Project Report - Energy Explorer

TNM048 Information Visualization 2009

Fredrik Rinman
820429 - 0277
freri429

Petter Grundström
?????? - ????
petgr138

Tobias Lundin
?????? - ????
toblu567

Table of contents

Background and problem definition

This report is the final result of a small project conducted in the course Information Visualization (TNM048). The aim of the project was to choose a topic of interest, gather appropriate data and build an information visualization application with which the data could be explored and analyzed, with a goal of gaining insight and new knowledge. The software was developed using the GeoAnalytics Visualization (GAV) framework, which is a framework that supplies a wide variety of information visualization tools.

The topic that was chosen was the energy use and consumption around the world, and how it affects human society and vice versa. The energy situation of the world is a burning issue, not just for our own way of life but for the longevity of the human race. In the last years the debate has grown bigger and come to not only be about energy efficient cars or better emission filters for factories but also about our way of life. An interesting question to ask is what would happen to the climate or energy consumption of the world if all countries had the same living conditions as Sweden. Is it really a fair assumption to make that a developed country is more energy efficient than a developing country? How does luxury really effect the energy consumption? Another question is what will happen as our reserves of fossil fuel starts to run out?

With this background we decided to look at the energy consumption of a country and measure it against a range welfare variables. We chose to make an exploratory analysis to begin with, as we did not know what possible welfare variables could correlate to the energy consumption.

The data

The data used in this application was taken from the public databases provided by the United Nations and the World Bank. Look in the appendix A for a complete and detailed list of the databases used.

As our approach was one of exploratory nature we tried to include as wide a spectrum of welfare variables as we could without them becoming to many. The welfare variables we chose, that are listed below, fall into a few main categories we thought were important to cover. These were IT, level of health care, finance, work force and population.

- Employment-to-population ratio, percentage
- Infant mortality rate (0-1 year) per 1,000 live births
- Personal computers per 100 population
- Internet users per 100 population
- Exports of goods and services (% of GDP)
- Imports of goods and services (% of GDP)
- GDP per capita, current international dollars
- GDP growth (annual %)
- Percentage urban (%)
- Population, total

The following energy variables were chosen to measure the use, consumption and environment effect.

- Energy use (kg of oil equivalent per capita)
- Electric power consumption (kWh per capita)
- Greenhouse Gas (GHGs) Emissions without LULUCF, in Gigagrams (Gg) CO2 equivalent

The data was downloaded in comma separated variable form, also known as *csv*, with all the available years and countries from the source database. The data was then processed, a selection of years and countries were made and then transformed via simple scripts into an excel sheet. Thanks to this repeatable transformation process it will be an easy task to change the data input into the application, even if it means to add or recalculate data.

The final data sheet used data from all available countries, as long as they had an unique ISO 3166-1 country code, and the time segment 1990 to 2005.

Choice of visualization methods

It was early on decided that the application would be based on the GAV framework as it provides many of the standard visualization methods in an easy to use fashion. In order to decide which components to use inside the framework we reasoned about the most effective use of the components and the impact it could have on our data. As we chose to conduct an exploratory analysis we wanted our software to be of a general nature, e.g. we wanted it to show as much information as possible and we also wanted to be able to change our variables should we realize we had chosen the wrong ones. Below is a short explanation of all the visualization components we chose to use.

Choropleth Map

The choropleth map is an interactive map, often connected to a color map. We chose to use it to get fast visual feedback in a geographical context, that also increases memorability. The choropleth map is also an effective component for selecting a specific country as opposed to selecting the correct tuple in a large dataset.

