Single view vs. multiple views scatterplots

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 2, April 2019, pp. 1426~1436
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i2.pp1426-1436

Single view vs multiple views scatterplots

Federico Manuri, Andrea Sanna, Fabrizio Lamberti
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

Article Info

Article history:

Received May 4, 2018
Revised Nov 6, 2018
Accepted Nov 21, 2018

Keywords:

Information visualization
Multidimensional
Multiple views
Multivariate
Scatterplot
Visual analytics

ABSTRACT

Among all the available visualization tools, the scatterplot has been deeply analyzed through the years and many researchers investigated how to improve this tool to face new challenges. The scatterplot visualization diagram is considered one of the most functional among the variety of data visual representations, due to its relative simplicity compared to other multivariable visualization techniques. Even so, one of the most significant and unsolved challenge in data visualization consists in effectively displaying datasets with many attributes or dimensions, such as multidimensional or multivariate ones. The focus of this research is to compare the single view and the multiple views visualization paradigms for displaying multivariable dataset using scatterplots. A multivariable scatterplot has been developed as a web application to provide the single view tool, whereas for the multiple views visualization, the ScatterDice web app has been slightly modified and adopted as a traditional, yet interactive, scatterplot matrix. Finally, a taxonomy of tasks for visualization tools has been chosen to define the use case and the tests to compare the two paradigms.

Corresponding Author:

Federico Manuri,
Dipartimento di Automatica e Informatica,
Politecnico di Torino,
Corso Duca degli Abruzzi 24, 10129, Torino, Italy.
Email: federico.manuri@polito.it

1. INTRODUCTION

Data visualization and data analytics are important areas of research due to the dramatic increment of information produced on a daily base by the information systems and infrastructures all over the world, also known as information overload. Moreover, the exponential growth of the Internet of Things (IoT) is providing even more services based on data, as they are collected from many kinds of sensors [1], [2]. The capability to correctly analyze these massive, typically messy and inconsistent volumes of data is crucial due to the insight they contain [3], [4], it can be the trend of the financial markets, the production efficiency of a big company, the most requested holiday destinations or consumer goods. Thus, a significant problem is how to comprehend and exploit these data so that they provide a cognitive advantage that can led to more aware choices than before. The scatterplot may be considered as one of the most effective and adaptable visualization tools [5]-[7], especially in statistics. The basic version of a scatterplot displays a set of data as a collection of points using Cartesian coordinates. The coordinates of each point correspond to two variables, x and y, calculated independently to form bivariate pairs (xi, yi). A scatterplot visualization of xi against yi can provide insight about the correlation and the dependence between x and y as well as it can help to identify outliers, clusters of points and many other useful information [8], [9].

Even if the scatterplot is considered a powerful tool for data analysis, one of the most significant and unsolved challenge in data visualization consists in effectively displaying datasets with many attributes or dimensions, such as multidimensional or multivariate ones [10]. Various techniques have been researched and adopted through the decades, following two main approaches [11]. The first one consists in adopting multiple views through a scatterplot matrix, whereas each scatterplot displays a combination of two attributes

from the dataset. Overall, the matrix displays all the possible attribute combinations. Another approach is to adopt techniques which increment the dimensionality of the scatterplot, representing the data through a single visualization.

The focus of this research is to compare the single view and the multiple views' visualization paradigms for multivariable datasets using scatterplots. The term multivariable is hereafter used to denote datasets which are multidimensional or multivariate. Since various techniques have been researched and adopted through the decades to increase the scatterplot's dimensionality, the first step consisted of a literature review of such techniques and a selection of the most used ones. Then, a multivariable scatterplot has been developed as a web application to provide the single view tool, whereas the selected graphic effects have been realized through the d3.js library [12]. For the multiple views' visualization, the ScatterDice [13] web app has been slightly modified and adopted as a traditional, yet interactive, scatterplot matrix. Finally, a taxonomy of tasks for visualization tools has been chosen to define the use case and the tests to compare the two paradigms.

The paper is organized as follows: Section 2 provides a detailed analysis of both history and state of the art for the scatterplot. The design and development of the multivariable scatterplot is presented in Section 3, followed by a description of its counterpart, the ScatterDice. Use case, tests and results are discussed in Section 4, whereas conclusions and future works are summarized in Section 5.

2. THEORETICAL BASIS

2.1. Information visualization

The first step towards the understanding of data is the way they are represented: an effective graphical visualization can provide a quick and intuitive way to comprehension. However, the effectiveness of the visual effects adopted to represent different variables deteriorates when the number of variables increases. Another problem to be considered is that too often visualizations become a product of the visual analysis instead of an exploration tools that the experts can use to extract meaningful information from their datasets. Thus, it is important that the tools enable users to interact with the represented information to further perform analysis on the data and gain insight useful for decision-making.

Overall, the goal of visualization tools is to reduce the user cognitive work needed to perform certain tasks, which usually involve retrieving information and deriving insight from massive, dynamic and eventually conflicting data. However, finding effective ways of presenting high-dimensional data is still a major challenge in information visualization. To this end, multivariable visualization techniques should further help users, turning the information overload produced by datasets that present multiple attributes into an opportunity to augment the discovery process.

2.2. Multidimensional and multivariate data

Datasets with a high number of variables have been defined by Wong and Bergeron as multidimensional multivariate data [14]: the dataset represents a set of observations X, where the i-th element xi consists of a vector with m variables, xi = (xi1, … , xim). Each variable m may be independent or dependent on one or more other variables. Independent variables are referred to multidimensional variables and dependent variables are referred to multivariate ones. This pose another problem, as the user might not know if the data are multidimensional or multivariate, thus if a correlation between the data exist. Eventually, this could be one of the question the user may want to address when analyzing the dataset with a visualization tool.

