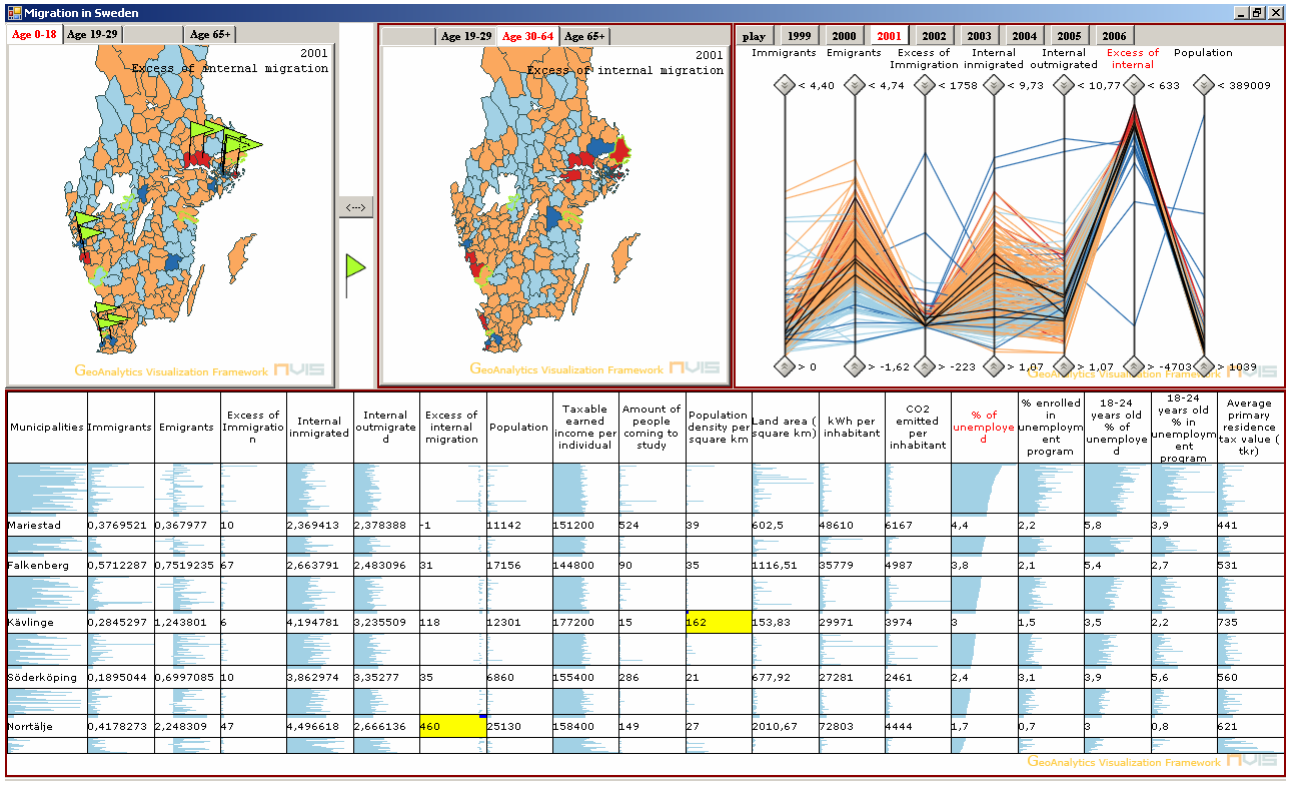
TNM048

Information Visualization

Migration in Sweden

By Cristina Vozian and Willem Frishert

2007-03-25



Municipalities	Immigrants	Emigrants	Excess of Immigration	Internal inmigrated	Internal outmigrated	Excess of internal migration	Population	Taxable earned income per individual	Amount of people coming to study	Population density per square km	Land area (square km)	kWh per inhabitant	CO2 emitted per inhabitant	% of unemployed	% enrolled in unemployment program	18-24 years old % of unemployed	18-24 years old % in unemployment program	Average primary residence tax value (tkr)
Mariestad	0,3769521	0,367977	10	2,369413	2,378388	-1	11142	151200	524	39	602,5	48610	6167	4,4	2,2	5,8	3,9	441
Falkenberg	0,5712287	0,7519235	67	2,663791	2,483096	31	17156	144800	90	35	1116,51	35779	4987	3,8	2,1	5,4	2,7	531
Kävlinge	0,2845297	1,243801	6	4,194781	3,235509	118	12301	177200	15	162	153,83	29971	3974	3	1,5	3,5	2,2	735
Söderköping	0,1895044	0,6997085	10	3,862974	3,35277	35	6860	155400	286	21	677,92	27281	2461	2,4	3,1	3,9	5,6	560
Norrtälje	0,4178273	2,248309	47	4,496618	2,666136	460	25130	158400	149	27	2010,67	72803	4444	1,7	0,7	3	0,8	621

