Exploring the Unknown: Analysing Data through Interactive Visualisations

Zena Wood, Clive E. Sabel, Niko Karvosenoja and Ville-Veikko Paunu

Abstract—Organisations, businesses and individuals are gaining access to an increasing amount of data, which could contain meaningful information that could impact their businesses or quality of life. Increasingly individuals want to be able to analyse their own datasets. However, extracting the information often requires the use of a data analytics expert or a piece of specialist software, which is not always easy or cheap to obtain. Interactive visualisations allow non-experts to explore their datasets to understand what meaningful information they might contain. This paper presents an online tool that has been developed using open source methods that will allow a user, without data analytics experience, to explore their datasets. The tool was produced as part of an EU FP7 project, URGENCHE.

Index Terms—Data analysis, Geospatial analysis, visualisation,

1 Introduction

The amount of data that is being created and collected by individuals and businesses is rapidly increasing, resulting in very large and complex datasets [15]. Within such datasets there is likely to be meaningful information that could create business value or a better quality of life but this information can only be extracted through analysis. Although the tools to collect and store large datasets have advanced to deal with the growth in the amount of data, the open-source tools that can be used by non-experts are still lacking.

There are many applications where users may have minimal analytical or statistics knowledge, or access to special tools, but must analyse datasets to extract meaningful information and make decisions (e.g., marketing, health organisations and campaign managers). Sometimes the information that might be contained within the datasets is time-critical and, therefore, must be extracted quickly (e.g., for emergency response and policy makers). Visual data exploration can facilitate extraction of meaningful information from large and complex datasets without the need for the user to have specialist knowledge of statistics or mathematical algorithms. The technique can also allow more efficient treatment of noisy or inhomogeneous data [14].

- *Z. Wood is in the Computing and Information Systems Department within the Faculty of Architecture, Computing and Humanities, University of Greenwich, UK*
  E-mail: z.m.wood@exeter.ac.uk
- *C. Sabel is with University of Bristol, N. Karvosenoja and V. Paunu are with SYKE.*

Visual Analytics is a growing research field that combines statistics, data mining and visualisation. It focuses on analysing datasets with the use of interactive techniques and visual representations [15], enabling insight and conclusions to be drawn for better decision making. An application of Visual Analytics, Exploratory Spatial Data Analysis (ESDA) is a collection of techniques that focus on the analysis of spatial datasets (i.e, describe and visualise spatial distributions, patterns and trends of spatial association and spatial outliers).

Although the integration of visualisations and interactions to assist in the analysis process is not a new concept, the advances in technology has enabled new and existing techniques to be combined to produce powerful analytical tools. Visual data exploration can be of particular use when there is minimal understanding of the dataset and there are only vague goals for exploration [14]. Allowing the user to directly interact with the data means that the exploration goals can adapt as necessary. Making use of the human user by allowing them to interact with the data is crucial ([3], [4]). They bring their expert background knowledge but also the inherent capability that the human brain has to identify trends and patterns within the dataset ([5], [15]).

Urgenche (Urban Reduction of GHG Emissions in China and Europe) was an EU FP7 project that focused on developing a methodological framework that would allow the benefits and risks of different GHG (Green House Gas) emission reduction policies for health and well-being to be evaluated in China and Europe. The project involved seventeen partners including six cities from Europe and China that were each been paired with a research institute from their

city. Each city collected and collated a number of datasets relating to health and well-being that could be analysed to evaluate the current state of these factors in each city and the effect different GHG reduction policies would have in the future. One objective of the project was to develop an online GIS tool that could be used by each city to interactively explore their datasets. The tool needed to be web-based, build upon existing open-source software and be able to handle health and environmental data with core datasets including data relating to population, socio-economic status, topography, regional climate, infrastructure and administrative areas. Users of the tool were expected to have minimal, if any, knowledge of analytics or statistics.

This paper will present the current version of the online tool, that has been developed as part of URGENCHE to allow a user without data analytics knowledge to dynamically explore their datasets through the use of interactive visualisations. The tool was implemented in d3, a JavaScript library that allows data to be brought to life through the manipulation the documents that form a webpage. It can be used to produce web-based dynamic and linked visualisations of large datasets.

An overview of the use of interactive visualisations for analysing data and the challenges that must be addressed is given in section 2 followed by a review of existing software that adopts interactive visual analysis (section 3). The requirements of the tool are outlined in section 4 leading to a description of the tool itself (section 4.2). The tool has been evaluated through a case study, which is presented in section 5. A discussion of the success of the tool in allowing users, without any expertise in data analytics, to dynamically explore their datasets is given (section 6). The paper concludes with the extensions to the tool that are currently being undertaken based on the results of the evaluation.

2 The Use of Visualisation and Interaction in Data Analysis

Existing research and systems that use interactive visualisations for data analysis have shown that there are key concepts and techniques, in addition to the associated challenges, that must be addressed.