Multidimensional visualizations are commonly used to display a sample dataset as points within the n-dimensional domain D in order to estimate the function f(x) which describes the dataset over the entire domain [15]. This is simple when the number of dimensions is small and matches the visualization tool dimensions, whereas it is a much harder problem when high-dimensionality datasets are involved. In such cases, the most common method is the Hyperslice from van Liere and van Wijk [16], [17]: this technique requires to slice the original n-dimension dataset into m 2D subspaces visualization, where m=(n(n-1))/2. This produces an n x n matrix of 2D visualizations displaying two variables, whereas the diagonal shows in 1D visualizations displaying only one variable.

Multivariate data refers to a set of data with dependent variables: usually, a multivariate dataset is collected as a table, each column representing an attribute and each row representing an observation of that attribute. The goal of multivariate visualizations, depending on the context, may consist of searching patterns, clusters, trends, behaviors or correlations among attributes, supporting the elaboration of hypothesis about the phenomenon represented by the data.

2.3. Scatterplots

The scatterplot visualization diagram is considered one of the most functional among the variety of data visual representations, due to its relative simplicity compared to other multivariable visualization techniques [18]. Scatterplot visualizations are usually appreciated because they easily reveal nonlinear relationships between variables. scatterplots are also used to establish correlations among variables within a certain confidence interval. Another common usage for the scatterplot is the comparison of similar datasets. Even if the scatterplot is considered a powerful tool for data analysis, one of the most significant challenge in data visualization consists in effectively displaying datasets with more than two variables. For this reason, various techniques have been researched and adopted through the decades to increase its dimensionality, since the effectiveness of the traditional scatterplot becomes ineffective as data grows. For example, additional variables may be displayed by correlating them to one or more graphical features of the plotted points. Overall, these graphical additions can either enhance existing capabilities of the scatterplot or provide additional capabilities that standard scatterplots do not have at all. Moreover, the development of these graphical additions is based on graphical principles that can be applied to graphics in general. However, these approaches usually produce visual representations that are tailored on specific dataset and/or tasks.

In 1984, Cleveland and McGill published a study regarding the enhancing of the scatterplot adding graphical information [19]. They proposed four different categories of graphical effects, which comprehend both traditional ones and new ones: sunflowers, coding categories, point cloud sizing and smoothing. After Cleveland and McGill, many other authors proposed graphical augmentation for the scatterplot. Even if a taxonomy or classification is out of the scope of this paper, a brief description of the most common 'augmentation' to the scatterplot visualization tool is proposed, since an analysis of the most used effects is necessary to develop a multivariable scatterplot.

2.3.1. Size

The simplest option to display an additional variable consists of varying the size of the point. This kind of scatterplot, which could display three dimensions of a dataset, is commonly known as 'Bubble Chart' [20]. Anyway, this option may lead to occlusion problems if the plot does not provide proper scaling on the two axes.

2.3.2. Color

Colored points on a scatterplot may suggest similarity among values of the same dataset or correspondence among points of different datasets. Moreover, this correlation may be perceived without drawing any connecting line. Colors can also be used to enhance the perception of a variable already displayed by another effect. Colors can be used for 'coding categories' or even to represent stereoscopic scatterplots as proposed by Wells [21].

2.3.3. Graph

Another possibility to increase the number of variables displayed by a scatterplot consists of displaying the points of the dataset by a shape or geometrical figure. A sunflower is a specific type of glyph represented by a dot and several line segments departing by it, whereas the number of lines depends on a variable. The usage of the sunflower technique has been suggested to overcome the problem of occlusion due to overlapping points [19]. Another usage for glyphs consists of using specific symbols to represent all the points pertaining to a given category or dataset. This technique is known as 'coding categories' and provides many possibilities, all with pros and cons: one consists of using letters, usually choosing the capitol one of each category. Other coding schemes consist of using different figures and/or colors to represent thepoints [19], [22], [23]. Another technique based on glyphs consists of using arrows or other shapes that can suggest a directional information to represent an additional variable [24].

2.3.4. Label

A string of characters can be placed adjacent to each point to provide an additional information. Elliot Noma proposed a heuristic method to solve the problem of overlapping when adding multicharacter labels to a scatterplot [25].

2.3.5. Point cloud sizing

Point cloud sizing refers to the proportion between the cloud of points and the scatterplot frame: if the point cloud size decreases respect to the size of the frame, the capability of the user to correctly identify linear association increases. Overall, the cloud should not get too far from the frame or too close to it, and Cleveland and McGill even proposed a procedure for correctly sizing the point cloud in a scatterplot [19].

2.3.6. Smoothing line

Another technique involves the usage of smoothing lines: firstly, it is necessary to compute one or more set of smoothed values from the original dataset; then, the functions describing the smoothed values distributions could be plotted as lines over the scatterplot. For example, it is possible to represent the middle of the distribution of y at x=xi, so that the smoothed points form a regression of y on x. The smoothed values can be obtained following different procedures, such as the lowess method (locally weighted scatterplot smoothing) or the robust locally weighted regression proposed by Cleveland [26]. The purpose of smoothing lines is to simplify the evaluation of the dependence of y on x. Moreover, many other types of smoothing lines may be computed to provide different kind of visual information, e.g. spread smoothing lines which display the spread of y given x, or upper and lower smoothings which represents another measure of the spread.

