Many-to-Many Relational Parallel Coordinates Displays

Mats Lind*, Jimmy Johansson† and Matthew Cooper†

* Department of Information Science, Uppsala University, Sweden
† Norrköping Visualization and Interaction Studio, Linköping University, Sweden
mats.lind@dis.uu.se, {jimmy.johansson, matt.cooper}@itn.liu.se

Abstract

An interesting property of the commonly used parallel coordinates display is the distinct overall pattern formed by the totality of lines between adjacent axes. These patterns have a direct correspondence to the type of relationship existing between the variables mapped onto the axes in question as well as a salient visual appearance. Parallel coordinates displays can therefore be used to visually investigate relationships between variables as well as investigating individual objects/lines. The problem with this approach is that, whereas each object is mapped in its entirety in a standard parallel coordinates display, only a small subset of the interrelations between variables is shown as the number of variables increase. To show all possible relations between variables multiple parallel coordinates displays are needed. In turn this means that each variable is duplicated several times, once per extra parallel coordinates display. To a viewer this increases the visual complexity and most probably the mental load. To aid users we have devised a new configuration of the axes in multiple parallel coordinates displays. Through an experiment we have also started to investigate the usability of this new configuration and the results are promising.

Keywords—**Parallel coordinates, multidimensional visualization, visual data mining, usability and user studies.**

1 Introduction

An interesting property of the commonly used parallel coordinates display [5] is the distinct overall pattern formed by the totality of lines between adjacent axes. These patterns have a direct correspondence to the type of relationship existing between the variables mapped onto the axes in question. Parallel coordinates displays can therefore be used to visually investigate relationships between variables as well as investigating individual objects/lines. This is especially useful in situations where the investigator has no prior knowledge of the types of relationships existing in the data since the patterns formed are especially indicative of this; positive linear relationships have a very different appearance from negative ones and both of these look different again from power functions etc. This was first described by Wegman [16] and examples of such patterns for common types of relationships are shown in figure 1.

Thus, visually investigating raw data from a set of variables by means of parallel coordinates displays has its distinct advantages to investigating pre-processed data. Such pre-processing usually involves some kind of ad hoc assumption about the nature of relationships. The validity of correlation coefficients, for instance, is based on the assumption that relationships are linear in nature.

The problem with using parallel coordinates displays to investigate relationships between variables is, however, that only a subset of the interrelations between variables is shown when more than two variables are of interest. This is in contrast to when parallel coordinates displays are used to investigate objects, each object being mapped in its entirety by a line in a display. A solution to showing all possible relations between variables is the use of several parallel coordinates displays simultaneously. This approach of course means that each variable is duplicated several times, once per extra parallel coordinates display. To a viewer this increases the visual complexity and most probably adds to the mental load.

The number of parallel coordinates displays needed to show all possible relations is given by the formula $\lfloor \frac{N+1}{2} \rfloor$ where N is the number of variables of interest [16]. So, for instance, having three or four variables require two parallel coordinates displays, having seven or eight variables require four such displays, and so on. In light of this, the goals of the work presented here have been:

- to find new forms of parallel coordinates displays that maintain the power of parallel coordinates displays to reveal hidden structures while minimizing the mental load induced by the need to duplicate variables.

- to begin an investigation of the usability of the found display type through a user study comparing some aspects of its efficiency with those of standard parallel coordinates by means of a controlled experiment.