Visualising large datasets is often difficult given the limitation of human visual perception. Graphs visualising datasets that contain a large number of records can often become quickly cluttered and difficult to interpret ([1], [15]). Techniques such as filtering and aggregation can be used to help reduce the amount of data being displayed but risk a loss in the amount of detail being communicated. Therefore, any visual representation that presents a high-level abstraction of the data must ensure that the amount of detail that is being presented is maximised. One option is to ensure that the user can 'drill-down' to obtain more details about the data where necessary [15].

Keim *et. al* define visualisation as 'the communication of abstract data relevant in terms of action through the use of interactive visual interfaces' and outline three overall goals: presentation, confirmatory analysis and exploratory analysis [15]. These three goals are often referred to as the Visual Analytics Mantra: "Analyze First - Show the Important - Zoom and Filter, and Analyze Further - Details on Demand" ([14], [15]). Essentially the user is presented with an overview of the data before identifying which patterns they wish to explore in more detail. Initial hypotheses relating to the data are presented for confirmatory analysis, which are accepted or rejected through the use of visualisations. Exploratory analysis is used to analyse the data to identify any useful implicit information. This last goal usually does not begin with a hypothesis and will involve an undirected search.

There are many different ways in which a dataset could be visualised and it is important that the most suitable method is chosen given the analysis that is to be undertaken. A three-dimensional classification of techniques for visualising information and data mining is presented within [14]. The three dimensions of the classification are: the type of data that is being considered; the techniques that can be used for visualisation; and the techniques that allow interaction or distortion. The data that is to be visualised could: be one, two or multidimensional. It may also comprise text, hierarchies or graphs. The visualisation techniques are categorised into standard two- and three-dimensional displays (e.g., scatter plots and bar charts), 'geometrically transformed displays' (e.g., landscapes), 'icon-based displays', 'dense pixel displays' and stacked displays.

Keim notes that the dimensions of the classification are orthogonal and that different types of data can be visualised in different ways with different methods of interaction possible. However, not all types of visualisation are suitable for different types of data. For example, it would not necessarily be suitable, or possible, to represent textual information as a scatterplot or a choropleth map. Therefore, having some understanding about the type of data that is being analysed is necessary when choosing the most suitable visualisation method. The type of data that is being analysed (i.e., qualitative or quantitative) and its features, should be carefully considered since it will govern the type of visualisation that is chosen [2]. For example, if the data is quantitive a symbol is needed that can express quantities such as size and position.

The method chosen for the visualisation must also be suitable for the task that the user is trying to achieve; understanding how the visualisations may be used may help with this decision ([16], [18]). Some types of visualisation lend themselves to certain tasks. For example, trends within datasets can typically be

identified using techniques such as quadrant counts or histograms and spatial autocorrelation using scatter plots or linked-windows.

An understanding of the data can be greatly affected by the quality of the datasets that are being considered. Missing and erroneous records are likely and must be dealt with. Pre-processing the datasets can help minimise the problems of data quality but it is important that any uncertainties or known errors within a dataset are explicitly shown in the visualisations [2].

The use of dynamic graphs that the user can interact with to immediately change how the data is being presented to them can assist in identifying trends [5]. Complex problems cannot be solved using static images alone. Instead, different views of the data are required with the user able to transform the images to see different views of the data [2]. As the exploratory goals change the user can use interaction techniques to adapt the visualisations accordingly. Within the classification presented by Keim the interaction component considers: filtering, zooming, linking and brushing, and distortion [14]. Interaction techniques can also allow multiple visualisations to be linked. Linking and brushing techniques can often be used to overcome the problems and shortcomings of a single visualisation method. Brushing can allow correlations and dependencies to be identified as any brushed points would be highlighted in all of the visualisations.

Given the adopted method the best parametrisation must be chosen, in addition to the most suitable visualisation method. However, when dealing with unknown datasets it can be difficult to establish the most 'suitable parameter settings' [1]. To overcome this, dynamic visualisations could be adopted that show the different parameter settings via animation; the user could then stop the animation when they believe that they have spotted an interesting pattern. There could also be an opportunity for the user to interact with the developed tool to change the parameter settings according to the task that they are trying to undertake.

To Dorling [11] visualisation means making visible what was obscure. By transforming large amounts of data into pictures, we can start to understand the underlying structure. The aims of Visualisation in Scientific Computing (ViSC) are not new to geographers. Much geographic enquiry seeks to discover and provide explanations for (spatial) patterns and relationships. Geographers have traditionally used visual analysis because the display of data within a spatial framework enables us to recognise patterns. ViSC has developed more recently than GIS, and might be loosely defined as 'exploring data and information graphically, as a means of gaining understanding and insight into the data' [20]. It is apparent therefore, that there are many areas of interaction between GIS and ViSC. In the context of statistics, the methods of scientific visualisation clearly build on the ideas of Exploratory Data Analysis, and therefore, ViSC is at the heart of a response to appeals for exploratory spatial data analysis in GIS [13].