2.3.7. Smoothing density

This technique consists in moving from plotting the individual dots to display them as an empirical and uniform distribution, to better represent the points' density [27]. Other researchers further investigated this technique and the problems related to density visualization [28]-[31]. For example, Bachthaler and Weiskopf proposed the Continuous scatterplot to combine a statistical visualization method such as the scatterplot with scientific visualization methods like volume or flow visualization [32], [33].

2.3.8. Scatterplot matrix

A scatterplot is a diagram showing a set of data as a collection of points using Cartesian Coordinates. A scatterplot matrix consists of a series of scatterplots, one for each pair of variables, displayed together on a single screen [34]. If the dataset consists of k variables, it requires k(k-1)/2 pairs and therefore scatterplots. Unfortunately, this solution presents a major problem: analyzing all the scatterplots may require a lot of time, depending on the number of variables, thus this solution is not optimal when dealing with time-related tasks.

2.3.9. 3D scatterplot

Another option to display multivariable data consists in adopting a 3D scatterplot visualization. 3D scatterplots exploit the third dimension, representing three data dimensions on the x, y and z coordinates, in a three-dimensional space. The third dimension enriches the visualization displaying an additional data dimension. Moreover, it allows the user to interact with the graphical representation changing the viewport. Unfortunately, it is not advisable to abuse multidimensionality if it is not absolutely necessary and the result is not visually illustrative. The extra dimension may greatly affect how information can be presented and interpreted, thus moving from a 2-dimension to a 3-dimension representation is not a simple task. One possible disadvantage from the use of three-dimensional objects is occlusion, which may occur if one object covers another or occupies the same spatial position for two coordinates in the 3D representation. This kind of problem usually occurs if the density of data items is large or when a very large object is displayed in front of smaller objects.

2.3.10. Other designs

Through the years, many researchers proposed different kinds of augmented scatterplots aimed at solving specific tasks or problems. In 2008, Elmqvist *et al*. proposed the ScatterDice, a visualization technique designed to explore large and multivariable datasets by navigation in data dimension space using 2D scatterplots, a matrix of scatterplots and 3D transitions [13]. Another interesting tool has been proposed by Chan *et al*. [35], the Sensitivity scatterplot: the idea is to display the data as star glyphs, thus the shape of each dot can represent four different variables at the same time, greatly improving the dimensionality of the scatterplot. Other example of scatterplot customization aimed at improving data perception and/or solve specific tasks are the binned scatterplot [36], the linkable scatterplots [37], the s-CorrPlot [38] and the columns scatterplot [39].

3. RESEARCH METHOD

The analysis detailed in the previous section depicts the ScatterDice as one interesting evolution of the scatterplot, since it is an interactive visualization tool that effectively represents the multiple views' paradigm. At the same time, a classical, single view scatterplot enhanced with additional graphical effects could be more generalist and more flexible, respect to the wide plethora of tasks it could be used for. For these reasons, it could be interesting to compare the ScatterDice with an implementation of the scatterplot that could visualize more dimensions than the two of the basic version and possibly even more

than four since many common solutions display up to four dimensions through the usage of color and size effects. To carry out this comparison, an implementation of a multivariable scatterplot has been developed, whereas the ScatterDice has been only slightly modified for a more realistic comparison.

3.1. Design

Multivariable visualizations try to address the challenges of displaying datasets with many variables. This fact suggests two kinds of problems: firstly, the majority of the charts usually adopted to visualize data cannot display more than three dimensions appropriately; secondly, the efficacy of the graphical effects adopted to represent different data dimensions deteriorates when their number increases. Visual exploration of multivariable data is relevant since it helps to find trends, patterns, outliers, and relationships among variables. When visualizing multivariable data, it is possible to map each variable to some graphical entity or attribute. The scatterplot visualization diagram is considered one of the most functional among the variety of data visual representations, due to its relative simplicity in comparison to other multivariable visualization techniques [18]. Moreover, multivariable visualization tools that feature scatterplots, such as GGobi [40], Tableau/Polaris [41] and XmdvTool [42], usually allow the user to map data dimensions to additional graphical properties such as point color, shape, and size.

Since the basic scatterplot may display only two variables, various techniques have been researched and adopted through the decades to increase its dimensionality. Additional variables may be displayed by correlating them to one or more graphical features of the plotted points, as detailed in section 2.3. It is possible to use simultaneously more than one of these techniques, independently, to obtain even better visual dimensionality. However, the graphical effects must be clearly distinguishable, otherwise the benefits of displaying more dimensions at the same time will promptly worsen due to a reduced visual clarity. Many studies, like [43], have been carried out to understand how visualization design can benefit from taking into consideration perception, as different assignments of visual encoding variables such as color, shape and size could strongly affect how viewers understand data.

Another important feature to take into account when dealing with visualization tools is the fact that different scenarios lead to disparate tasks when dealing with multivariable visualization techniques. As defined by [44] and further described by [45], five major tasks can be considered as objectives a user might want to fulfill when using a visualization tool to display or analyze multivariable data: identify, determine, compare, infer and locate. Scatterplots can be used to assess all these different tasks and have been applied to data in many fields of use, such as automotive, finance, pharmacology, environment, weather forecast, telecommunication, food and many others. The five tasks defined by Valiati are: identify, determine, comparison, infer and locate. Identify refers to any action of finding, discovering or estimating visually. Determine corresponds to the action of calculating, defining or precisely designating values (such as mean, median, variance, standard deviation, amplitude, percentile). Comparison tasks take place when the user wants to compare data that have been previously identified, located, visualized or determined. Infer refers to the action of inferring knowledge from the visualized information, such as defining hypotheses, rules, probabilities or trends, attributes of cause and effect. Locate refers to the actions of searching and finding information in the graphic representation: they can be data points, values, distances, clusters, properties or other visual characteristics.