GeoAnalytics Visualization Framework

Table Of Contents

Assignment

The aim of this group assignment is to develop an information visualization application in order to solve a visual analysis task on multivariate data. Public data is available at SCB statistical database (www.scb.se). The group decides which tools and techniques to use to make a visualization that fulfils the assignments goals (in-house developed GeoAnalytics Visualization GAV, .NET, C# with DirectX, Flash, OpenGL, VR etc.

The group assignment includes the following three tasks (percentages of contribution to final grade):

- Application (40%).
- Report and documentation (written in English) of the application including references, justification to used information visualization methods including found patterns and discoveries (40%).
- Public application demonstration and presentation of group assignment (20%).

The following techniques and features must be implemented.

- Visual representation (map and 2-3 multivariate graphs)
- Data transformation (for example: dimensionality reduction that reduce the number of variables considered for the application, such as, multi-dimensional scaling MDS or alternative, a clustering method for reducing the number of observations.
- Interaction including brushing, direct manipulation, coordinated and linked views

1 Introduction

Our project is about exploring the use of information visualization techniques and tools to visualize the relations between content and structure of large collection of data. The aim of our work is to illustrate how visualizations techniques can be used to give the user a better understanding of large quantities of information and amplify cognition. Software tools and language used are GeoAnalytics Visualization GAV (in-house component framework) and C# with DirectX.

This project uses the Analytical Reasoning as a basis for the Sense-making loop. Making assumptions, hypotheses, planning scenarios and discussing various issues are critical parts that carried our project along.

We set our sights on the migration data, the main reason being the fact that we are foreign students ourselves and move around a lot. It is captivating to analyze and discover the focal locations for migration within Sweden and the places foreigners move to. We also plan to uncover the motivations for migration, which are the major elements that force people to choose a specific municipality instead of another one and to move away.

2 Gathering Information

Our first choice was to analyze migration within Europe available from www.eurostat.com. However, not all European countries provide complete information, some data being present for one year and missing for another one. On top of that we ran into the problem of extracting the available data in a suitable format. Therefore we decided to analyze data from the Swedish statistical database (www.scb.se).

At the later website, we found two data categories of migration, both by municipality.

1. First category describes migration to and from Sweden from abroad.

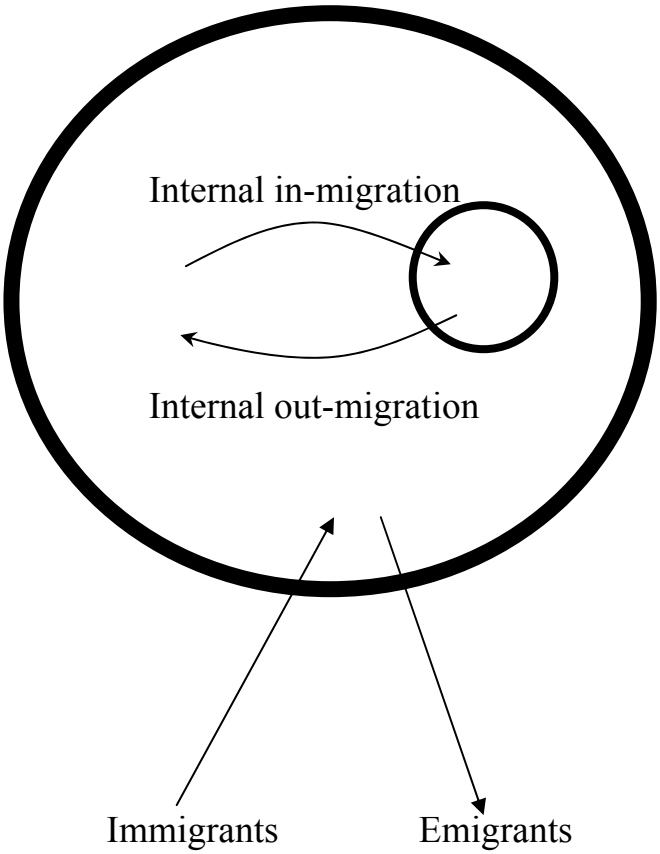
Immigrants – amount of people who come in from a foreign country.
Emigrants – people leaving the country to settle permanently in another place.
Excess of immigration – represents the difference between the amount of people coming into Sweden and the amount of people leaving Sweden. A positive excess shows that there are more people coming in than leaving. A negative excess shows that more people are moving abroad.