MacEachren *et al.* [17] prefer to focus upon the term Geographic Visualisation (GVis) as a sub-discipline of ViSC. GVis, takes its principles from cartography, GIS, exploratory data analysis and information visualisation, 'to develop and assess visual methods that facilitate the exploration, analysis, synthesis and presentation of georeferenced information' [17]. GVis involves the use of computer graphics to stimulate the human visual system to recognise patterns that would not otherwise be obvious. GVis can therefore be extremely beneficial in all stages of (interactive) exploratory data analysis (EDA).

3 Existing Software

Online publishing of georeferenced data is not new; spatial data has been published online since the early 1990s, shortly after the web became publicly available. The first internet mapping solutions were simple, only able to produce maps as static image files for example a Graphics Interchange Format (GIF), then shortly after, in the mid 1990s with the appearance of JavaScript (which allowed users to execute client side requests), interactive services became available such as pan, zoom, and query. The rapid growth in popularity of the internet, emerging opportunities to publish georeferenced data and the growing public interest in accessing spatial information, combined to spawn a large number of web mapping applications such as MapGuide, Mapquest, MultiMap, and ArcIMS; some commercial, and others, free and open to the general public.

Standalone pieces of software have been developed that use interactive visualisations for data analysis in addition to specialist modules that extend existing software. Software modules that allow interactive ESDA within existing GIS software such as ArcView have been developed by a number of groups (e.g., XGobi dynamic graphics software [19] and SpaceStat [7]). A brief review of existing software in terms of dynamic interaction through the use of linked windows can be found in [5]. This section will briefly review two standalone systems: IRIS and GeoDa.

IRIS is a system that allows spatially referenced data to be explored using automatically generated visualisations, referred to as thematic maps, that the user can interact with and manipulate [2]. Two versions of the system are available: a web-based version or a standalone version. The data that is to be analysed is stored in uniform tables where each field within a record forms a column. Meta-information is required, which specifies the associations between columns and any concepts relating to a column. The

user must select the columns that they would like to visually analyse resulting in one or more maps being automatically generated and displayed. All possible visualisations of the selected data are generated and displayed according to the types of data that the user has selected and the relations and associations that are known about the data. The user can select additional columns within the table to visualise. Once new columns have been selected all current windows will be closed and the new visualisations displayed to avoid overcrowding. However, the user can reopen previously generated maps for comparison.

The way in which data is encoded into visualisations has been carefully considered. The visualisations are generated using the principles and rules of visualisation; many of the visualisation techniques that are found in cartography have also been included. The main principle between selecting the most appropriate visualisation technique is to ensure the the properties of a symbol, which they refer to as a 'visual variable', is consistent with the data type that it is representing. For example, if dealing with quantitative data the symbol must be able to depict a size or a value; unordered data can be visualised using different shapes and colour. It is not just the data type that affects the choice of visualisation technique, the meta information is also important. The system will group together components that need to be visualised according to comparability and each group is then visualised using an appropriate technique.

IRIS contains specialised facilities to interact and manipulate the data; these facilities enhance the capabilities of the visualisation technique that has been chosen. The facilities include: visual filtering, allowing multiple maps to use the same colour scheme and allowing for dynamically hidden data (i.e., certain qualitative values may be switched on or off to determine whether they are displayed in the current visualisation). The user can select the relevant portion of the map (i.e., the displayed visualisation) to view the exact data values. Variables can be restrained by setting upper and lower bounds. Arithmetic formulae can be constructed by the user within IRIS to derive information from the data. The derived information can then be visualised, filtered or used in further calculations. Logical operators can be used to produce more complex visualisations.

IRIS has many merits. The user does not have to think about how to present the data and multiple views are given through the use of different visualisations to help different properties of the data become evident. The goal of the tool is not to simply display the final visualisations but allow the user to interactively explore a dataset. A good evaluation of the tool is given within [2] by walking through an example of what a typical analyst (i.e., a non-expert) has done, why they have done what and what they discovered. The evaluation showed that the user did discover previously unknown facts about the dataset.

GeoDa is a program that has been built to introduce spatial analysis to non-GIS specialists [6]. The tool was developed following the recognition that an important component of facilitating and disseminating spatial analysis would be through an interactive piece of standalone software designed for a non-specialist that would be visual and easy-to-use. A key feature of GeoDa is the use of linked windows that the user can interact with where each window present a different perspective of the data presented as maps, table or graphs; there is no limit to the number of windows that can be linked. Anselin *et. al* note that the 'user-friendliness' of the GeoDa that distinguishes it from similar software.

GeoDa guides the user through a typical spatial analysis route: the data is presented through visualisations that the user can interactively explore. Following exploration the user can use the inbuilt methods for spatial autocorrelation and regression analysis. The user is required to do no programming and, instead, is offered a point-click interface. Similar techniques for spatial statistics are offered in GeoDa to those found in R functions. However, the mapping and visualisations are more advanced than those offered within the R environment. Unlike this environment, GeoDa can not be extended or customised by the user. Given that the users are meant to be non-specialist this additional functionality seems unnecessary.