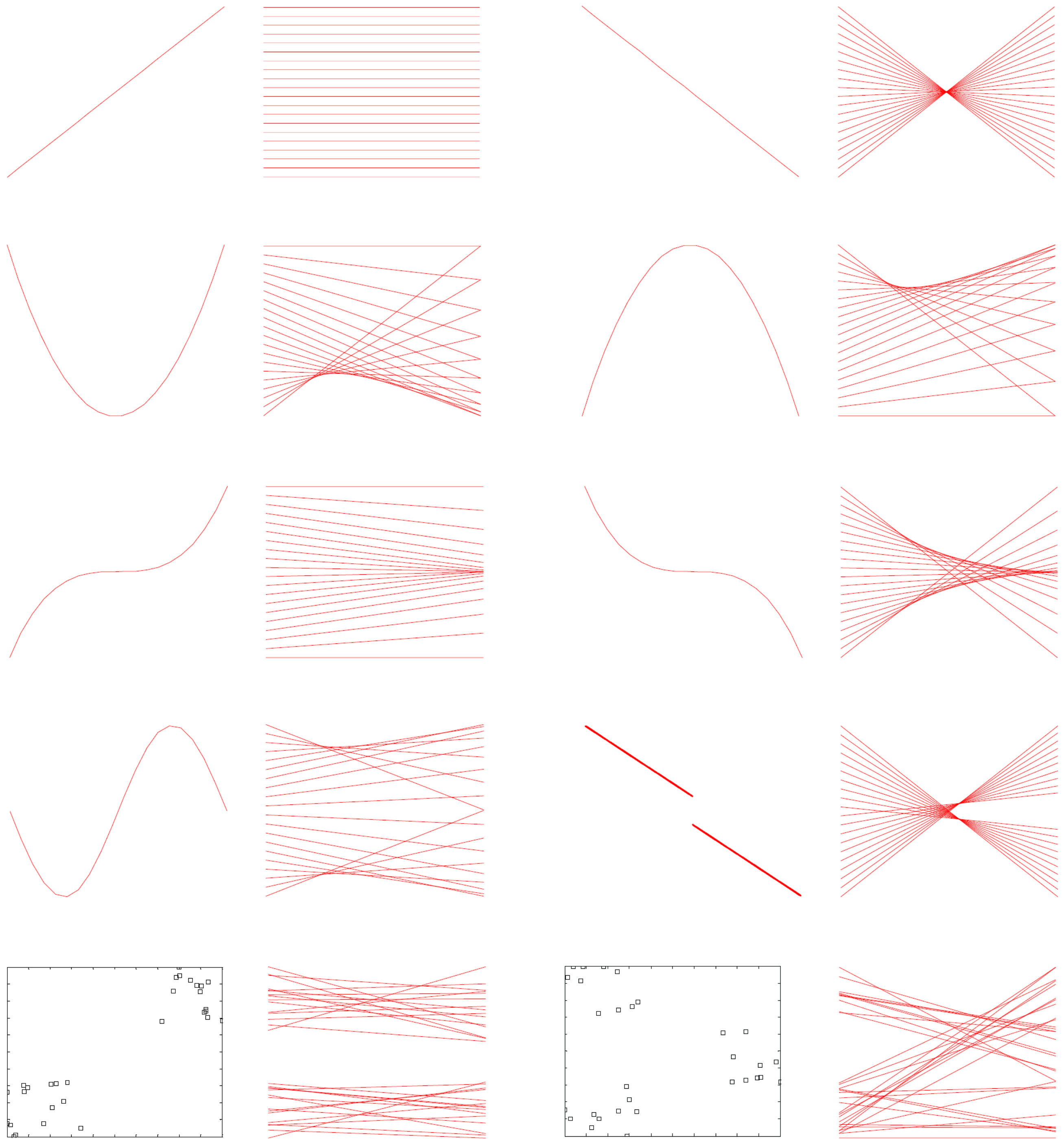

Figure 1: An illustration of some of the visually salient patterns formed by common relationships. The leftmost graph in each column shows the relationship between two variables in a standard Cartesian coordinate system while that to the right shows the resulting pattern in a parallel coordinates display. The first row thus illustrates positive and negative linear relationships, the second row quadratic relationships and the third row cubic relationships. The left column of the fourth row illustrates a sinusoidal relationship while the right column of that row illustrates that any discontinuities in a functional relationship results in a gap in the pattern. The fifth row illustrates the appearance of groups of scattered points.

2 Background and Related Work

When faced with the challenge of analysing data where each item contains values of a large number, tens or even hundreds, of variables, there is a problem with exploring the enormous number of inter-variable relationships which may be present in the data. One possibility is to reduce the number of variables using a dimensionality reduction approach. Many such techniques exist and common ones are multidimensional scaling, principal component analysis and self-organizing maps. While this approach is effective in extracting the most significant features of the data set, it has the drawback that the original data and the individual relations between variables are not shown and so exploring the relationships is impossible.

Direct visualization techniques which are designed for exploring relationships, such as parallel coordinates, remove the need for dimensionality reduction since the multivariate data set can be visualized using only a 2D display. The order in which variables are mapped to axes, however, gives completely different visual representations of the data and only relationships between adjacent axes can be directly analysed. One approach to deal with this is to order the axes according to some criterion [1, 17, 10] which it is hoped will highlight the most significant relationships. Depending on the task, this criterion could be chosen in order to reduce cluttering in the display or to position the axes such that they have the highest correlation with their neighbours. This approach is effective but has the drawback that it could lead the user to overlook less distinct, but equally important, relationships that should be brought to the user's attention.

