Linköpings universitet

TNM085
Information Visualisation
Reported crimes and socio-economic conditions in Sweden

Johan Beck-Norén, johbe559
Viktor Axelsson, vikax333
Alexander Johansson, alejo849

2012-03-18

Abstract

This project is aimed to develop insight into if certain socio-economic parameters correlated with the rates of various categories of crimes in the municipalities of Sweden monthly during the years 2000 through 2010. The original hypothesis was that municipalities with a low rate of employment and a high population density would have an overall higher rate of crime all year round.

Using a number of visualization methods and different visual components the data could be explored interactively, and trends concerning the nation as a whole down to specific municipalities could be studied over time. An own component, a radar plot, was designed and implemented to enable comparison between socio-economic data and reported crimes for a specific municipality.

A usability evaluation was performed to find issues, learnability and usability problems in the application. As a consequence, a number of features were added, removed completely or modified.

The application allowed us to confirm our initial hypothesis, as well as presenting a number of previously unknown patterns in the data. For example, municipalities with a small number of permanent residents received a huge influx of reported crimes during vacation periods where the number of visiting people exceeded the permanent residents. In short, known vacations spots in Sweden all had a huge peak in reported crimes during holiday seasons.

Contents

1 Background and problem definition

The purpose of the project in the course Information Visualization was to visualize large multivariate data sets having both a spatial and a temporal dimension. In this project we chose to study if there is any correlation between a municipality's reported crimes and its socio-economic situation. Our general hypothesis was that the rate of reported crimes in municipalities with a high population density would be higher than that of municipalities with a lower population density. We were also curious if this was simply because of the high population density, or if there was another underlying socio-economic factor causing the high rate of reported crimes.

With this broad hypothesis we decided a map was needed in the application, as well as a way to examine the data set over time. To aid us in this we had the GAV (GeoAnalytics Visualization) framework for C#, using DirectX. The problems with handling this kind of data sets is both the large volume of the data set, and the multiple dimensions it possesses.We had to find a way to make the data easy to survey, yet at the same time give the user the possibility to study, in detail, specific municipalities. Our goal was to combine a spatial dimension and a temporal dimension, together with additional multi-variate visualisation components, all linked together.

2 Data and Information Gathering

Population- and crime statistics are both easily available from the homepages (Appendix B) of Brottsförebyggande Rådet (hereinafter BRÅ) and Statistiska Centralbyrån (hereinafter SCB). Unfortunately the statistics from BRÅ were arranged in a way that made it very hard to use with the Excel-file reader from the GAV framework. Luckily the VBA code language used in macros for most Office applications are not very advanced, so after some coding of the raw data is was ready to be read and processed.

The BRÅ database contains very detailed data on all crimes committed (for example; in 2010 one person with a disability was mugged by an assailant using a firearm). Using that level of detail would result in an application too crowded with categories, so instead we chose six major categories to examine:

- Total number of crimes
- Violent crimes
- Theft
- Car crime
- Vandalizing (including arson)
- Alcohol and drug related crime
- Weapons offences

From the population statistics obtained form SCB another six categories were chosen so that the data could be displayed on the same radar plot as the reported crimes data, and use all axis.

Population statistics:

- Populations average age
- Populations average income
- Education above upper-secondary education
- Employment rate
- Proportion of residents with foreign origin
- Population density

The time range used, 2000 to 2010, was decided upon due to the ranges of data available from BRÅ, which was only available on a monthly basis from 2000 and onwards and the population statistics from SCB that was not available after 2010 and only annually. Also any larger periods of time would result in rather large sets of data, with each year containing some 20 000 posts.