GeoDa has been designed to analyse discrete geospatial data (i.e., objects that have point or boundary coordinates). Shape files are used to store spatial information. The majority of the visualisations focus on specialised chloropleth maps that can be used to identify outliers. The inbuilt mapping techniques include chloropleth, map animation and cartograms. The statistical graphics that are often found within EDA are also included (i.e., histograms, scatter plots and box plots) in addition to the less common graphs: parallel coordinate plots (PCP), conditional plots and three-dimensional scatter plots.

The implementation of GeoDa that was documented within [6] was limited to MS Windows platforms due to the use of Microsoft Foundation Classes (MFC) as the basis for the graphic windows. Extending the system to be cross-platform and open-source were listed as items for further work in addition to extending the limited functionality in terms of spatial regression.

Evans and Sabel [12] present a design and development of an open source web-based Geographical Information System allowing users to visualise, customise and interact with spatial data within their web browser.

4 The New Tool

The tool that was to be developed needed to be web-based, build upon existing open-source software and

allow users, who do not necessarily have detailed knowledge of statistics or analytics to interactively explore health and environmental data. Within [18], Robinson *et. al* present seven factors to be considered when designing maps: the purpose of the map, reality (i.e, the phenomena that is being mapped), the characteristics of the available data, map scale, audience, conditions of use (i.e., the environment that the map will be used in) and technical limitations (e.g., whether it is to displayed digitally or not) [18]. Although these were considered for map design many of these factors are relevant for designing visualisations in general and they were considered when identifying the design choices of the tool.

- Purpose:
  - The tool should read in a dataset (health or environmental) and present different visualisations of the data that the user can interact with.
- Reality:
  - It is not known what might be recorded within a dataset but there may be some spatial information present.
- Available data:
  - It is not known which datasets will be used. However, discussions with potential users (i.e., the cities within the URGENCHE project) revealed that the datasets that they would wish to analyse with the tool could exist in three different formats: comma-separated values files (.csv), JavaScript Object Notation files (.json) and shapefiles (.shp).
- Audience:
  - Users with little, if any, expertise in statistics and analytics that may or may not be decision makers.
- Conditions of use:
  - The tool is web-based and it is assumed that it will be used on a desktop.
- Technical Limitations:
  - The computer used by the user to run the tool will need to be connected to the internet. Since the tool is web-based, the connection speed will bring some limitation to the size of data file that can be uploaded; the current version of the tool will have to be tested with different file sizes and connection speeds to establish the relevant limitations.
- Map Size:
  - Although map size is not directly relevant, the resolution of the displays where the system will be used must be considered. A minimum resolution of $1366x768$ will be assumed; this is the most popular screen resolution of visitors to W3 schools [10].

The review presented in section 2 clearly showed that dynamically linked visualisations are an important way for a user to establish any correlations or trends between variables. The tool should simultaneously present the data to the user in different formats with the multiple 'views' of the data linked (i.e., any interaction on one graph should affect the remaining graphs); this should allow patterns and impacts to be clearer to a user and aid with scenario analyses. For example, if the user chooses to highlight some of the data in one graph, the same data would be highlighted in the remaining graphs of that dataset.

There are many common visualisation methods that are used for geographical datasets including a range of charts, plots and graphs. Choosing the most suitable visualisation given the dataset and what you are trying to show is important but not always trivial especially since given the lack of knowledge about the datasets that the tool will be used for. Therefore, a variety of visualisation options will be made available to the within the GIS tool. This is a similar approach to IRIS. Visualisation methods should include tables, a variety of charts and choropleths. The user should be made aware that the different visualisations that are displayed may be used to identify initial trends but further exploration is likely to be needed to confirm those patterns.

The Visual Analytics Mantra should be followed. Therefore, the tool should initially present an overview of the data. The user should be able to select which parts of the data they would like to explore in more detail and the visualisations that they would like to use. Where possible, a visualisation should clearly show any known errors or uncertainty.

Consultation with the cities indicated the following types of analysis would be useful within an online environment.

- Filtering:
  - Data is filtered to only display those matches a specific criterion (e.g., where values are greater than, equal to or less than a given value).
- Proximity analysis:
  - Identify features that are near to another feature.
- Point density maps:
  - Splits the map into a neighbourhood and calculates the density of points per cell where a point will represent a given feature.
- Clipping tool
  - Allows a user to focus on a target region.

4.1 Summary of Requirements

Following the discussion in the previous section, the requirements for an initial version of the tool are summarised below.