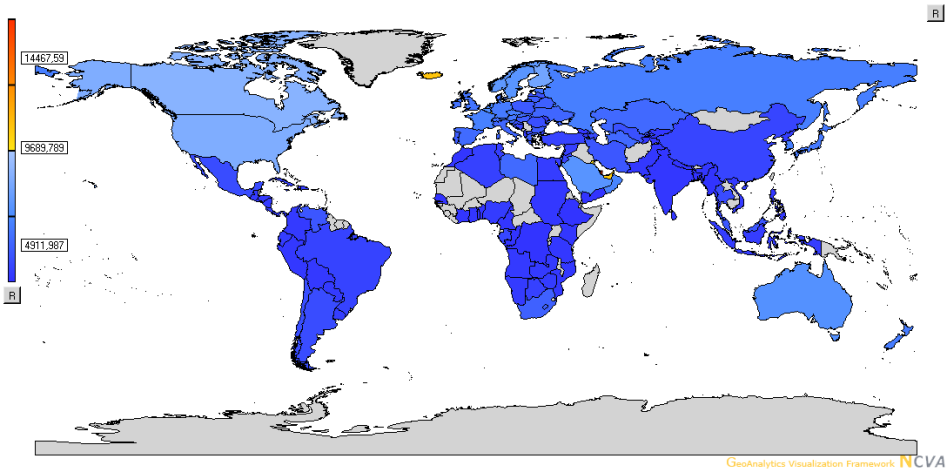

Illustration 1: Choropleth map, as used in the application

Scatter Plot

The scatter plot shows data as points in a coordinate system (we chose to use a two dimensional grid), where the user can choose what data to display at each axis. Relative positions of data is among the best methods to put quantitative data into relation. Since this method is used in the scatter plot we chose to use it to be able to easily get an overview of how well two variables correlate to each other and within the variable itself. We also made it possible to to visualize a third variable as the size of the point.

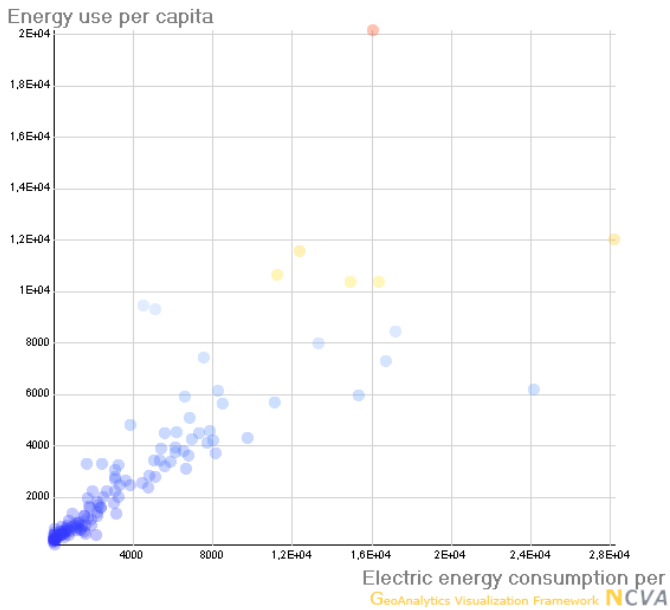

Illustration 2: Scatter plot, as used in the application

Parallel coordinates plots

The parallel coordinates plot is a visual component that shows whole data tuples in two dimensions. The columns in the dataset are displayed on vertical axises with a scale and then a line is drawn between the axises to be able to see which values belong to which tuple. The axises also have two sliders to enable easy filtering in the form of threshold values for the column. As a parallel coordinates plot is one of the best methods of filtering a data set for information visualization it was

chosen for this purpose. It also has an element of displaying correlations between variables, but that was a minor concern. We chose to use two parallel coordinates plots. One for visualizing all the data for the chosen year and one for visualizing the data from all the years of the chosen variable. The downside of parallel coordinates plots is the visual clutter created by many tuples. To work around this problem we implemented a zooming function for each column to get more information out of otherwise cluttered areas of the plot. We also enabled the use of K-means clustering in the parallel coordinates plots to be able to get a quick look at possible correlations.