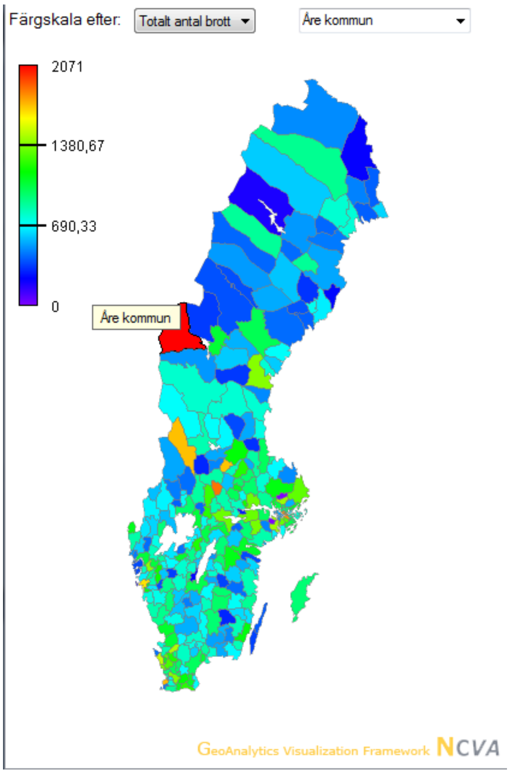
3 Choice of visualization methods

With data categorised after municipality a map was an obvious choice in what components to use in the application. The multiple dimensions of the crime data would be represented by a parallel coordinates plot, which can show a large number of dimensions at once and also supports dynamic filtering of the different dimensions. A scatter plot was implemented to discover correlations between different types of reported crimes. Finally, an own component was developed, a radar plot. This was used to compare reported crimes with a municipality's socio-economic situation. A track bar with a PLAY/PAUSE button was implemented, allowing the user to watch the data set as it unfolds over time or skip to a specific year and month on the time line.

3.1 Choropleth map

Using the GAV framework for C# a choropleth map is easily created and can be used both for visualizing data and to provide interaction for the other components by allowing searching and selection of items of interest. The color-scale used on the map can be set to represent any of the crime data categories and also shifted to change the threshold for different colors to easier group and filter.

The map interacts with all other components, selecting a municipality in the map will highlight it in the other components and if clustering is being used the corresponding cluster will be selected (figure 3.1).



Färgskala efter: Totalt antal brott

Figure 3.1: Choropleth map

3.2 Parallel coordinates plot

A parallel coordinates plot provides an easy way to filter values by each dimension (figure 3.2). It also easily show deviating values in each category which can be helpful since it makes it easy to monitor all dimensions at the same time. It shows correlation between dimensions but only to adjacent ones which is a limitation. Having a lowered opacity on the lines in the parallel coordinates plot also provides a sort of data mining method, since large concentrations of data will cause a brighter color where they converge.

The values on the axis of the parallel coordinates plot are set frame by frame based on the data at that time. This can cause confusion since the highest value of a category of crime in January and July will appear the same way in the plot, even if they differ greatly, and would require a look at the axis values to see which is the highest. But having the axis set to the highest values in the complete set of data would cause all but the highest values to appear at the bottom of the plot, so it was decided that the local values would determine the axis max/min values. Instead, to find how municipalities compare over time a scatter plot with set axis is used.

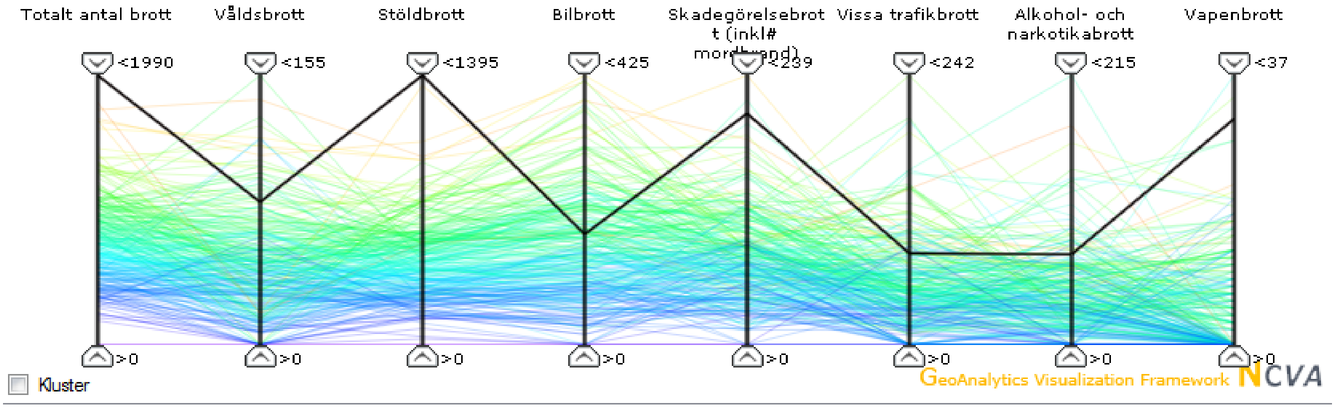


Figure 3.2: Parallel coordinates plot

3.3 Scatter plot

The scatter plot can represent data in four dimensions using the axes, color and size of the glyphs (figure 3.3). By changing the dimensions shown on the axes it can be used to find correlations between different dimensions. Linear correlations are easy to find using the axes. In the scatter plot a check box determines if the axis are dynamic or static, using the static axis will make it easier to compare absolute data between different time steps. When using the dynamic axes there is clutter in the low end of the plot, but easy to distinguish outliers.

