of the municipalities for this task. Together this results in the following formal definition of task 2:

(confirmatory|exploratory, searching|localization, T_{attr}, all)

Though the specific question for this task was already mentioned earlier ('Which municipalities in The Netherlands suffer from a high degree of aging?'), we may also ask ourselves 'Is there a certain trend in the municipalities that suffer from a high degree of aging?'.

3 Techniques

3.1 Techniques Task 1

Because the tasks 1a and 1b are similar in the sense that they aim at observing quite a lot of different data attributes on a rather high level, we will argue for the use of one general technique for both of these tasks. Techniques that can be chosen for this broad exploration of the data include a Parallel Coordinate Plot (PCP) and Scatterplot Matrix (SM). This is because these two techniques enable us to show a lot of different attributes in one large overview. An advantage of the PCP over the SM is however that the PCP tends to take up less space. Also the axes of the SM tend to become very small when numerous attributes are used. This makes it hard to distinguish between single data objects when analyzing the data. When data objects can be be categorized, this problem can be partly overcome by specifying each category with a certain color. However, for our municipalities data set this categorization is not really possible.

Other visualization techniques such as bar and pie charts, were not considered suitable for these broad overview tasks because they can not really create a nice overview. Different types of hierarchical or network visualizations were quickly disregarded as the data does not posses such structure. Even maps were not deemed very useful, because they make it hard to compare different attribute values with each other. For these reasons, the PCP and SM techniques were chosen to be implemented [5][4].

3.2 Techniques Task 2

Unlike the first task, task 2 only has a few attributes that, when combined, exactly add up to one hundred percent for each municipality. Hence, complex visualizations are not necessarily needed, and simple visualization may suffice. The biggest challenge in finding a good visualization is managing the great amount of municipalities. A pie chart or sunburst are therefore not very suitable, because they will be very hard to read. As mentioned before, because the data has no real hierarchical structure, a sunburst would have at most two levels.

A normalized bar chart will be a much more suitable technique to visualize the data [3]. The proportion of aging inhabitants for each municipality will be visible at once, giving a good overview of the total aging population. By also sorting the data on the 65+-percentage, we can easily find the municipality with the highest degree of aging population. Using a tooltip [6], the exact percentage of each age group can even be shown per municipality. A regular bar chart will result in the same visualization as the normalized bar chart, because the attributes are percentages (always adding up to a hundred). We can however use the population count per municipality to show the actual number of aging people (65+). With this visualization the absolute values are show for each municipality. If the data is sorted on the 65+-percentage, it will for instance give insight in whether the aging percentage is higher in bigger cities. With this bar chart it will on the other hand be hard to get an overview of the total aging population. We will therefore implement a combination of both bar charts, giving us the best of both worlds [1].

The final visualization technique we will implement for performing task 2 is a choropleth map. If the aging population of a municipality is high, a dark shade of a color is used, if it is low, a light shade is used. This will give a quick insight in which municipalities have a high degree of aging population. An extra bar chart [2] may be added when hovering over a municipality, giving more detailed information about that municipality. This bar chart will show the percentage of inhabitants for each 5-year age group. We choose to this in order to investigate whether this is additional information may be useful to the user.

4 Observations

After implementing the different techniques for the tasks posed in section 2, we will now describe our observations from the analysis of these different techniques. We will elaborate on the information that was retrieved from the different techniques, list their pros and cons and come up with possible improvements for the used techniques.

4.1 Observations Task 1

For task 1 (as described in section 2.2.1) we implemented a Parallel Coordinate Plot (PCP) and a Scatterplot Matrix (SM). We will first argue how both techniques performed for the subtask 1a 'analyze the relations between the different attributes'. When looking at the PCP, we first try to confirm the first question we posed in section 2.2.1: 'is there a positive relation between the attributes `AANT_INW` and `BEV_DICHTH`?'. When the axes of these two attributes are placed next to each other, it pretty quickly becomes apparent that this question can be answered positively. However, we also see that municipalities like Amsterdam and Rotterdam do not adhere to this. Apparently these municipalities cover such a large area that their vast number of inhabitants still not get them in the top of the `BEV_DICHTH` scale. Municipalities like Leiden, 's-Gravenhage and Haarlem on the other hand have much less inhabitants, but score much higher on the `BEV_DICHTH` scale because of their small surface area. If we then compare the attributes `OPP_TOT` and `BEV_DICHTH`, we can see there is a somewhat stronger (negative) relation between these two attributes.