Moreover, the analysis of the existing visualization tools highlights several utilities that the proposed tool should comprehend to enhance the user experience, such as:

a. the tool should allow the user to input numerical, alphabetical and discrete variables;
b. the tools should enable the user to define filters to simplify the visual exploration of the considered dataset;
c. a zoom function should allow the user to explore the scatterplot, enhancing points evaluation and comparison, especially when occlusion occurs;
d. the user should be enabled to customize the mapping the data dimensions to the available graphical effects;
e. instruments for interactive refinement of the dataset should be available.

To obtain the most possible visual clarity, the visual effects have been chosen and implemented considering both the literature review summed up in Section 2.3 and the analysis hereby proposed. A brief enumeration of the chosen visual effects adopted to enhance the scatterplot consist of:

1. showing a third dimension through a color map; colored points on a scatterplot may suggest similarity among values of the same dataset or correspondence among points of different datasets;
2. varying the size of the points to display an additional dimension; unfortunately, this option may lead to occlusion problems if the plot does not provide proper scaling on the two axes;
3. changing the shape of the points, drawing each element of the dataset as different kinds of glyphs depending on a variable of the dataset;

4. changing the orientation of the shape; usually a dot or line is drawn orthogonally to the perimeter of the shape to better identify the reference point for the orientation;
5. adding to each point a texture, which density may vary depending on a variable;
6. adding to each point an outline, which width may vary depending on a variable;
7. adding a label inside the point, depending on the size of the point itself.

3.2. Multivariable scatterplot

The multivariable scatterplot was developed as a web application using HTML, CSS and JavaScript. The D3.js [12] JavaScript library has been used to both load and manage the input dataset and to provide the visualization. The D3.js library is designed to provide efficient manipulation of documents based on data, supporting large datasets and dynamic behaviors for interaction and animation. Moreover, D3.js makes use of SVG to graphically display the data on the HTML canvas. The algorithm behind the multivariable scatterplot fulfills two core steps: loading the dataset from the source file and displaying it on the screen as shown in Figure 1.

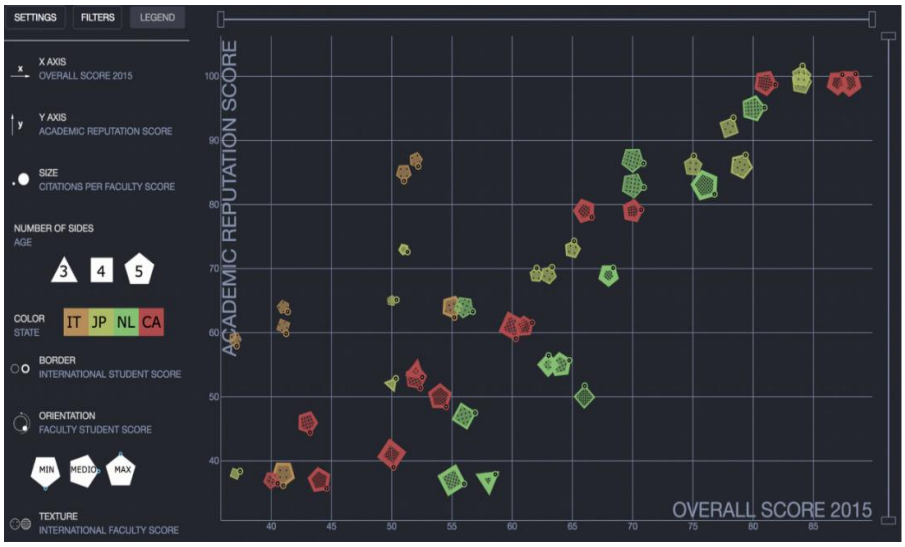

Figure 1. Multivariable scatterplot

The default configuration maps two parameters on the *x* and *y* axis, with a standard size and color for all the point displayed on the chart. The points are displayed with a 30% value of transparency to simplify the understandability of the chart and easily identify overlapped points or occlusion. All the points have a solid border of 1 pixel to distinctly identify each point on the chart. As displayed in Figure 1, the proposed tool, in addition to the scatterplot representation of the dataset, assigns a vertical section on the left of the viewport for the available configuration panels. Moreover, two range bars displayed on the sides of the scatterplot allows zooming on the x and y axis of the chart. The settings panel provides two dropdown menus, one for the list of available visual effects and another one for the parameters. The list of visual effects includes: position on the x or y axis, size of the point, number of sides, color of the point, thickness of the border, orientation and texture. The other dropdown menu lists all the parameters that are available for the current dataset. Through this menu, the user can choose how to map one or more parameters on the different visual effects. Each visual effect, when selected, allows the user to define the default value (except for the x and y axis), such as the default color of the points, the default size and so on. Likewise, when a parameter is mapped on the selected visual effect, it is possible to set the corresponding effect for the minimum and maximum values of the dataset, e.g. the size of the point for the minimum and maximum values of the dataset. Moreover, when a parameter is mapped on a visual effect, other two flag buttons are available for the user: the first one allows to switch the graphical representation of the minimum and maximum value; the second one allows the user to apply the visual effect with an absolute or relative scale. Absolute scale means that even if the user zooms in the chart through the side bars, the effect on each point of the chart will be the same. Relative scale means that the scale for the current graphical effect is applied only to the visible points. This implies a minimum and maximum value will always be displayed on the screen. This option, that should be avoided when looking out for absolute value, becomes useful when performing a comparison

task, since it is easier to distinguish between points with almost identical values, enhancing their differences through the graphical effect. The filter panel allows the user to defined one or more filters. It is possible to define more than one filter for each parameter. The interface allows choosing between range filter and list filter. A range filter defines a set of values that should be included or excluded through a minimum and maximum values, which delimit the range. The delimiter values may be included or excluded from the selection, singularly or both. This kind of filter is often used when dealing with numerical values. A list filter allows to define a specific list of values that should be included or excluded from the visualization. This filter is usually adopted for alphabetical or discrete parameters. The legend panel resumes all the selected mappings with a miniaturized representation of the graphical effect and the name of the current parameter mapped on it.