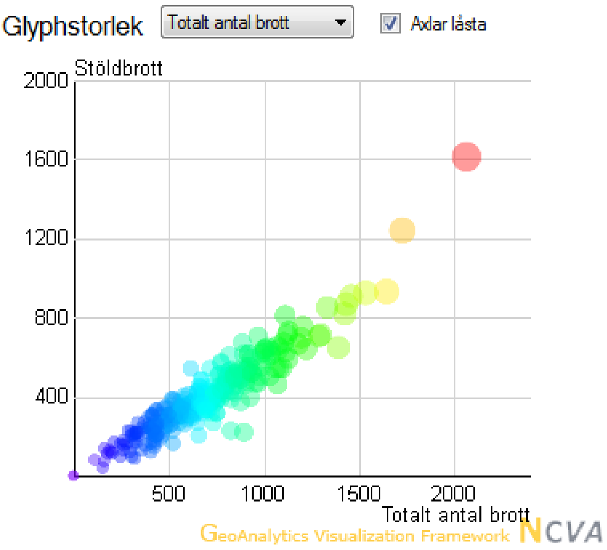


Figure 3.3: Scatter plot

3.4 Radar plot

The radar plot is a component we developed. It's intended use is to compare a specific municipality's reported crime rates and socio-economic situation. The plot's basic shape is an hexagon, which means that it can show six dimensions from each data set, ergo the radar plot can show a total of twelve dimensions (figure 3.4). The radar plot is linked to all other views and visual components in the program (parallel coordinates plot, scatter plot, choropleth map and the municipality search field). It is also linked together with the temporal dimension, so the hexagon representing reported crimes is also smoothly interpolated at the same rate as the rest of the reported crimes data set.

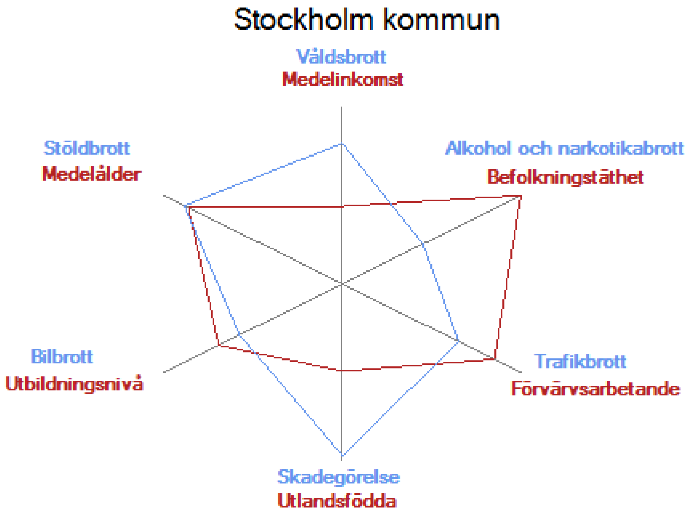


Figure 3.4: Radar plot showing Stockholm

3.5 Temporal dimension

A powerful part of the application is the track bar which allows the user to quickly scroll through the whole time frame of the data set. The ”Play”-button automatically scrolls through time, allowing the user to see how the data changes and progresses through time.

Since the data can change rapidly between the data sets it can be hard to follow how an item of choice actually changes. Therefore an interpolation method was implemented to make the transitions smoother, easier to follow and as a bonus it is quite aesthetically pleasing. The interpolation is a quite simple linear interpolation, for each interpolation step the intermediate value is calculated (3.1).

$$Interpolatedvalue = \frac{previousvalue * (totalsteps - currentstep) + nextvalue * (currentstep)}{numberofsteps} \tag{3.1}$$

4 Data mining methods

The data mining method of choice was a clustering method called k-means clustering. The k-means clustering algorithm arranges the data items in different clusters using centroids. The centroid’s position is calculated from the mean value of the data items in the current cluster. When this is done the Euclidean distance from the centroid to the data item is calculated. The centroid with the shortest distance to the data item is the one assigned to the current data item. This is repeated until no change occurs from the last iteration of the loop.

5 Exploring the data

To begin with we only had the general conception that are probably shared by many, that most crimes occur in larger cities where there is a lower level of education, more unemployment and lower salaries. This could be confirmed to some extent by studying the position of municipalities with high population density in the scatter plot. It becomes even more clear when we use the k-means method for clustering the data. Using that, the municipalities with high population density belong to the same cluster, and the cluster lies in the high end of the scatter plot. For example, Stockholm, Malmö, Göteborg and Huddinge belong to the same cluster (figure 5.1).