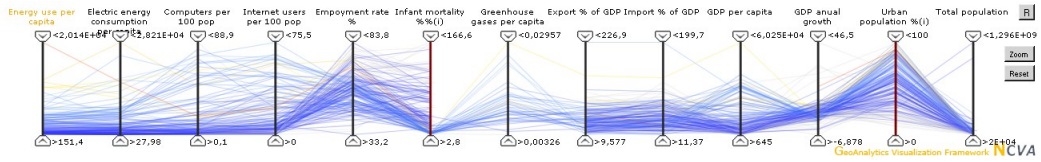

Illustration 3: Normal parallel coordinates plot, as used in the application

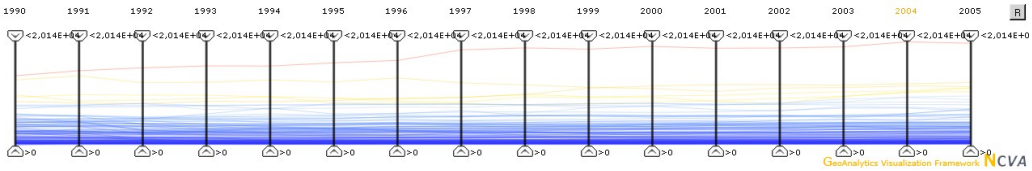

Illustration 4: Year parallel coordinates plot

Table lens

The table lens shows data graphically in a table where each column consists of a bar chart over one variable. Each column can be sorted in both an ascending or descending order. This functionality makes the table lens good for finding correlations between specific variables. We chose to use it to be able to easily sort each variable and find correlations.

It is worth to note that all of the visual components were connected to a color map, a selection and picking manager as well as a visibility manager. This functionality makes it easier to connect the data between the different components as the same color map, selected countries and visualized data is shown in all components.

To further work around the problem of cluttering, which became a problem in the scatter plot and not only for the PC plot, we made it possible to change the opacity of the lines and points in the PC and scatter plot.

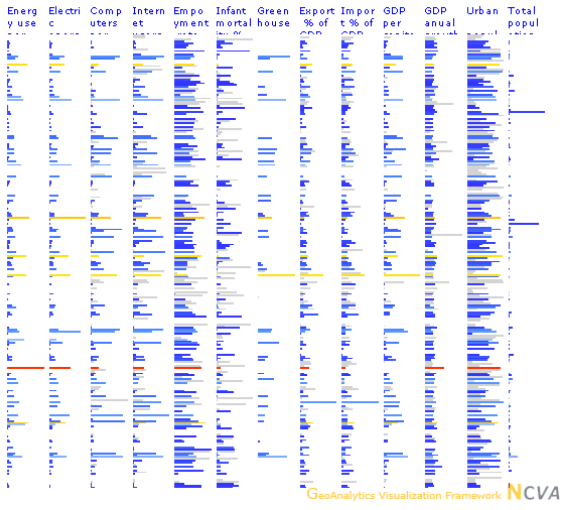

Illustration 5: Table lens plot, as used in the application

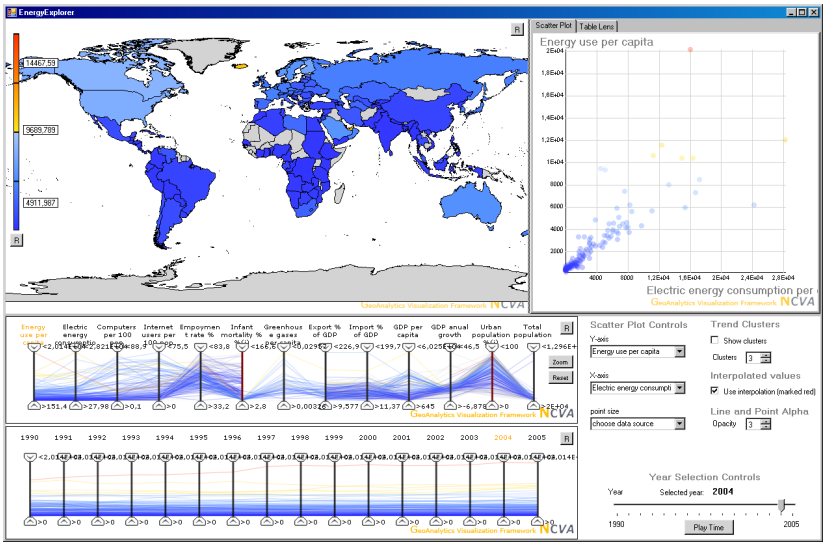

Illustration 6: Overview of the the application

Exploring the data

If not otherwise stated the chosen year is 2005.

When looking at the energy use and electric energy consumption around the world at 2005 two groups can be identified. The first group consists of USA, Canada, Luxembourg, the Nordic countries and the Arabic countries west and south of the Persian golf, which have a high value for either or both energy variables. The second group consists of the rest of the world. Within the first group there are outliers such as, Iceland, Norway and Qatar which are among the few countries to have a strong inverse correlation between energy use and electric energy consumption. The other countries within the first group seem to follow a more direct correlation between the two energy variables.
The second group of the countries has a more obvious correlation between electric energy consumption and energy use than the first group. The energy consumption of the second group is also significantly lower than in the first group.