In the SM it is much harder to see the positive relation between the attributes `AANT_INW` and `BEV_DICHTH`, and the negative relation between `OPP_TOT` and `BEV_DICHTH`. We can however quickly detect that there are some outliers in both of these relations. However, in the SM it is unfortunately not possible to retrieve which municipalities these outliers actually are by hovering over the dots. It was tried to implement this feature (similar to the tooltip hover in the PCP), but because the cursor already allows for brushing in the SM, it was not accomplished to also enable the tooltip feature. Being able to implement this tooltip feature could be regarded as a large improvement of the SM. On the other hand, when dots lie very close to each other, it may become hard to select a specific one.

When we return our attention to the PCP and want to compare other attributes with each other, we notice that quite some attributes have a few outliers that greatly enlarge the scale of the axis. These attributes include `OAD`, `AANTAL_INW`, and `AANTAL_HH`. Because of this, the majority of the lines cross the axis in a rather small area, making it very hard to distinguish between them. In its turn, this makes it harder to detect relations between attributes of the more 'general' municipalities. In order to solve this problem, a new feature is proposed. When a certain range of some attribute is selected, it may be convenient to stretch this range over the entire axis. In this way, more space is available to distribute the different lines, making it easier to distinguish between them. We should however ensure that the user is clearly notified of the fact that some data objects (municipalities) are left out in that specific view. The SM on its turn suffers from a similar usability problem. Because the plots are rather small, it is also hard to distinguish between different data objects. Here we propose to implement a zoom feature. Clicking on one of the plots would enlarge it to the size of the entire matrix, making it much more readable. However, as the implementation of both these proposed features was considered rather time-consuming, it was chosen not to do so.

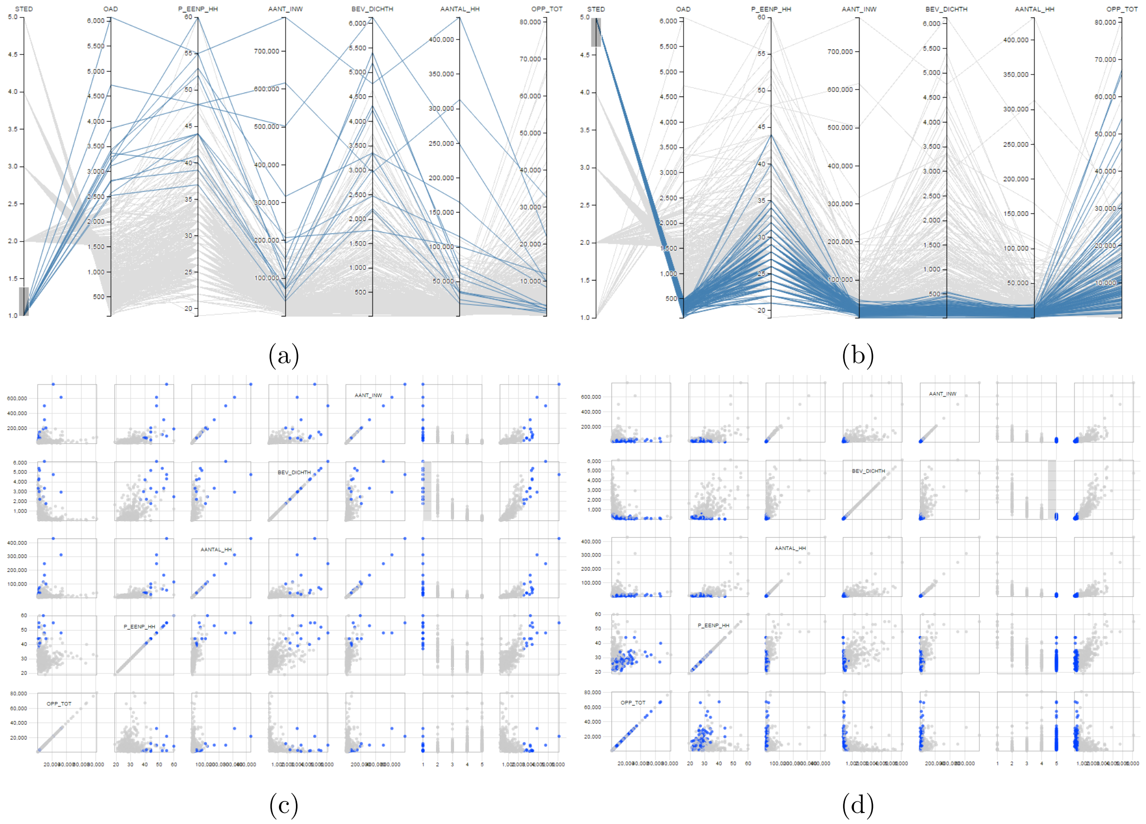
Another observation that we can make when considering the PCP and SM for the task of comparing different attributes with each other, is that the SM makes it slightly easier to compare multiple attributes at once. Because all pairs of attributes are shown in one (or actually two) of the plots, we can quickly switch between comparing two attributes. When using the PCP, we need to shuffle around with the different axes and place the axes of two attributes next to each other in order to compare them.