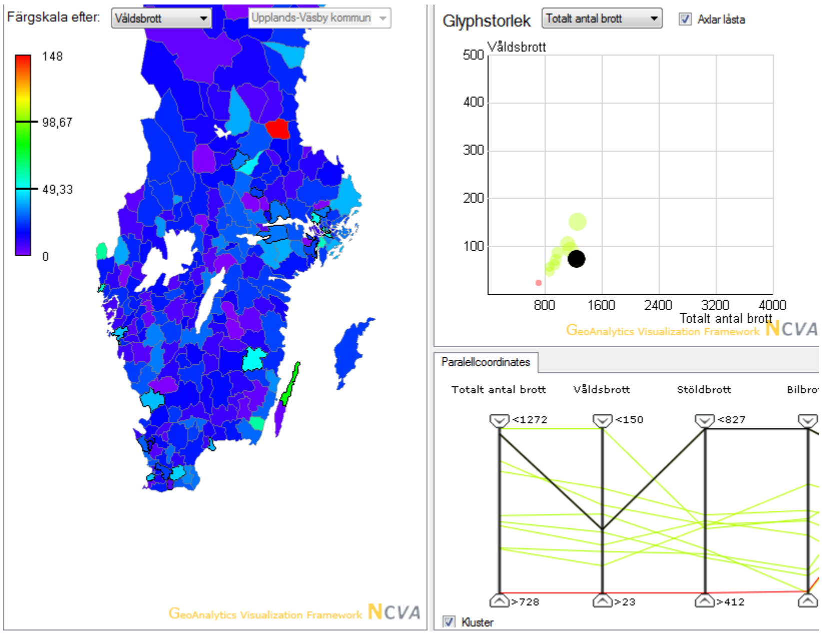


Figure 5.1: Cluster

The most noticeable trends found in the data set depended on people travelling to certain municipalities for holidays and events. It is especially easy to see when the municipality in question has a low amount of inhabitants, making the relatively large number of visitors and their influx of crime noticeable.

6 Usability evaluation

6.1 User evaluation

To detect design flaws and usability issues in our application a usability evaluation was performed. The main aim of the evaluation was to measure effectiveness, receive feedback about the GUI and test learnability. The radar plot had not yet been implemented when this evaluation was performed, so no feedback on it's design or usage was received.

Two groups consisting of two people were given three specific tasks to complete and a few broad questions about the user interface. A short interview was then conducted about the perceived impressions the users got when using the application. The users all had experience of using GAV components but none had seen this specific application or interface before.

We found that for the users to get started they had to be given a short introduction to the application and the data it represented. Short explanations about the different visual components were given before the user got started using the application. The conclusions draw from this first part of the evaluation are that the interface has to be more self-explanatory to the user.

The three tasks given to the users (Appendix A.1) where formed in a way so the users had to utilize all the different components of the application. The time it took for the users to complete the tasks were recorded. For example, one user had problem with a task involving following a specific municipality's crime trend over time because the choropleth map filled a selected region with black. The user wanted to be able to view a selected map region's color map over time instead of following the data item in the parallel coordinates plot or the scatter plot. Conclusions draw from this part resulted in using a thick border layer for the choropleth map's selected region instead of a black fill. The function to drag and drop headers from the parallel coordinates plot on to an axis of the scatter plot was expanded to allow bigger areas for the headers to be dropped on to the scatter plot. A bigger and more clear label for the time axis was also added, along with real values on the color legend for the choropleth map. All users liked the fact that tool tips showing the municipality was used on both the choropleth map and the scatter plot, and the clear links between the choropleth map, parallel coordinates plot and scatter plot. None of the users used the table lens during any stage of the evaluation, so the decision was made to remove it completely from the program.

The short interview conducted consisted of broad questions about the perceived look and feel the user experienced during the evaluation (Appendix A.1). The overall impressions were positive with a few minor points to consider.

6.2 Expert review

The expert review was done continuously throughout the development of the program. We constantly evaluated and re-evaluated different components and visualisation methods with the end user in mind. A scale with four grades of severity was used, running from cosmetic problems to catastrophic. A top 10 heuristic list was used (Nielsen, 1994) and updated with new entries as problems arose (Appendix A).

7 Conclusions

Some conclusions could be drawn by just exploring the data for a short period of time. Summer months generally mean a large influx in overall reported crimes in municipalities known to be popular vacation spots. For example, the music festival *Hultsfredsfestivalen* attracts over 30 000 visitors to the small town of Hultsfred which has roughly 14 000 inhabitants. The large number of non-permanent residents and the large quantities of alcohol consumed at the festival makes Hultfred clearly stand out in the plots, not only as clearly the most alcohol and drug related offences per capita but also thefts and weapons offences in July when the festival takes place.