3.3. ScatterDice

ScatterDice is a visualization technique designed to explore large and multivariable datasets by navigation in data dimension space using 2D scatterplots and a matrix of scatterplots [13]. For each dimension of the dataset, the ScatterDice creates one scatterplot per every combination of dimensions and arrange them in a large scatterplot matrix, used as an overview of the whole dataset. Next to the matrix, a standard bidimensional scatterplot is displayed, representing the currently selected scatterplot from the matrix. 3D rotation visually represents the transitions from one scatterplot to another. Since this feature in not achievable in the multivariable scatterplot, it has been removed for the comparison tests. Furthermore, the user may perform visual queries using bounding volumes and interactively refine a query changing viewpoints. Figure 2 shows an example of ScatterDice visualization.

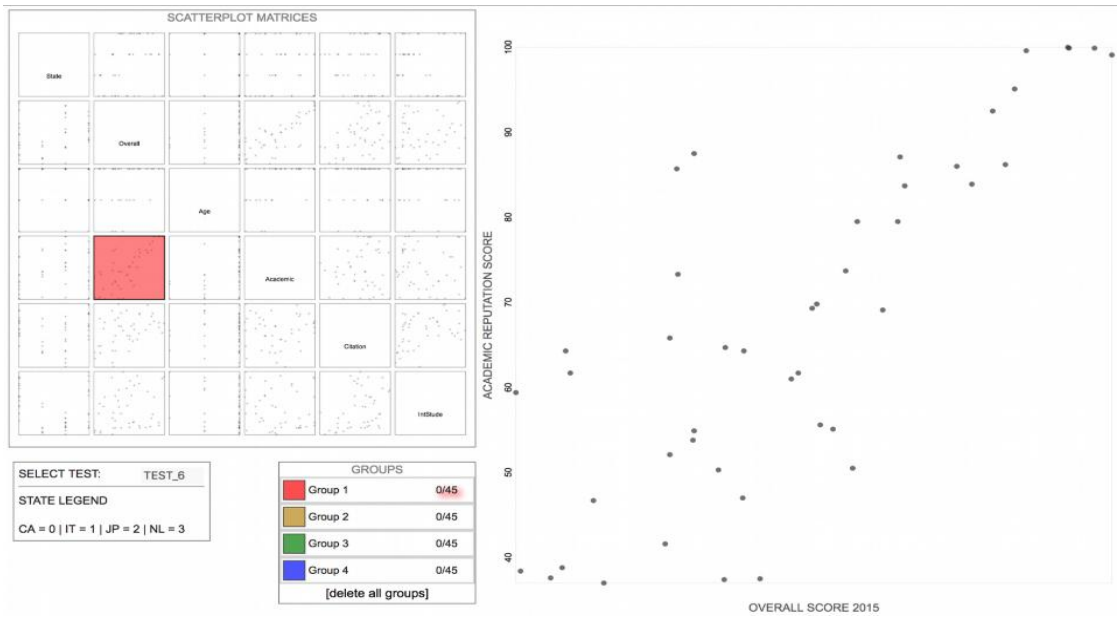

Figure 2. ScatterDice

4. RESULTS AND DISCUSSION

To compare the performance and usability of the multivariable scatterplot and the ScatterDice, a use case has been defined following the methodology described by Valiati *et al*. in [44], in order to address all the different visualization task categories described by their taxonomy. Since the ScatterDice provide a selection tool based on bounding volumes, a similar tool to select relevant point has been developed for the multivariable scatterplot: the use can highlight relevant points by clicking on them, removing the transparency effect and thus allowing the point to stand out.

4.1. Dataset

For the proposed use case, the dataset from QS World University Ranking have been used [46] this is a classic example of multivariable dataset, since for each university are provided fourteen attributes,

some dependent and some independents. For example, the overall score for a university is determined by the following six attributes: academic reputation (weighted 40%), employer reputation (10%), faculty/student ratio (20%), citations (20%), international faculty ratio (5%), and international student ratio (5%). Moreover, other unknown dependencies and correlations among attributes may exist. The other attributes are the overall score for the current year and for the previous year, five discrete values that categorize faculty areas, research intensity, age, size and the State where the university is located.

4.2. Use case

The test consisted of three questions that required the user to perform research in the dataset through subsequent refinements. For each question, the user had to choose among four possible answers. To successfully perform the given task, the user must locate and compare items, identify differences between them, determine specific values and infer the answer. Finally, it may be necessary to visualize the value of a specific attribute through the tooltip. The questions involved only some among the fifteen available attributes: the current year overall score and the six attributes which define it, plus the state and the age attributes. The attributes were mapped to the graphical effects following the perception criteria described in section 3.1. For example, the State attribute was coded as a two characters label, which was displayed inside the shape of the points in the multivariable scatterplot, whereas continuous values were mapped on continuous effects such as the coordinates.