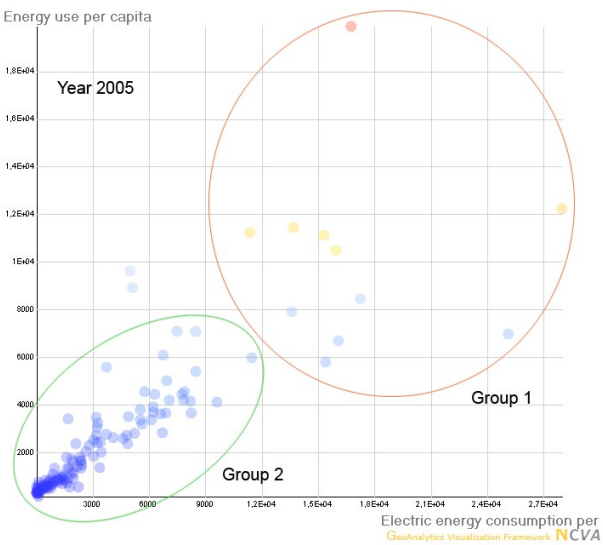

Illustration 7: Energy use vs electric energy consumption, 2005

In comparison there were three distinct groups in 1990, the first one being almost the same as the first group mentioned above with the exception of Luxembourg. This group shows about the same correlations as the first group of 2005.
The second group consists of the rest of the European countries (both west and east), the former Soviet Union, Australia, New Zealand, Japan, South Korea and South Africa. The third group consists of all the countries of South America, the rest of Africa, the rest of the Middle East and Indo-China.
The second and third group also has about the same correlations as the second group of 2005 but a distinctive gap can be seen between them. This gap shows that the second group has a higher energy consumption than group three.

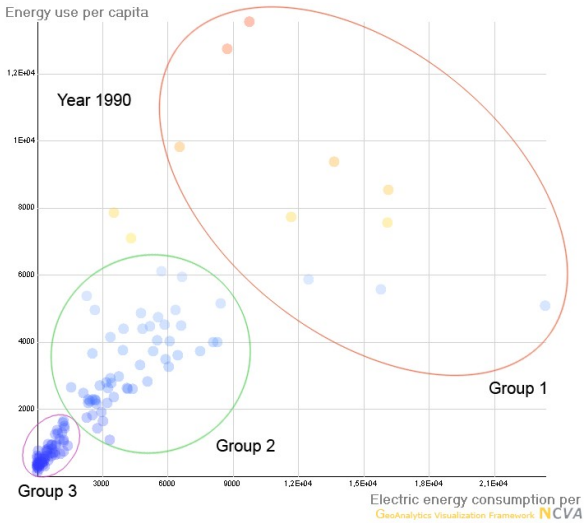

Illustration 8: Energy use vs electric energy consumption, 1990

Another interesting fact is that the correlation between the two energy variables shows a shift in favor for electric energy consumption over the span of 1990-2005.

When comparing the energy consumption variables to the rest of the dataset the correlations become more interesting. A strong correlation between both energy variables can be found with GDP per capita and urban population percentage. An inverse correlation for the energy consumption is also visible when put in relation to the infant mortality rate. For the number of computers per 100 population and Internet users per 100 population there is also a rather strong correlation with both energy variables, but the correlation for the electric energy is somewhat stronger than for the energy use.
The number of data points for the green house gas emissions are to few to give a clear picture of possible correlations. However, a visible trend that possibly could be a correlation can be seen. There are four countries that greatly diverge from this trend. Sweden, Norway, Iceland and Finland. These countries have significantly higher energy consumptions, especially electric energy, than countries with the same amount of greenhouse gas emissions.

The employment rate, the export and import as percentage of GDP and GDP growth seems to have a very weak correlation with the energy consumption.