2. The second category describes migration within Sweden
   Internal in-migrated – coming into the municipality from another place in Sweden.
   Internal out-migrated – amount of people leaving the municipality and moving to some other places in Sweden.
   Excess of internal migration - represents the difference between the amounts of people coming in and moving out within Sweden.

Both categories are represented in the figure below.

Excess of internal migration =

= (Internal in-migration) – (Internal out-migration)



Excess of immigration (abroad) =

= Immigrants – Emigrants

Figure 1
Explains the main migration variables in the application
The big circle represents Sweden, the small one a municipality.

Population movement is a process dependent on time. We have chosen to analyze the recent period between 1999 and 2006, one of the reasons being the fact that for this period the data is more complete. We have split the migration data into four age groups.

1. Age 0-18
2. Age 19-29
3. Age 30-64
4. Age 65+

The age division was driven by the following rationalizations:

- Age group 0-18 represents children and teenagers. They are categorized as a separate group because unlike other age groups, generally they do not have resources or rights to move away by their own will.
- Age group 19-29 is dominant by people involved in higher education. Essentially, the movement within this age group is determent by the presence of a university and the level of education.
- Age group 30-64 represents people finishing their studies and the main reason to move could be in search of a better job.
- 65 is the average age of retirement. Possible the motivation to move for this category could be a safer environment.

Before proceeding with our application we visualize the migration data using GeoWizard2Time. We could already see some interesting trends and similarities in our data. For example we noticed that the outliers for age group 0-18 resemble the ones for age 30-64. We assumed that this similarity is a consequence of the fact that parents (30-64) move together with their children (0-18).

Using GeoWizard2Time also helped us realize another important side of the data. The rendering of the data on the Parallel Coordinates resulted in a cluttered representation because the extreme values where much bigger than the rest of the data. That was the reason why we transformed the initial data in percentage. The calculation of percentage is based on the total amount of population in the municipality by age group.

In order to discover the motivations for migration, in addition to the general population data, we selected more information such as environmental, energy, labour market, public finances, education, living conditions and household finances. Each of these data sets is responsible to answer a user specific task.

3 Re-representation

Different user tasks require different visual methods. Before deciding what visual tools to use we performed a task analysis. Our user is a person interested in analyzing what are the main points of attraction to move into a municipality. It can be the government of a foreign country, whose citizens are moving to Sweden, a real estates agency willing to know where is there market, or a company interested to find a better location for their central office.

First off all we thought that the user would be interested to compare between different age groups. That is why our application includes two maps representing different age groups. The maps are set side by side and it is possible to switch between them. We also added tabs on top of each map that allow the user to choose between age groups. The map on the centre of the screen is considered the active one, which means that its data is connected to other components of the application. To amplify cognition we added flags to the passive map. The flags indicate the presence of common extremes in both groups. For example while comparing ages 30-64 and 0-18, eight flags pop up indicating the same outliers.

Because of the multidimensionality of our migration data we chose to represent it on the Parallel Coordinates Plot. This representation gives a good possibility to compare between 2 categories of migration data: Internal Migration and Immigration. On top of that, PC is a helpful tool for filtering data and noticing relationships between 2 adjacent axes.

The colours selected for the components assist the user to quickly identify the main municipalities with the largest relocations of population:

Colour significance for the excess of migration:

Orange- indicates municipalities where there are more people coming in.

Light Blue – places from where more people move out.

Dark Blue - indicate extremes. The municipalities from where more people move away than come in

Red - displays the positive outliers, municipalities with a bigger amount of people coming in.

Colour significance for the rest of data represented on the map and parallel coordinates.

There is a transition from Blue to Red.

Blue - representing the minimum

Red – the extremes.