Since the taxonomy defined by Valiati *et al*. defines many possible tasks for each category, the proposed use case involves only a small part of the numerous possible combination of tasks. However, an important achievement is that all the questions involve tasks from almost all the available categories. The "Configure" category was the only one excluded, mainly for two reasons: first of all, the multivariable scatterplot provides a wide set of available customizations, not comparable with the ScatterDice. Secondly, to properly evaluate the usability of the "Configure" tasks, it would have been necessary to extensively use both tools for a long time to acquire a reasonable competence, thus it has been avoided and may be investigated in future works.

For each question, the user had to answer using firstly the multivariable scatterplot and then the modified ScatterDice. Some settings such as the filter panel were disabled to provide a more realistic comparison. The two tools were automatically configured to provide a comparable view and the dataset was pre-filtered to avoid overplotting in the ScatterDice, since a zoom function is not available for this tool.

The first question is hereby described to provide an example of the analysis performed by the user with the two tools. Since the question requires to perform research through subsequent refinements, it has been proposed as a sequence of actions to perform to find the answer:

1. among the Japanese universities with maximum Age value, select those with Overall Score lower than 70;
2. then, exclude the one with the highest International Student Score;
3. then, among the two with the lowest Citation Score, select the one with the highest Academic Reputation Score.

With the scatterplot matrix, the users had to select the most appropriate view to perform the first refinement, selecting the relevant universities through the bounding volumes tool; then, they had to answer to the following requests, selecting the next view to perform the refinement, till they found the answer. The users always had to perform the refinement basing their analysis on the coordinate's position of the points of interest. This was not always easy, since each time the user had to change view, they had to remap the attribute corresponding to at least one axis, or even both.

With the multivariable scatterplot, the users had to focus on the graphical effect corresponding to the first attribute, thus highlighting relevant points through the selection tool, then move on to next attribute, one by one. Depending on the point distribution and on the graphical effect, some users were able to perform the first selection taking into account two attributes at the same time.

4.3. Usability evaluation

At the end of the test, users had to compile a usability questionnaire to provide a qualitative evaluation of the two tools. For the qualitative study we recruited eleven participants (9 male, 2 female) between 22 and 34 years old. They are all students with a background in computer science or engineer, thus with previous knowledge on information visualization or more generally accustomed to graphical representation of datasets.

In a pilot study with two additional participants, we ensured that the overall process runs smoothly and that the tasks are easy to understand. Before performing the test, we checked the participants for color blindness; in one case we needed to verify that the palette used for the color effect was reliable to correctly perform the test.

Then, the multivariable scatterplot, the ScatterDice and the study dataset have been introduced to the testers. The proposed task had different levels of complexity: the first question was easier, with the purpose of making the user comfortable with the visualization tools; the difficulty of the second question was average, whereas the third one was the most difficult. Moreover, the number of visual effects displayed increases in each task, with the first one based on 6 parameters, the second one on 7 and the last one on 8. Participant were instructed to use the first question to familiarize themselves with the software and to ask questions concerning the concept and interactions. Overall, the introduction and warm-up phase took about 10 minutes per subject. We advised the participants to 'think aloud' during the study, to make it simple to identify problems or misunderstandings in the analytical process required to find out the answer of each question. In addition to measuring task completion time for the answers, we took notes on the participants' approaches to the tasks and what problems they encountered. After they had finished all tasks, we gave the subjects a questionnaire with 18 questions, 9 for each tool. Table 1 shows the list of questions and the data gathered through the qualitative evaluation. The questions were designed to evaluate different aspects of the tool through Likert scales. Additionally, we concluded every session by asking open questions to collect detailed feedbacks and suggestions for improvements

Table 1. Questionnaire Results: for Questions 1-8, the Values are Provided on a 4-point Likert Scale, from very few (1) to Very Much (4), whereas the Overall Score is Provided on a 5-point Likert Scale, from very bad (1) to Great (5)

	Questions	Scatter Dice										Mean	Multidimensional Scatterplot										Mean
1	Was it easy to use the tool for the first time?	2	2	4	3	2	3	3	3	2	2	2.6	3	4	3	4	3	4	3	4	4	3	3.5
2	Was it easy to use the tool after the first time?	3	4	4	4	3	4	3	4	3	4	3.6	4	4	3	4	4	4	4	4	4	4	3.9
3	Was it easy to complete the proposed use case?	2	3	4	3	2	2	3	4	3	4	3	4	3	3	4	3	3	3	4	4	4	3.5
4	How much it seemed difficult to you the first task?	4	3	1	1	1	3	3	2	3	2	2.3	2	1	1	1	1	2	3	1	3	3	1.8
5	How much it seemed difficult to you the second task?	3	3	2	1	2	3	3	2	3	2	2.4	1	2	2	1	2	2	3	2	3	3	2.1
6	How much it seemed difficult to you the third task?	2	3	2	1	3	3	3	2	4	2	2.5	1	2	2	1	3	3	2	1	4	4	2.3
7	Was the tool quick to use/explore?	2	2	3	3	3	3	2	3	2	4	2.7	4	3	3	4	4	3	4	4	4	4	3.7
8	Does the tool required a high cognitive demand to be used?	2	3	2	3	3	3	2	3	4	2	2.7	1	3	2	3	2	2	2	2	3	2	2.2
9	Provide an overall score to the tool	3	3	4	3	3	4	4	5	4	5	3.8	5	4	4	4	4	5	4	5	5	5	4.5