Let us conclude the observations of subtask 1a with mentioning that performing this high-level task on the attribute set S_{attr}, mainly brings forward some rather obvious relations, or no clear relation at all. Examples of obvious relations are the positive one between `AANTAL_INW` and `AANTAL_HH`, and the negative relation between `OAD` and `STED`. A clearly unrelated pair of attributes would be `BEV_DICHTH` and `P_EENP_HH`. Choosing different attributes might bring forward more interesting relations. However, with a lot of different attributes it might be hard to find those pairs that are truly valuable.

Next we continue to the second subtask: 'compare the attribute values of the different municipalities with each other'. Let us start with again shortly addressing the fact that because of the few significant outliers, the scale of some axes is stretched quite a lot, making it hard to distinguish between different data objects. This emphasizes that the solutions proposed earlier may be very valuable.

To answer to the specific question posed in section 2.2.1, we can easily confirm from both the PCP and the SM that the four largest cities in the Netherlands score significantly highest on the `AANT_INW` scale. Regarding the general, explorative question concerning interesting attribute distributions, we turn our attention to the `P_EENP_HH` attribute (i.e., municipalities with the highest percentage of single households). Looking at the top 8 ($> 50\%$), we see that these spots are taken by the municipalities Wageningen, Groningen, Amsterdam, Delft, Utrecht, Nijmegen, Leiden, and Maastricht in descending order. Shortly considering this list, we may find it quite logical, as this list represents the top student cities in The Netherlands!

When comparing the values of a cluster of objects, there is one observation that can me made from both techniques. This observation can be made when comparing the cluster of the municipalities with a STED value of 1, with those with a STED value of 5. As becomes apparent from Figure 1), the municipalities with a low STED value (i.e. the more urban municipalities), have much more fluctuating values for the other attributes. Municipalities with a STED value of 5 have values for the other attributes that lie much closer to each other. This again shows that there is a handful of municipalities that significantly influence the scaling.



(a) (b)

(c) (d)

Figure 1: Parallel Coordinate Plots highlighting the municipalities with a STED value of 1 (a), and a STED value of 5 (b), as well as the Scatterplot Matrix highlighting the municipalities with a STED value of 1 (c), and a STED value of 5 (d)

Summarizing, we may conclude that both techniques are rather effective in answering the two subtasks. If we were to chose one of them, there may be slight preference for the PCP. However, when all the proposed improvements were to be implemented, the SM might perform just as well, if not better. In general, we can conclude that this broad and high-level exploratory task may remain rather difficult with any visualization technique.

4.2 Observations Task 2

For this task we implemented a bar chart and choropleth map visualization. The bar chart visualization contains a combination of a regular and normalized bar chart. The choropleth map also contains an extra bar chart to give detailed information about the age groups of a certain municipality.

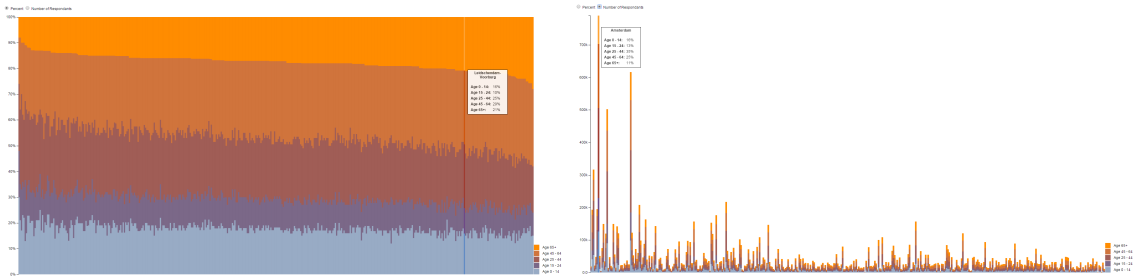
The specific question that these visualisations should confirm is whether there are municipalities in the Netherlands that suffer from a high degree of an aging population, and which ones do so the most. Both visualizations achieve this task in their own way. The bart chart shows that the degree of aging population in some municipalities is more than half the size of the working group, which is quite a lot. It also shows that the degree of aging population is smaller in large municipalities. An example of this is Amsterdam, which is the seventh municipality when sorting the data on the percentage of degree of aging, in ascending order (see figure 2).