The outliers are calculated using the data mining algorithm called Inter Quartile Range. The interquartile range (IQR) is the distance between the 75^{th} percentile and the 25^{th} percentile. The IQR is essentially the range of the middle 50% of the data. Because it uses the middle 50%, the IQR is not affected by outliers or extreme values.

Example:

Compute the interquartile range for the sorted continue data:

18, 33, 58, 67, 73, 93, 147

The 25th and 75th percentiles are the

0.25*(7+1) and 0.75*(7+1)

= 2nd and 6th observations, respectively.

IQR = Q1-Q3=93-33 = 60.

To calculate the outliers we the use the following formulas:

Q1 - (1.5 * IQR) - top outliers

Q3 + (1.5 * IQR) - bottom outliers

To visualize our additional data: Environmental, Energy, Labor Market, Public finances, Educational, Living Conditions and Household finances, the PC is not sufficient. It doesn't work too well on more than 10 data variables. This is why we decided to build a new component.

It should be able to deal with vast amounts of multivariate data. We selected the table lens because it gives an overview of all data in the same window and detail on demand for the municipality of interest. In other words, the table lens allows us to show the data in focus and context at the same time. In the table lens, every variable is shown as a column and the values show up as lines which are scaled according to the size of that specific column. Our main interested is in the municipalities that are at the extremes of migration data. We want to know what makes some of these municipalities appealing to people and others unattractive. That is the reason why, in the table lens, we colour the background of a cell which contains extreme values with yellow as soon as this cell is in focus. In this way the user task to spot particular information becomes easier. We also order data by one column in ascending or descending order.

All components are connected to each other meaning that if a municipality is selected in one component then the other components are updated with the same item selected. The linked view of

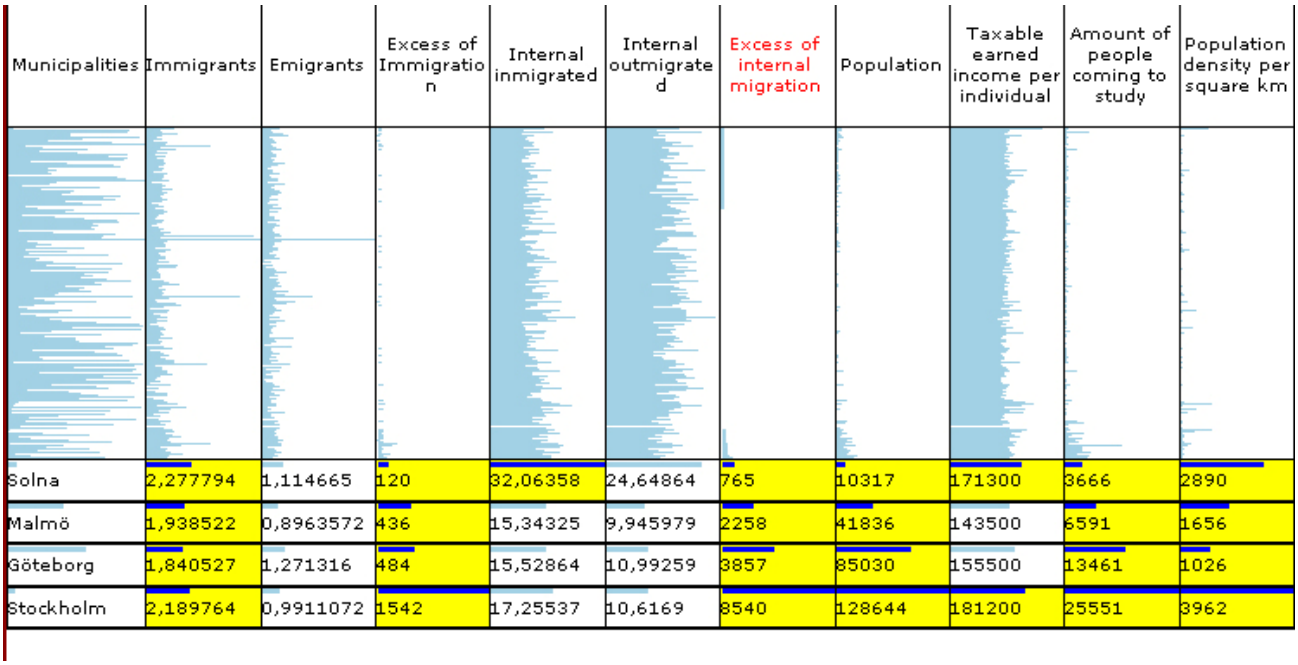
the same data make it easier to explore data, in essence it offloads cognitive work of the human visual perception system.