[R1] The tool would be web-based and build upon existing tools.
[R2] Two types of dataset should be accepted: .csv, and .json.
[R3] Initially only quantitative data will be accepted. Files containing only text and graphics will not be processed.
[R4] Visualisation methods should include tables and a variety of charts.
[R5] The user should be able to select which parts of the data they would like to explore in more detail.
[R6] The tool should automatically extract the labels of the variables that are contained within the dataset and display these to the user.
[R7] The data should be simultaneously presented different formats (e.g., in a table, as a bar chart or scatterplot) with the multiple 'views' of the data linked to allow brushing (i.e., any interaction on one graph should affecting the remaining graphs).
[R8] The user should be able to select different variables to visualise in different ways.
[R9] The tool should provide the user with the option to filter the data that they are analysing.

4.2 Implementation

Initially three suitable technologies were identified as being potential solutions: D3.js, Python and R. D3.js, a JavaScript library, has been recently developed as a web-based technology that focuses on the visualisation of data allowing documents to be driven by the data on which they are based [8]. Python is a scripting language that is becoming increasingly popular with GIS developers; a number of libraries have been developed specifically for use in GIS applications. Applications need not be web-based if developed in Python. R is a programming language that has been developed for statistics and graphics. Like Python, libraries have been developed for R that are specific to GIS applications. Discussions with potential users suggested that the interactivity and powerful visualisations that D3.js offered made it the most suitable option. Therefore, the tool GIS tool was developed using D3.js and the necessary web technologies (e.g, php).

The remainder of this section will describe the functionality of the tool according to the requirements listed in section 4.1. To illustrate the functionality of the tool, a sample dataset will be used throughout; the dataset has been chosen because it allows all of the capabilities of the tool to be shown and incorporates a variety of information relating to population and socio-economic status.

For development purposes the tool was split into three components: uploading, processing and visualisation.

4.2.1 Uploading

Initially the user must upload the dataset that they wish to analyse (figure 1). The current version of the tool is limited to analysing one dataset at a time; this point is discussed further in section 6. The system can currently process .csv (comma-separated values files); the tool will need to be extended to accept .json (JavaScript Object Notation files) (see section 6).

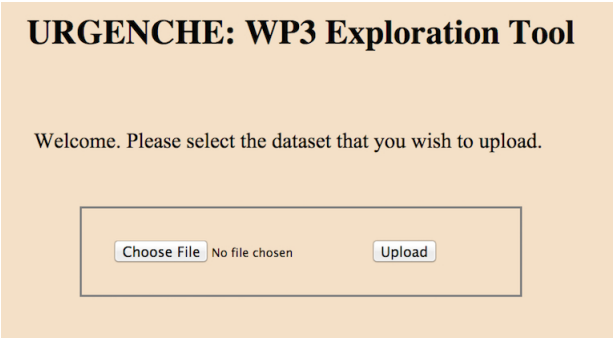

Fig. 1. The interface for uploading a dataset

4.2.2 Processing

Once a dataset has been successfully uploaded the tool will identify the categories within the dataset that could be analysed. Similar to IRIS it is assumed that the data will be in a column format where each field within a record forms a column. The tool extracts the column headings and displays them to the user who can then select which ones they may wish to include in later analysis.

4.2.3 Visualisation and Analysis

A range of visualisations have been developed within the tool that allow a user to interactively explore their datasets. These visualisations include different charts: bar, scatter and line. A user can select different categories to compare and display, focusing on a small section of the dataset or the entire dataset (i.e., filtering). Up to three charts can be shown simultaneously within the browser to help identify patterns and anomalies. In the current version of the tool these charts are not fully linked but some brushing and filtering is enabled (i.e., the user can zoom in on a section of a graph to see it in more detail).

5 Case Study

A sample .csv dataset will be used to evaluate the tool. The dataset contains information relating to the health and well-being of individuals living within a region of England. For each area within the region the dataset records scores for factors such as employment, income, health and crime levels. The dataset also records the health of the men and women in each area based on their weight and the proportion that

are considered healthy and obese. This section will show how the tool can be used to explore this sample dataset.

The dataset was formatted so that each record formed a row with columns separating the different variables that have been recorded; column headings existed within the dataset to indicate what each value in a row represented. Once read in by the tool, these column headings were automatically extracted and presented to the user using a set of drop down lists. The user must select two variables that they wish to initially compare and the way in which they wish to visualise the data. The current choice of visualisations include a single visualisation (bar, scatter or line), the raw data shown in a table, or a set of visualisations. The categories are always presented along the top of the screen. At any point the user can choose a different set of visualisations by using the 'Update' button.

The initial choices that are given to the user for the sample dataset can be found in figure 2.