4.4. Test results

In terms of task completion, results show that the multivariable scatterplot is slightly better that the ScatterDice: with the first tool, only one user was unable to provide the right answer for the most difficult question, whereas another user provided a wrong answer to the easiest question. In one cases, the error was due to the selection tool, in the other one the user simply forgot to select one value in the selection step, thus affecting the following refinements. With the second tool, three users were unable to provide the right answer for the first question, suggesting that maybe multiple views present a more abrupt learning curve, and two users provided the wrong answer to the second question. However, all the errors were due to the fact that the

users made error performing the selection, thus forgetting one or more relevant points. Overall, the users were unable to provide the right answer using the multivariable scatterplot only in two cases out of thirty, whereas with the ScatterDice it happened five times (one out of six). From a usability point of view, as showed in Table 1 the multivariable scatterplot was evaluated better than the ScatterDice in every question, especially in terms of speed and cognitive load, even if the gap between the two visualization tools was very low for questions number two and number six. Overall, even if the testers' group was too small to obtain results with statistical validity for all the proposed questions, performing the T-Test for questions one, seven and nine proved the null hypothesis that the multivariable scatterplot is better than the ScatterDice.

The feedback provided by the user for the multivariable scatterplot was very positive and eventually they proposed some improvements: most of them were related to configuration functionalities, already available but disabled for the comparison, whereas some suggestions regarded the graphical effects adopted. Based on the users' feedbacks, the most significant problem when using the scatterplot matrix was that the mapping of the attributes on the axes was different on each view.

5. CONCLUSION

In this paper, a single view and a multiple views scatterplot for displaying multivariable datasets have been compared. The analysis of the state of the art allowed to define a set of guidelines to support the design and development of the multivariable scatterplot: a single view visualization tool which implements many graphical effects to increment the number of attributes visible at the same time on a scatterplot. The aim of the comparison was to investigate if such a tool could provide an understandable and functional representation of the data, comparable to the multiple views' paradigm. The chosen use case allowed to test the multivariable scatterplot for different visualization tasks, trying to cover all the task categories proposed by Valiati *et al*. in their taxonomy. Finally, a qualitative evaluation of the proposed tools pointed out that not only they are comparable but also that the multivariable scatterplot is more appreciated by the users. Future works may include further assessments of the multivariable scatterplot with different datasets since the distribution of values for the different data dimensions may greatly affect the visualization. Moreover, it would be possible to develop additional graphical effects, for the visualization of both static and dynamic data. Finally, further tests may point out which graphical effects perform better together and which ones are most suited for specific tasks in order to provide pre-selected configurations.

REFERENCES

[1] Pflaum A.A., Golzer, P., "The IoT and Digital Transformation: Toward the Data-Driven Enterprise," *IEEE Pervasive Computing*, vol. 1(1), pp. 87-91, Jan 2018.

[2] Rghioui A, Oumnad A., "Internet of Things: Surveys for Measuring Human Activities from Everywhere," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7(5), pp. 2474-82, Oct 2017.

[3] Sacha D., Stoffel A., Stoffel F., Kwon BC., Ellis G., Keim DA., "Knowledge Generation Model for Visual Analytics, " *IEEE transactions on visualization and computer graphics*, vol. 20(12), pp. 604-13, Dec 2014.

[4] Sindhu CS, Hegde NP, "A Novel Integrated Framework to Ensure Better Data Quality in Big Data Analytics over Cloud Environment," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7(5), pp. 2798-805, Oct 2017.

[5] Sarikaya A, Gleicher M., "Scatterplots: Tasks, Data, and Designs," *IEEE Transactions on Visualization & Computer Graphics*, vol. 1(1), pp. 402-12, Jan 2018.

[6] Saket B., Endert A., Demiralp C., "Task-Based Effectiveness of Basic Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, May 2018.

[7] Zhao Y., Luo F., Chen M., Wang Y., Xia J., Zhou F., Wang Y., Chen Y., Chen W., "Evaluating Multi-Dimensional Visualizations for Understanding Fuzzy Clusters," *IEEE transactions on visualization and computer graphics*, 2018.

[8] Anscombe FJ., "Graphs in Statistical Analysis," *The American Statistician*, pp.17-21, 1973.

[9] Chambers J, Cleveland W, Kleiner B and Tukey P., "Graphical Methods for Data Analysis," Wadsworth, Ohio, pp. 128-129, 1983.

[10] Milman I., Pilyugin VV., "Interactive Visual Analysis of Multidimensional Geometric Data," *WSCG*, 2016.

[11] Lamberti F., Manuri F., Sanna A., "Multivariate Visualization Using Scatterplots," *Encyclopedia of Computer Graphics and Games*, pp. 1-2, 2017.

[12] D3.js, Data Driven Document, https://d3js.org/

[13] Elmqvist N., Dragicevic P., Fekete JD., "Rolling the dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation," *IEEE transactions on Visualization and Computer Graphics*, vol. 14(6), pp. 1539-148, 2008.

[14] Wong PC, Bergeron RD., "Multivariate Visualization using Metric Scaling. InProceedings of the 8th Conference on Visualization'97," *IEEE Computer Society Press*, pp. 111-ff, Oct 1997.