4 Developing insight

Once the gathering and creating of the components is done the prototype application is build. With this prototype a user can start to analyse the data. Before starting to re-represent the data, we made several hypotheses about the migration and other data. The next sections discuss the hypotheses and the assumption about the data. This implies that we are performing confirmatory analysis on the data to confirm, reject or make new hypotheses. To prove that the hypotheses are confirmed, we provide screenshots of the program.

As stated we made several hypotheses about the data which are listed below. The result of the analysis, and whether these hypotheses are true, will be discussed as well.

Hypothesis 1:
Young people in the age group of 19 to 29 are attracted to big cities to study and work there; this hypothesis needs to corrected. There are four municipalities which are attracting a lot of Swedes: Stockholm, Malmo, Goteborg and Solna. These places are also outliers when looking at the amount of people coming there to study. However, we can not see any reason for people of this age group to move to big places to find work.



Municipalities	Immigrants	Emigrants	Excess of Immigration	Internal inmigrated	Internal outmigrated	Excess of internal migration	Population	Taxable earned income per individual	Amount of people coming to study	Population density per square km
Solna	2,277794	1,114665	120	32,06358	24,64864	765	10317	171300	3666	2890
Malmö	1,938522	0,8963572	436	15,34325	9,945979	2258	41836	143500	6591	1656
Göteborg	1,840527	1,271316	484	15,52864	10,99259	3857	85030	155500	13461	1026
Stockholm	2,189764	0,9911072	1542	17,25537	10,6169	8540	128644	181200	25551	3962

Figure 2
Shows part of the table lens to prove hypotheses 1

The columns supporting this hypothesis are “Excess of internal migration” and “Amount of people coming to study”.

Hypothesis 2:
Elderly (65 and over) will move away from big cities. They want to avoid places where the population density is very high; confirmed. Elderly are trying to move away from places with big population density.

Municipalities	Immigrants	Emigrants	Excess of Immigration	Internal inmigrated	Internal outmigrated	Excess of internal migration	Population	Taxable earned income per individual	Amount of people coming to study	Population density per square km
Stockholm	0,1221636	0,1419739	-24	0,5926586	1,112679	-630	121149	191500	27032	3997
Göteborg	3,651539	5,946792	-44	23,57851	36,93487	-257	1917	163000	14294	1036
Malmö	2,22087	1,703681	17	9,005172	12,01704	-99	3287	149500	6617	1669
Huddinge	0,08782523	0,142716	-5	1,240531	1,899221	-60	9109	180400	6321	644
Sundsvall	0,0421509	0,0240862	33	0,529897	0,7647378	-39	16607	160400	1698	29

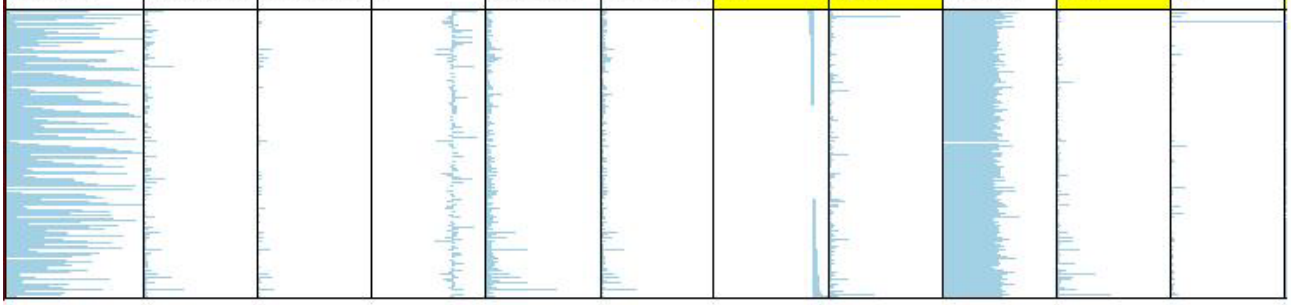


Figure 3
Shows part of the table lens to prove hypotheses 2

This hypothesis is confirmed by looking at the column of "Excess of internal migration" and "Population density". Note that the elderly are moving away from municipalities which attract people to study.