When looking at the energy use and electric energy consumption over the years from 1990 to 2005 it is not difficult to see an overall trend towards more energy being consumed, but the pattern is not shared by all countries. Some countries have increased their electric energy consumption with over

130% or more, like Angola. Other countries, have decreased their energy use. Sweden for example has decreased the energy use by 3%.

While the energy use has increased somewhat over the period from 1990 to 2004 the number of computers and number of Internet users per 100 population has increased drastically. The number of computers per 100 population has increased more than 500% for most developed countries since 1990. Many countries that by 2004 had below 40 computers per 100 population have had an even bigger increase since 1990 than that. Brazil, for example, had an increase with over 3500% and by 2004 had the same amount as Sweden had 1990.

Two other clear trends over the 15 years from 1990 to 2005 is that the infant mortality rate is falling and that the GDP per capita is rising.

Usability evaluation

To find any usability problems areas in the design of our software we performed a user evaluation with three persons followed by an expert review, performed by us using lists of usability heuristics.

User evalutation

A user evaluation was performed, mainly aiming at testing effectiveness, learnability and satisfaction. No element of efficiency was included in the test.

We thought it was important to perform the evaluations both with users that had previous experience of working with GAV-components and users with no previous experience. Two of the users were participants of the Information Visualization course and one user a student with no previous experience of the theory or software associated with information visualization.

The evaluation consisted of two parts. The first part was an exploratory evaluation where the user got to use the software to answer questions we posed. The second part consisted of questions regarding quality of the software and what general experience the user got from using the software.

The conclusions we came to from the first part of the evaluation mainly consisted of design issues both of a general and specific nature.

Even though we had a parallel coordinates plot that showed the development of a variable over time, none of the users used it when asked to find out how a specific variable developed over time. Instead they used the choropleth map and changed the year to see the development. This suggests that the users did not see or understand the parallel coordinates plot. From this we draw the conclusion that the user interface must be designed so all the components get equal amounts of attention. We did not have a label or a tooltip explaining what the parallel coordinates plot should be used for. This is a suggestion that might have changed the outcome from the evaluation.

The selection of a specific tuple in the dataset was performed with the selection and picking manager that is built into GAV. During the evaluation all participants had a hard time selecting the correct tuple from the scatter plot, table lens and parallel coordinates plots. This problem was mainly caused by the cluttering that occurs when you visualize a large dataset in the mentioned components. Another problem that came up was that the selected tuple sometimes was deselected in a rather inconsistent way. For example, if you selected a country in the choropleth map and then zoomed in the map the country could be deselected. These problems suggests that more care has to

be taken to visualizing the selected tuples. For example in the parallel coordinates plots the selected tuple could have been a thicker line that was always placed in the front for easier detection.

A general conclusion we came to was that many of the components we used had a low learnability. That is, users with little or no previous experience of using the components had a hard time figuring out how to actually use them. All three of the users asked questions about how the components worked and they seemed lost. The conclusion we drew from this was that help on demand is a very important part of the software, especially with custom designed components. This could be performed with either help functions (such as a help menu or help buttons in the user interface) or tooltips with explaining texts.

Expert review

The expert review was performed using both a general usability heuristics list and a more specific information visualization heuristics list. The heuristics that where used where Nielsens general usability heuristics from 1994 and Shneiderman information visualization heuristics from 1996.

We found quite a few problem areas, many of whom should be considered major and even a few who are to be considered catastrophic. The issue were graded according to the scale, which was four point scale from cosmetic to catastrophic, with an end user that was semi experienced and a context that did not involve the application falling into disuse after this course is finished.

Many of the problems we found, from a usability stand point, are problems with learnability, documentation and help, issues that are not big problems for us while using the application for exploring the data within the context of the course. These issue would however need to be fixed should the application be released to other users.

There where though a fair number of issues that concerned the effectiveness and efficiency of the application. These issues would have to be fixed in a future iteration of the application even if the users would only be ourself context would remain the data exploration for the course. The most serious problem we found was that the scatter plot rescales the axises as you change the selected year, which makes comparison within the plot all but impossible. Heavier use of percentiles and percentage comparison would also have help in the data analysis, along with being able to save the application state. More, but also more well chosen, details on demand would also have been helpful.