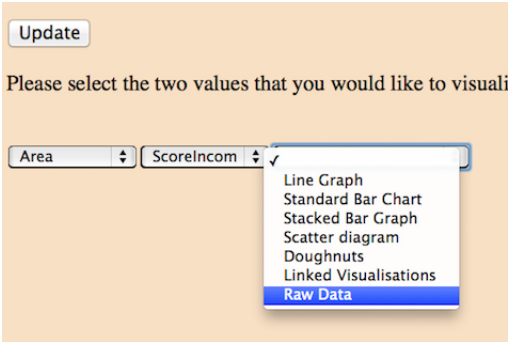

Fig. 2. The interface for selecting the data to visualise

The current version of the tool offers the user the option to compare two variables using either a bar, line or scatter plot. The user can select up to three visualisations to simultaneously display in the browser; these visualisations could consider the same two variables or different ones. Figures 3 and 4 depict examples of the scatter and bar charts. Below each graph a table shows some basic statistics (i.e., the minimum, maximum and mean) relating to the variables that are being considered (figure 5).

Figure 6 shows an example of two visualisations being used to present different 'views' of two variables in the same browser. Figure 7 shows an example of two visualisations being used to come three variables in the same browser; common variables are depicted in the same colour to aid comparison.

In addition to the visualisations, the user can choose to view the data as a table. It is important to allow the user to see the actual values when they wish to.

6 Evaluation

The case study has been useful in showing how the current version of the tool satisfies the majority

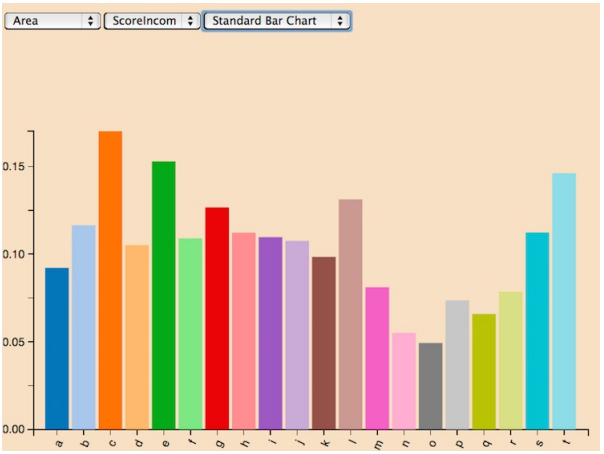

Fig. 3. Example visualisation I: a bar chart

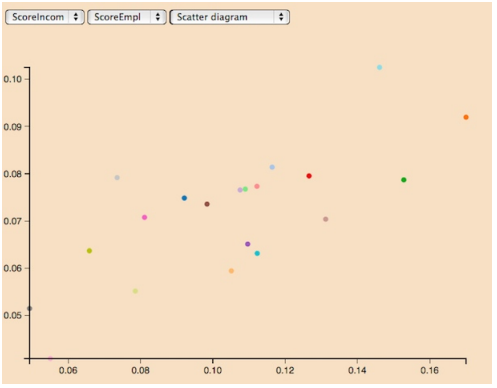

Fig. 4. Example visualisation II: a scatter plot for two variables

of the requirements set out in section 4.1. Through examining the example dataset benefits of the tool have been highlighted in addition to features that must be extended or developed for the next version of the tool and before the software can be released.

The web-based tool allows a user to upload a .csv file and explore their dataset by comparing different variables simultaneously using a range of visualisation types. Although a user needs to have Javascript enabled on their machines to use the tool, no specialised software or libraries need to be installed. This increases its accessibility to a wider community of users and is likely to encourage more people to use it.

Considering large datasets can be problematic and visualisations showing only a high-level of abstraction can lead to a loss of detail. The tool allows the user to drill down by viewing the raw data. Simple statistics have also been used to give context to the data that the user is visualising. The user can filter the dataset by only considering a subset of variables.

Not all visualisation methods are suitable for a given dataset. However, the best ones are difficult to identify without knowledge of the dataset. Offering

Variable	min	max	mean
X	0.04929	0.17003	0.10464
Y	0.04086	0.10251	0.07161

Fig. 5. Showing simple statistics

a variety of visualisations allows the user to quickly identify unsuitable charts and discard them. The variety of visualisation methods offered to the user can also allow erroneous data to be identified. For example, when considering one variable within the sample dataset, a recurring dramatically low number was quickly identified. Going back to the raw values showed that, where data was not available, a very low negative number had been recorded. The tool is currently being updated to allow the user to filter out such values.

The dataset used in the case study clearly indicates the need for choropleths to allow spatial relations to be easily identified. Therefore, an extended version should be able to accept shape files. This would allow information to be viewed on a map but also allow proximity analysis and point density maps.

The target user is expected to have little or no knowledge of statistics. The tool currently does not allow uncertainties or errors to be visualised [9] unless they are an explicit category within the dataset. It would be useful if a future version of the tool could overcome this by calculating some simple descriptive statistics (e.g., distribution, correlation) and presented them to the user. Given the lack of statistics knowledge additional graphics could be added to help explain the different concepts and why they are important.

Challenges arise from both the volume and complexity of the datasets, which often come from various heterogeneous data sources, which can be difficult to combine. The tool should be able to process and combine more than one dataset. This is particularly important with the push towards Big Data and the unceasing availability of multiple datasets to an individual or business.