[15] Dos Santos S., Brodlie K., "Gaining Understanding of Multivariate and Multidimensional Data Through Visualization," *Computers & Graphics*, vol. 28(3), pp. 311-25, Jun 2004.
[16] Van Wijk JJ, van Liere R., "Hyperslice Visualization of Scalar Functions of Many Variables," *Department of Computer Science [CS]*, (R 9449), Jan 1994.
[17] Van Liere R., Van Wijk JJ., "Visualization of Multi-Dimensional Scalar Functions Using Hyperslice," *Centrum voor Wiskunde en Informatica*, Jan 1994.
[18] Tufte ER., Schmieg GM., "The Visual Display of Quantitative Information," *American Journal of Physics*, vol. 53(11), pp. 1117–1118, 1985.
[19] Cleveland WS., McGill R., "The Many Faces of a Scatterplot," *Journal of the American Statistical Association*, vol. 79, pp. 807-822, 1984.
[20] Tufte ER., "The Visual Display of Quantitative Information," 1983.
[21] Wells NA., "Quantitative Evaluation of Color Measurements: I. Triaxial Stereoscopic Scatter Plots," *Sedimentary Geology*, vol. 151(1-2), pp. 1-15, 2002.
[22] Reese A., "Scatterplots Revisited. Significance," vol. 5(2), pp. 87-89, 2008.
[23] Chae M., Lee Y., Lee J., "Sawtooth Bubble Chart".
[24] Fiaschi D., Gianmoena L., Parenti A., "Local Directional Moran Scatter Plot-LDMS," p. 197, 2015.
[25] Noma E., "Heuristic Method for Label Placement in Scatterplots," *Psychometrika*, pp. 463-468, 1987.
[26] Cleveland WS., "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, vol. 74, pp. 829-836, 1979.
[27] Eilers PH., Goeman JJ., "Enhancing Scatterplots With Smoothed Densities," *Bioinformatics*, vol. 20(5), pp. 623-628 2004.
[28] Janetzko H., Hao MC., Mittelstädt S., Dayal U., Keim D., "Enhancing Scatter Plots using Ellipsoid Pixel Placement and Shading," *46th Hawaii International Conference on System Sciences (HICSS)*, pp. 1522-1531, 2013.
[29] Mayorga A., Gleicher A., "Splatterplots: Overcoming Overdraw in Scatter Plots," *IEEE transactions on visualization and computer graphics*, vol. 19(9), pp. 1526-1538, 2013.
[30] Chen H, Chen W., Mei H., Liu Z., Zhou K., Chen W., Gu W., Ma KL., "Visual Abstraction and Exploration Of Multi-Class Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20(12), pp. 1683-1692, 2014.
[31] Staib J., Grottel S., Gumhold S., "Enhancing Scatterplots with Multi- Dimensional Focal Blur," *Computer Graphics Forum*, vol. 35(3), pp. 11-20, 2016.
[32] Bachthaler S., Weiskopf D., "Continuous Scatterplots," *IEEE transactions on visualization and computer graphics*, vol. 14(6), pp. 1428-1435, 2008.
[33] Bachthaler S., Weiskopf D., "Efficient and Adaptive Rendering of 2- D Continuous Scatterplots," *Computer Graphics Forum*, Blackwell Publishing Ltd, vol. 28(3), pp. 743-750, 2009.
[34] Fisherkeller MA., Friedman JH., Tukey JW., "Prim9, an Interactive Multidimensional Data Display and Analysis System," *Dynamic Graphics for Statistics*, pp. 91–109, 1988.
[35] Chan YH., Correa CD., Ma KL., "The Generalized Sensitivity Scatterplot," *IEEE transactions on visualization and computer graphics*, vol. 19(10), pp.1768-1781, 2013.
[36] Hao MC, Dayal U., Sharma RK., Keim DA., Janetzko H., "Variable Binned Scatter Plots. Information Visualization," vol. 9(3), pp. 194-203, 2010.
[37] Nguyen QV, Simoff S, Qian Y, Huang ML., "Deep Exploration of Multidimensional Data with Linkable Scatterplots," *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, 2016.
[38] McKenna S., Meyer M., Gregg C., Gerber S., "s-corrplot: An Interactive Scatterplot for Exploring Correlation," *Journal of Computational and Graphical Statistics*, vol. 25(2), pp. 445-463, 2016.
[39] Ogino K., Oishi A., Oishi M., Gotoh N., Morooka S., Sugahara M., Hasegawa T., Miyata M., Yoshimura N., "Efficacy of Column Scatter Plots for Presenting Retinitis Pigmentosa Phenotypes in a Japanese Cohort," *Translational vision science & technology*, vol. 5(2), pp. 4-4, 2016.
[40] Swayne DF., Lang DT., Buja A., Cook D., "Ggobi: Evolving from Xgobi Into an Extensible Framework for Interactive Data Visualization," *Computational Statistics & Data Analysis*, vol. 43(4), pp. 423-444, 2003.
[41] Stolte C., Tang D., Hanrahan P., "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8(1), pp. 52–65, 2002.
[42] Ward MO., "Xmdvtool: Integrating Multiple Methods for Visualizing Multivariate Data," *Proceedings of the IEEE Conference on Visualization*, pp. 326–333, 1994.
[43] Demiralp Ç., Bernstein MS., Heer J., "Learning Perceptual Kernels for Visualization Design," *IEEE transactions on visualization and computer graphics*, vol. 20(12), pp. 1933–1942, Dec 2014.
[44] Valiati ER., "Taxonomia De Tarefas Para Técnicas De VisualizaçÃo De InformaçÕes Multidimensionais," *PPGC/UFRGS*, 2005.
[45] Pillat RM., Valiati ER., Freitas CM., "Experimental Study on Evaluation of Multidimensional Information Visualization Techniques," *In Proceedings of the 2005 Latin American conference on Human-computer interaction* ACM, pp. 20-30, Oct 2005.
[46] QS World University Rankings, https://www.topuniversities.com/university-rankings/.