Conclusions

One of the first things we realized when we asked what correlations exists between the energy use and the variables affecting it, is that it is a very complex question. The modern society is highly dependent on energy in all its forms. This means that the number of variables affecting and being affected by the energy consumption is vast. We quickly realized that we would only be able to scratch the surface of this question, and only find general clues how energy is being consumed.

What our data shows us is that we are moving towards a more homogeneous world where the available energy is more equally spread between the entire human population. This does not only apply to energy use and consumption, but as noted above the same trend can be seen with infant mortality, GDP per capita and the use of IT.

It is also rather clear that a modern society like the one we have in Sweden with a high level of welfare, high availability of IT and high GDP use significantly more energy than many of the

underdeveloped countries. The global trend seems to be that countries like Sweden that have high values on the welfare variables also have high values of energy use. However, there are also some countries, like South Korea, that have a fairly low energy consumption but have comparable, in some cases better, values on the welfare variables than countries like Sweden.

There is a strong correlation between the number of computers and Internet users per 100 population and the energy variables. However, the number of of computers and Internet users has increased drastically compared to the increase of energy use, which suggest that energy is being used less by some variable. To be able to draw any conclusions about this it would be valuable to have the energy consumption data divided into different categories to see how the energy is actually consumed.

We also noted that a high value GDP per capita seems to require a high value of energy per capita, but the annual GDP growth rate does not seem to be affected. This is perhaps not surprising but still interesting. This suggests that the energy efficiency does in fact not increase with a higher value of GDP per capita.

The strongest correlation with the energy use is the negative correlation with the infant mortality rate. Norway, for example, has a very low value of infant mortality and a very high value of energy use, while Angola, Congo and Mozambique has the reverse values of these variables. Looking at the greenhouse gas emissions there is an apparent correlation to energy use. This is, however, not true for the Nordic countries (Sweden, Finland, Norway, Iceland) which have a significantly lower amount of greenhouse gas emissions compared to countries with the same level of energy usage. It would be interesting to investigate how these countries have managed to keep the gas emissions so low, but it is most likely due to low dependency of fossil fuel and high usage of water, nuclear and geothermal power.

Further Work

Our approach to the project was of an exploratory nature, that is, our initial questions were of a general nature and we intended to find out more about the data before posing more specific questions. Because of this approach, and even more so as it became apparent that the initial question was broader than we first had anticipated, our answers are also of a general nature and more work is required in order to find more precise answers.

If a further study would be conducted the following should be taken into consideration.

The dataset should be expanded with more precise data that divides the energy production and use into categories. That way it would, for example, hopefully be possible to see if there is indeed a correlation between the increase in electric energy use and the greatly increased number of computers. The dataset could also be reduced with variables that does not correlate to the energy consumption and replacing it with data similar to the data that correlates, as this should be more interesting for an in-depth study.

Components to ease the identification of correlations would be a good addition in future version of the application. An overview map, as a complement to the normal map, would also be good.

Should any of the visual components be removed from the user interface? Some of the components did not serve the function we intended as the user evaluation (see Appendix B User Evaluation) suggested, they could be removed from the user interface to make it less cluttered or to leave room for a more useful component. Perhaps the parallel coordinates plot showing years should be

replaced by another method for showing data over time.

The feedback from the usability studies also gives a fair number of problem areas that should be addressed would any further studies be conducted.

There are a large number of bugs in the current version of the software that needs to be fixed in future versions. Many of these bugs are found when using the PC plot zoom function, which while it may have been a good idea is way to unstable and confusing in it's current form.

Appendix

Appendix A - Data sources

UN Data

http://data.un.org/

- Employment-to-population ratio, both sexes, percentage (*Millennium Development Goals Database (1.5)*)
- Greenhouse Gas (GHGs) Emissions without Land Use, Land-Use Change and Forestry (LULUCF), in Gigagrams (Gg) CO2 equivalent (*Greenhouse Gas Inventory Data*)
- Infant mortality rate (0-1 year) per 1,000 live births (UNICEF estimates/MDG) (*Key Global Indicators*)
- Percentage urban (%) (*World Population Prospects: The 2006 Revision*)
- Personal computers per 100 population (ITU estimates/MDG) (*Key Global Indicators -> International Telecommunications Union (ITU)*)
- Internet users per 100 population (ITU estimates/SYB51) (*Key Global Indicators -> International Telecommunications Union (ITU)*)
- GDP per capita, current international dollars (PPPs) (WB estimates) (*Key Global Indicators -> World Bank, World Development Indicators (WDI)*)