7 Conclusion

Individuals and business have increasing access to large amounts of data that could improve their lives or business income. However, access to tools that would allow them to explore these datasets without knowledge of data analytics or access to specialist software can be difficult. An initial version of a tool has been presented that allows users with little or no expertise in data analytics to dynamically and visually explore their datasets. Developed as part of an EU FP7 project, the tool makes use of open-source technologies that could allow a wider community of users than existing software. The use of open-source technology will allow a user to extend and customise the tool if they wish. A sample dataset was used to illustrate the benefits of the tool and how it might be used. Evaluation of the tool has indicated how it should be extended in the future before being released as the final version.

References

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-orientated data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008.

[2] G. Andrienko and N. Andrienko. Iris: an intelligent tool supporting visual exploration of spatially referenced data. In *Proceedings of 3rd ERCIM Workshop on User Interfaces for All*, pages 61–74, Abernai France, 1997.

[3] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *IEEE Visual Analytics Science and Technology (VAST 2009)*, pages 3–10, Atlantic City, New Jersey, USA, October 12-13 2009. IEEE Computer Society Press.

[4] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A systematic Approach*. Springer, Berlin, 2006.

[5] L. Anselin. *Geographical Information Systems*. Wiley, 2 edition, 1999.

[6] L. Anselin, I. Syabri, and Y. Kho. Geoda: An introduction to spatial data analysis. *Geographical Analysis*, 38:5–22, 2006.

[7] Luc Anselin and Shuming Bao. Exploratory spatial data analysis linking spacestat and arcview. In Manfred M. Fischer and Arthur Getis, editors, *Recent Developments in Spatial Analysis. Spatial Statistics, Behavioural Modelling, and Computational Intelligence*, pages 35–59. Springer Berlin Heidelberg, 1997.

[8] Mike Bostock. D3 data driven documents.

[9] D. J. Briggs, C. E. Sabel, and K. Lee. Uncertainty in epidemiology and health risk assessment. *Environmental Geochemistry and Health*, 31(2):189–203, 2009.

[10] Refsnes Data. http://www.w3schools.com/browsers/browsers.

[11] D. Dorling. Cartograms for visualizing human geography. In H.M. Hearnshaw and D.J. Unwin, editors, *Visualization in Geographical Information Systems*, pages 85–101. John Wiley & Sons, 1994.

[12] B. Evans and C.E. Sabel. Open-source web-based geographical information system for health exposure assessment. *International Journal of Health Geographics*, 11(2), 2012.

[13] A.S. Fotheringham. Exploratory spatial data analysis and gis. *Environment and Planning A,*, 24:1675–1678, 1992.

[14] D. Keim. Information visualization and data mining. *IEEE Transactions on Visualization and Computer Graphics*, 7(1):100–107, 2002.

[15] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Tenth International Conference on Information Visualization*, pages 9–16, 2006.

[16] P. Longley, M. Goodchild, D. Maguire, and D. Rhind. *Geographic Information Systems and Science*. John Wiley & Sons, 2010.

[17] A.M. MacEachren, F.P. Boscoe, D. Haug, and L. W. Pickle. Geographic visualisation: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Proceedings IEEE Symposium on Information Visualization (InfoVis '98)*, North Carolina, Oct 18-23 1998.

[18] A. Robinson, J. L. Morrison, P. Muehrcke, A. Kimerling, and S. C. Guptill. *Elements of Cartography*. John Wiley & Sons, 6th edition, 1995.

[19] Deborah F. Swayne, Dianne Cook, and Andreas Buja. Xgobi: Interactive dynamic data visualization in the x window system. *Journal of Computational and Graphical Statistics*, 7(1):113–130, 1998.

[20] D. J. Unwin, J.A. Dykes, P.F. Fisher, K. Stynes, and J.D. Wood. Wysiwyg? visualization in the spatial science. In *Proceedings of 1994 AGI Conference*, Birmingham, UK, 1994.

ACKNOWLEDGMENTS

This study was conducted as part of the URGENCHE project coordinated by Prof Sabel and funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 265114.