Hypothesis 3:
People in the age group of 0 to 18 are moving to the same areas with the people in the age group of 30 to 64. In common language this means that children are migrating together with their parents; this hypothesis seems to be correct. The municipalities which are outliers when it comes to excess in inmigrated seem to be similar for both age groups. The flags show the regions with similar outliers.

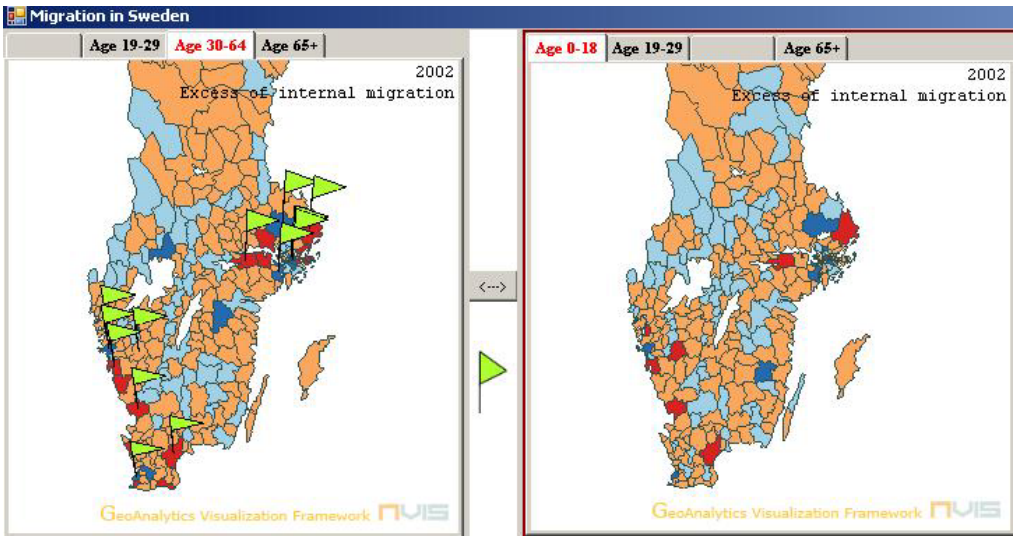


Figure 4

The image shows the active and passive map.
The flag indicate the common outliers in both age groups.

Hypothesis 4:
People of the age of 30 and over will move away from municipalities with a lot of CO_2 emissions; this hypothesis is rejected. It seems that there is no correlation between the emission of CO_2 and the migration of people

Hypothesis 5:
People engaged in the job market will try to move away from regions where there is a lot of unemployment; this hypothesis is confirmed. Most municipalities with a large number of unemployed also have a negative amount of excess which means there are more people moving away from this region than there are coming in.

Municipalities	Immigrants	Emigrants	Excess of Immigration	Internal inmigrated	Internal outmigrated	Excess of internal migration	Population	Taxable earned income per individual	Amount of people coming to study	Population density per square km	Land area (square km)	kWh per inhabitant	CO2 emitted per inhabitant	% of unemployed	% enrolled in unemployment program	18-24 years old % of unemployed	18-24 years old % in unemployment program
Pajala	0,3992629	-0,49140053		1,934889	2,825553	-29	3256	132300	18	1	7916,52	42925	5134	8,5	8,9	9,9	13,9
Haparanda	1,728696	2,643888	23	3,070978	2,155786	45	4917	134300	26	11	927,26	51028	5657	8,3	7	12,2	10,2
Övertorneå	0,9848174	0,4924087	8	2,256873	2,749282	-12	2437	131400	12	2	2381,35	47379	6394	8,1	8,7	11,2	12,6
Malmö	1,270456	0,6515161	718	2,876958	3,495898	-722	116651	158200	7041	1687	155,56	28400	3560	6,4	2,8	6	2,2
Bengtsfors	0,8563074	0	27	2,046784	2,903091	-41	4788	144500	184	12	889,16	95530	9384	6,3	3,7	7,5	3,9
Kiruna	0,2987878	-1,271982	-1	1,33174	2,90251	-184	11714	169400	226	1	19371,12	122398	18849	6,3	4,4	9,6	7,1

Figure 5

Shows part of the table lens to prove hypotheses 5

This hypothesis is confirmed when looking at the percentage of unemployed and looking at the excess of internal migration.