World Bank

http://ddp-ext.worldbank.org/ext/DDPQQ/member.do?method=getMembers&userid=1&queryId=135

- Electric power consumption (kWh per capita)
- Energy use (kg of oil equivalent per capita)
- Exports of goods and services (% of GDP)
- Imports of goods and services (% of GDP)
- GDP growth (annual %)
- Population, total

Appendix B - User evaluation

Tasks

1. Välj ett antal valfria länder i kartvyn.
2. Sortera länderna på "energy consumption" och se hur förhållandet mellan dem ser ut.
3. Använd "electric energy consumption" och ändra färgskalan så att bara de tre länderna med högst konsumtion ligger i den orange-röda delen av skalan.
4. Välj valfri variabel och titta på hur den variabeln utvecklar sig över tid för alla länder.
5. Hitta vilken variabel som du tycker korrelerar mest mot "antal internetanvändare". Kan du hitta en tredje variabel som korrelerar bra mot din andra?
6. Underlätta att se överliggare genom att ändra genomskinligheten på linjer/prickar.
7. Vad betyder värdena som står bredvid färgskalan till vänster i applikationen?
8. Ändra ordningen på axlarna i PC plotten för att avgöra negativ eller positiva korrelationer.
9. Hur många internetanvändare hade israel 1995?

Fritt användande

Saker som fattas i programmet?

Features som man inte förstod att de fanns?

Allmän uppfattning/känsla vid användande?

Appendix C - Expert review

The result of the expert review. Issues that are known bugs are not included in the evaluation result. Cosmetic issues have mostly been skipped due to their sheer number.

Issues are graded on the following scale:

- Cosmetic problem only (need not be fixed unless extra time is available on project)
- Minor usability problem (fixing this should be given low priority)
- Major usability problem (important to fix, so should be given high priority)
- Usability catastrophe (imperative to fix)

General usability heuristics - Nielsen (1994)

- Visibility of system status
Minor - Map is not immediately updated when filtering

- Match between system and the real world
Minor - Use complete values instead of scientific E short version
Minor - Function descriptions in settings box, interpolation, clusters etc.

- User control and freedom
Minor - No reset button for the table lens
Minor - No reset button for the scatter plot
Major - No undo redo functionality

- Consistency and standards
Minor - No normal program menus
Minor - Possibility for confusion with axis colors in the PC plot
Major - Year headers are not clickable
Minor - The reset button of the zoom function conflicts with the reset button for the whole PC plot
Catastrophe - The rescaling of the scatter plot over time makes comparison impossible

- Error prevention
Major - The zoom function is rather prone to error and confusion, should be redesign or skipped
Minor - Filtering is reset when going to cluster mode
Minor - Filtering is reset when switching interpolation on and off

- Recognition rather than recall

- Flexibility and efficiency of use
Major - No keyboard shortcuts
Major - No way save program status

- Aesthetic and minimalist design
Minor - Replace some UI elements based on text with icons, for example the play button

- Help users recognize, diagnose, and recover from errors
Major - No error checking is performed on the excel data sheet (this might not be applicable)

- Help and documentation
Catastrophe - There is no help or documentation, at all

Specific information visualization heuristics - Shneiderman (1996)

- Overview

- Zoom
Major - The zoom function on the PC plot is incomplete and confusing
Major - No easy way to focus/zoom on chosen countries
Major - No way to make easy multiple countries grouped together
Major - No zoom at all in the table lens or the scatter plot

- Filter
Major - Filtering only possible via the PC plot (except for years)
Major - Difficult to restart a filtering with the PC plot because the arrows are inexact
Major - The zoom function on the PC plot is incomplete and confusing
Minor - The PC plot has problems with filtering when there exists NaN values in the data set
Major - Lack of percentiles filtering

- Details on demand
Major - The tooltip for the map is almost voided of information
Major - A lot more details on demand in all tool tips, perhaps connected to the first year or max/min values of all countries etc
Major - More detailed info in percentage form

- Relate

-History
Major - No undo redo functionality

- Extract
Major - No possibility to extract and/or save the state of the application