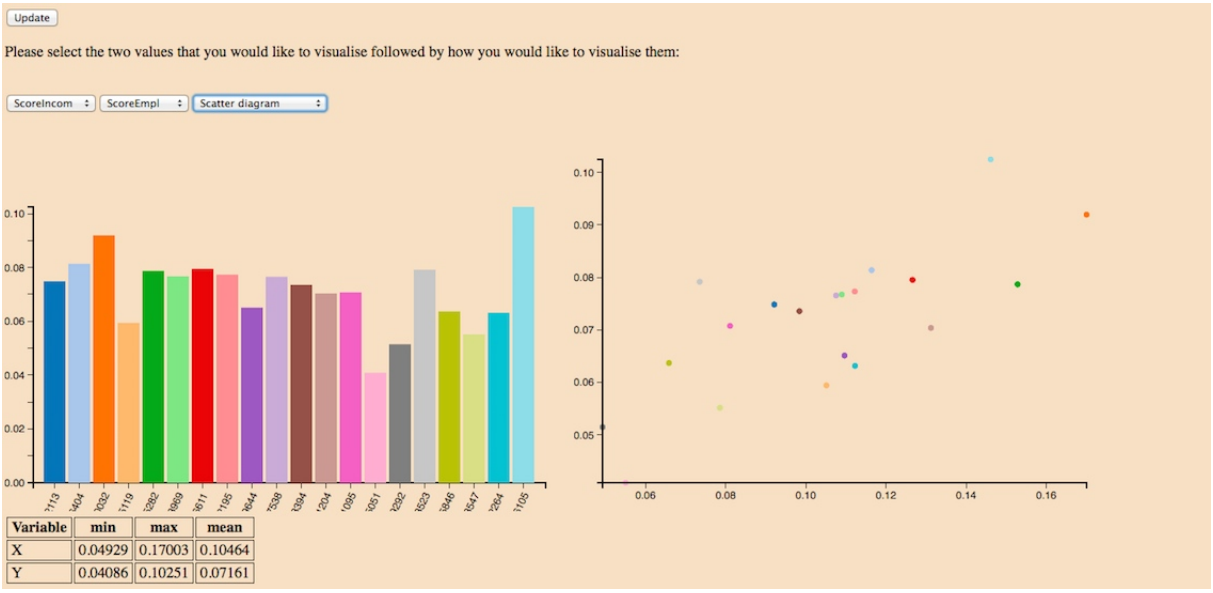

Fig. 6. Example of simultaneous visualisations for two variables

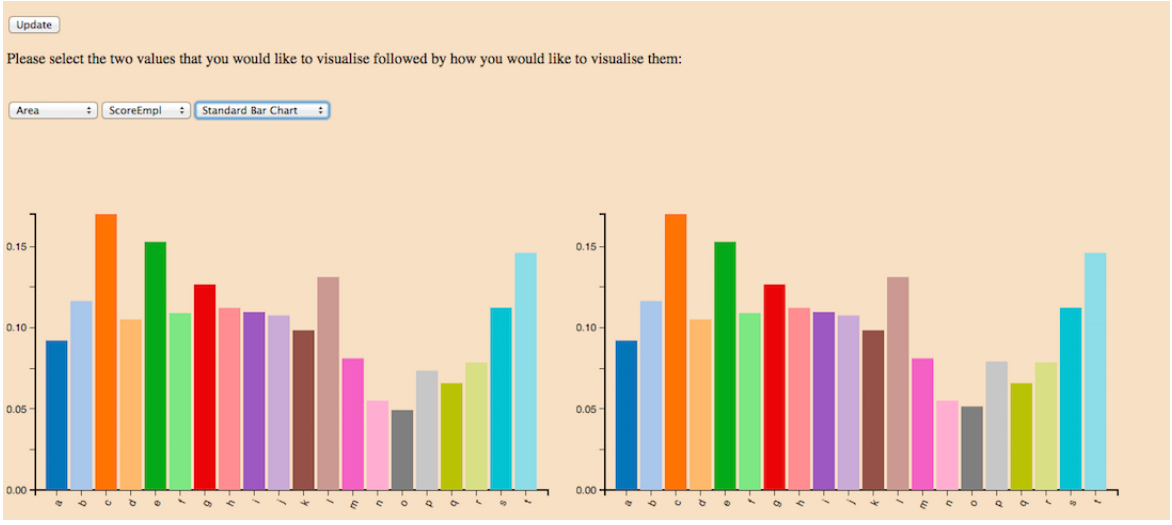

Fig. 7. Example of simultaneous visualisations for three variables

RE: TVCG-2015-02-0058, "Exploring the Unknown: Analysing Data through Interactive Visualisations"
Manuscript Type: Regular
Authors: Wood, Zena; Sabel, Clive; Karvosenoja, Niko; Paunu, Ville-Veikko

26-Feb-2015

Dear Dr. Wood:

Your manuscript entitled "Exploring the Unknown: Analysing Data through Interactive Visualisations" has been successfully submitted online and is presently being given full consideration for publication in IEEE Transactions on Visualization and Computer Graphics.

Your manuscript ID is TVCG-2015-02-0058.

Please mention the above manuscript ID in all future correspondence or when calling the office for questions. If there are any changes in your contact information, especially your e-mail address, please log in to ScholarOne Manuscripts at https://mc.manuscriptcentral.com/tvcg-cs and edit your user information as appropriate.

You can also view the status of your manuscript at any time by checking your Author Center after logging in to https://mc.manuscriptcentral.com/tvcg-cs.

Thank you for submitting your manuscript to Transactions on Visualization and Computer Graphics.

Sincerely,

Joyce Arnold
IEEE Transactions on Visualization and Computer Graphics