A little more exploration of the data also reveals a linear correlation between reported burglaries and vehicle crimes.

The hypotheses we had in the beginning of the project could be confirmed, showing that municipalities with high population density such as Stockholm and Malmö had a high number of crimes all year around compared to other municipalities. But there was no correlation that could be confirmed between socio-economic parameters and certain crime types. Municipalities with high influx and low population density are shown as outliers in the program,.

A Usability evaluation

A.1 User evaluation

A.1.1 Tasks (timed)

- Find the municipality with the highest rate of violent crimes in July of 2005.
- Find the municipalities with the high amount of total reported crimes the first quarter of every year.
- Examine the trend of reported crimes in Hultsfred over a few years. Can any conclusions be drawn?

A.1.2 Perceived look and feel

- Was it easy to orientate in the program?
- Would you like to add anything?
- Would you like to remove something?

A.2 Expert review

A.2.1 Top 10 heuristics, Nielsen (1994)

- Visibility of system status
  - Minor - We would have liked to have the playback speed and number of K-MEANS clusters visible in the main window of the GUI instead of in the tool strip. Unfortunately, there wasn't enough space in the GUI to allow this information to be displayed directly.
- Match between system and the real world
  - Major - We should make it clearer to the user that the numbers presented are per capita and not absolute numbers.
  - Minor - The interpolated values displayed when viewing the data over time could be confused with being real values. We added a label to the date telling the user the data shown is interpolated, but it could still be confused with real data values.
- User control and freedom
  - Minor - If the user want to reset the program to it's initial state it has to be done manually. No reset button exists. This is not a major issue since there are only a small number of parameters to reset and these are rather easy to find.

  - Major - There is no functionality implemented to reset the zoom of the choropleth map. It is quite easy to find yourself zoomed in and lost on the map and it would have been nice to have a reset button for the map.

- Consistency and standards

  - Minor - The implementation of dragging and dropping headers from the parallel coordinates plot to the axes of the scatter plot was modified so the headers could be dropped anywhere in the scatter plot, but the axes assigned with the header is not always consistent. For example, a header can be dropped relatively near the Y-axis but still end up on the X-axis.

  - Catastrophic - There is no support from dragging and dropping the axes headers from the radar plot to any other part of the program. The easiest way to examine correlations between reported crimes and socio-economic situations would for example mean average income on one axis and violent crimes on another axis of the scatter plot. At present, the only way to compare socio-economic data to reported crime data is for a specific municipality in the radar plot.

- Error prevention

  - Major - Due to the implemented clustering methods, interpolations and the sheer volume of data, the application is rather demanding to run. If the computer used to run the application is out-dated or slow, the program is prone to locking and sometimes even crashing. No error message is shown, but the user will notice the program running slow for a few moments before locking up.

- Recognition rather than recall

  - Minor - There is now way for the user to directly see the number of clusters being used or the current playback speed. This information is located in the tool list instead.

- Flexibility and efficiency of use

  - Major - There is no way to save the programs current state. That means there is no way to take a snapshot of the application's state at a given time if something interesting is found.

- Aesthetic and minimalist design

  - Major - There are no fixed labels on the track bar axis. Permanent markings indicating years could be added to facilitate faster browsing to a specific year or month.

- Help users recognize, diagnose and recover from errors

  - Minor - No error messages exists. On the other hand, we could not produce a situation where an error message would be necessary.

  - Minor - The Excel document containing data sets has to be formatted in a certain and specific way to work in our program. This is a minor issue for the end user, since we assume the end user will not be concerned with acquiring data sets.

- Help and documentation
  - Major - In the tool list there is a menu item called ”Help”. This displays a panel with general information about the application and the different views and functionalities. This is only very broad and introductory information. More detailed help is not available. The user is forced to explore data using the application and get familiar with it. If this application is to be used outside the context of this course, a comprehensive help system has to be implemented.

B Data sources

- Statistiska Centralbyrån - *www.scb.se*
  - Born abroad by municipality per 100 000
  - Higher education by municipality per 100 000
  - Average age by municipality per 100 000
  - Average income by municipality per 100 000
  - Gainful employment by municipality per 100 000
  - Population density per km^2 by municipality
- Brottsförebyggande rådet - *www.bra.se*
  - Violent crime per month by municipality
  - Burglary per month by municipality
  - Vehicle crimes per month by municipality
  - Vandalizing per month by municipality
  - Traffic violations per month by municipality
  - Alcohol and narcotics per month by municipality
  - Weapons offences per month by municipality