As shown with this confirmatory analysis, our hypothesis are confirmed or rejected. However, we do not know whether the user can easily manipulate the data and work with the application.

5 Usability study

To find the shortcomings of the prototype application we performed a usability test on it. Testing the usability was done by inviting several test candidates, who were not familiar with the data and the application, to sit in front of the computer and get a first impression of the application. One thing that needs to be mentioned about these test candidates is that all of them have experience with Windows and are used to the look and feel of applications running under Windows.

After this first glimpse, we explained to them what the data is representing, what the application is showing and which main tasks can be executed. Once the prototype had been explained, we allowed the test candidate to play around with the different views and asked them to fill out a questionnaire. The questionnaire consists of several closed questions and one open question and can be found in the appendix of this report. Using the answers from the questionnaire we are able to deduct the usability of the application and see which parts needs to be changed or improved. The following paragraphs discuss the usability of the prototype according to the test subjects and will list recommendations for changes.

- The test subjects said it was quite easy to work with the application and could easily manipulate the data but only after receiving a thorough explanation. All of them informed us that they would like to see some help functionality in the program explaining what the views represent and what the data shows. This help functionality should also make it clear which actions the user can take and what happens if the user hits a certain button. Looking at the layout of the screen, most of the test subjects were in agreement that it is clear and understandable for a new user. However, one of the test subjects mentioned that there was a lot of data on screen to look at and it was overwhelming when taking a glimpse at the program. When the candidates were explained the usage of the flags on the passive map and were asked if they found it useful, mixed reactions came from them. Several candidates thought the flags would be practical but were scared that they would clutter the map. They suggested making it possible to turn the flags on and off. One candidate said that he would also like to see the flags on the active map. Every interviewed person was in agreement that the buttons in the application are intuitive to use and prototype got an overall positive grading.

Conclusion

With the tool that we constructed, we were able to execute our confirmatory analysis to verify our set of hypotheses. The table lens has been applicable to our data and works well together with the GAV components.

In our usability test we posed the test candidates the open question which features they would like to see changed or would like to see added, they give us several suggestions. The list of suggested changes to the program and additions features is mentioned below. Feature list:

- Eliminate the lag in the program; make the components update faster
- Change the flag glyph to another type of glyph which won't overlap/clutter
- Emphasize the value of a variable in the table lens by using colors for the bars
- Create a group of migration data which represents all ages so users can compare age groups to the total migration data
- Center the test in the table lens and add tool tip to the column headers to show the meaning of the column

From usability test and list of features we can conclude that the application is able to respond to the user tasks but there is still space for improvement.

Appendix

User Survey

Conducted by **Willem Frishert** and **Cristina Vozian**

It is only 9 questions and takes you only 10 minutes to answer.
All your answers will be treated confidentially.

Instructions: Please answer the following questions by circling a letter or number that most closely represents your answer. Opportunity is also provided for your written comments.

1. **How easy is it to work with the application? (Please rate it.)**

A. Very easy	B. Easy	C. Normal	D. Hard	E. Very hard

2. **How easy is it to manipulate the data in the application to find useful information? (Please rate it.)**

A. Very easy	B. Easy	C. Normal	D. Hard	E. Very hard

3. **What do you think about the layout of the screen (i.e. is it clear and understandable)? (Please rate it.)**

A. Very good	B. Good	C. Normal	D. Bad	E. Very bad

4. **Do you find the flags on the passive map useful to compare different age groups? (Please rate it.)**

A. Very useful	B. useful	C. Normal	D. Not useful	E. Take them off

5. **Is the interface clear (i.e. do you understand the actions you can take within the application) ? (Please rate it.)**

A. Very clear	B. clear	C. Normal	D. unclear	E. Very unclear

6. **Do you find the buttons intuitive? (Please rate it.)**

A. Very intuitive	B. Intuitive	C. Normal	D. Not intuitive	E. Far from intuitive

7. Is it clear what the data represents? (Please rate it.)

A. Very clear	B. Clear	C. Normal	D. Unclear	E. Very unclear

8. Which features should be added to/changed in the program?

_____

_____

_____

_____

9. How would you scale the application? (Please rate it.)

A. Very good	B. Good	C. Normal	D. Bad	E. Very bad