The choropleth map does not directly show that there are municipalities in the Netherlands that suffer from a high degree of an aging population. This is because the color scale of the map is determined by the highest and lowest value of the degree of aging. However, when hovering over municipalities the degree of aging is shown using a tooltip. Next to this, a bar chart shown, giving a good insight about the aging degree of the municipality. Because the bar chart has a bar for each 5-year age group, it shows a detailed distribution of the age groups, which gives extra information for analysis. The downside of these 5-year bars is that the user has to combine the bars mentally in order to 'view' the 65+ group and working group. Then again, the colormap already takes care of indicating the height of the degree of aging, and the tooltip shows the 65+ percentage of the selected municipality. Even thought the map itself may not give direct insight in the aging degree, it does provide geographic information about the aging population. The map for instance shows that the north and south of the Netherlands suffer from a higher degree of aging population than the center. Another interesting observation are the few low-degree outliers. Investigating these municipalities more closely, can again find that these are the municipalities that contain a large student city. The most notable example of this is Groningen (see figure 3).

When looking at the choropleth map, there are a few improvements and design decisions that we should shortly address. A clear improvement would for instance be adding a color scale to the map, showing the range of values that the colors in the map represent. Currently the bar chart or tooltip is required to get an idea about the value of a certain color. This color scale would even answer the question of task 2 without any need of the bar chart, because the percentage of aging population is immediately visible to the user. Furthermore we could argue that changing the hover function by a mouse click function would be an improvement. With the current hover function it is hard to compare two specific municipalities, because the moving from one municipality to another one, triggers the map to show the bar charts of all intermediate municipalities. This creates a chaotic view, and could be prevented by using a mouse click selection, rather than the hover selection. Finally we can argue about the color palette of the map. In the current choropleth map the color scale is green-red. By using these colors, the map also emphasizes municipalities with a low degree of aging population, like student municipalities. Though this is interesting information, it is not necessary to perform the specified task. Instead of the green-red color palette we could use a white-red color scale. This gives a more organized and calm impression, as shown in figure 3.

The bar chart on the other hand is more straightforward. However, a feature that would be nice to add is an absolute sort-on-population option. The bar chart is currently sorted on the percentage of 65+ population per municipality. If the bar chart could be sorted on the absolute value of 65+ inhabitants this might provide valuable additional information.

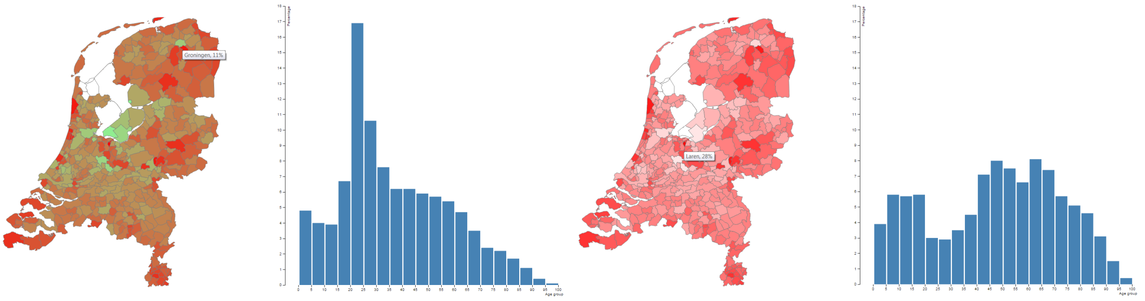
When comparing both visualizations, we may conclude that the bar chart is a better visualization technique for the task that was formulated. The user can immediately answer the question because of the sorting ption, which is not possible using the choropleth map. However, if the choropleth map were to be updated with the proposed improvements, it would also be a very suitable technique for performing task 2.



(a)

(b)

Figure 2: Bar chart with the age group percentages of the municipalities (a), and the absolute population values (b)



(a)

(b)

Figure 3: Choropleth map that shows the degree of aging population using a green-red color scale (a), and the same map using a white-red color scale (b)

References

[1] Thomas Apodaca. Stacked to normalized stacked bar chart, 2013.

[2] Mike Bostoc. Bar chart, 2012.

[3] Mike Bostoc. Normalized stacked bar chart, 2012.

[4] Mike Bostock. Scatterplot matrix brushing, 2012.

[5] Jason Davies. Parallel coordinates, 2011.

[6] Vinicius G Rocha. d3-bilevellabeltooltip, 2014.

[7] Hans-Jorg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. A design space of visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2366–2375, 2013.