Instead of ordering the axes in the parallel coordinates display, alternative methods can be used to permit all possible relations to be displayed. As mentioned previously, a first approach to this was presented by Wegman [16] where an algorithm is described to automatically permute the axes in a number of standard parallel coordinates displays so that all relationships present are displayed in the minimum number of axis arrangements. Using this algorithm, the resulting displays are shown either one after another or simultaneously. This shows the complete set of relationships but either the user must remember the relationships between displays or must refocus on the many instances of each variable scattered in different places. This dramatically decreases its usefulness since, in either case, the short term memory load on a user will be high when trying to get an overview over a data set containing more than a few variables.

A quite different approach is to extend the parallel coordinates displays into a 3D display. Wegenkittl et al. [15] introduced extruded parallel coordinates and 3D parallel coordinates to visualize dynamical systems. A similar view, the 'Cube', was presented by Falkman [3] for visualizing clinical data. Tominski et al. [13] introduced a 2D radial parallel coordinates layout that can be used to simultaneously visualize the relationships between a single focus variable and all other variables. Johansson et al. [6] extended the radial layout to a 3D display. This 3D layout has been shown to be superior to standard 2D parallel coordinates, particularly in revealing more complex multiple relationships in the multivariate data [4, 7].

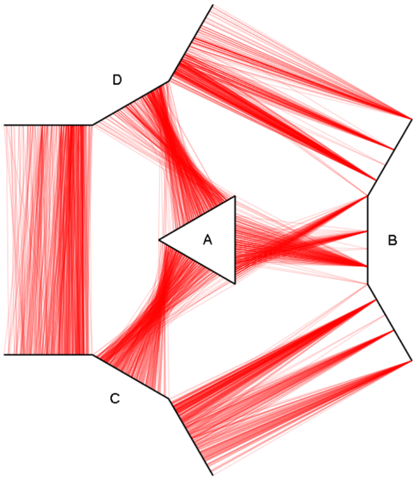

Figure 2: The basic four variable arrangement of the new proposed visualization. Here illustrated with data from the cars data set [2]. Variable A refers to MPG (miles per gallon), B number of cylinders, C horsepower and D weight.

3 The New Proposed Visualization

As stated above, the main weakness of multiple simultaneous standard parallel coordinate displays is that each variable/axis is presented multiple times and spatially separated. In turn this requires users to sequentially fixate on many parts of the visualization keeping many bits of information in short term memory during this exploration. However, in that individual objects/lines are no longer of primary interest, other spatial arrangements of the axes than in standard parallel coordinates displays are possible. In particular we can re-arrange the axes in a multiple simultaneous standard parallel coordinates display such that the axes associated with one variable are placed adjacent to one another. By doing this the number of axes needed remains the same or increases, and, indeed, this is a logical necessity, but the number of visual replications of the actual variables decreases. Looking at the case of four variables, shown in figure 2, illustrates this in its simplest form.

Here the number of axes is 12 compared to 8 which is the logical minimum number but each variable is represented only once instead of twice, as would be the case in a standard set up. Of course, the number of axes could have been reduced to 8 but this would have led to multiple lines from each axis, making the patterns visually interfere with one another. Therefore we have chosen to use the condition of non-overlapping lines when constructing these visualizations. Unfortunately, using this condition and aiming for non-duplicated variables, the four variable arrangement in figure 2 depicts the maximum number of variables that can be shown. However, these shapes tessellate, expanding both the number of axes and the number of relationships which can be included in the display but at the cost of replicating variables and relationships in the display, see figure 3. The replication of variables is, however, modest in size and the replicated relationships can be removed, which makes identifying the number of positive, negative or other relationships easier, though at the cost of some loss of symmetry, see figure 4. Depicting seven variables in this manner requires six of them to be duplicated once. To show the same amount of interrelations in a standard parallel coordinate display configuration, four such displays are needed in which four of the variables will exist in four copies and three in three copies. This is illustrated in figure 5. It should be noted that replicating variables in a parallel coordinates display is not a new concept. This has previously been attempted in [11] but with a different aim—to show correlation between more than two axes.

Our proposed seven variable configuration also has some other interesting properties:

- One variable is not replicated, placed in the centre of the display and can be used as a variable of central importance.
- The replicated variables exist in both a triangular form and a rounded form, that is, the different shapes of the representations could help a user to navigate in the visualization.
- The replicates are placed on opposite sides of the central variable making orientation in the visualization easy with practise.

We have not so far investigated further possibilities to tessellate the basic pattern in order to present an even larger number of variables in this manner. The main reason for this is that we want to be certain of the practical usefulness of this complex layout. Already the seven variable version is complex enough to merit a usability examination and a first user study is presented below. Also, it is easy to envisage a potentially very effective system for visual data exploration using these visualizations even if it is restricted to the case where a user, from a larger set, could choose subsets of seven variables to investigate at a time.

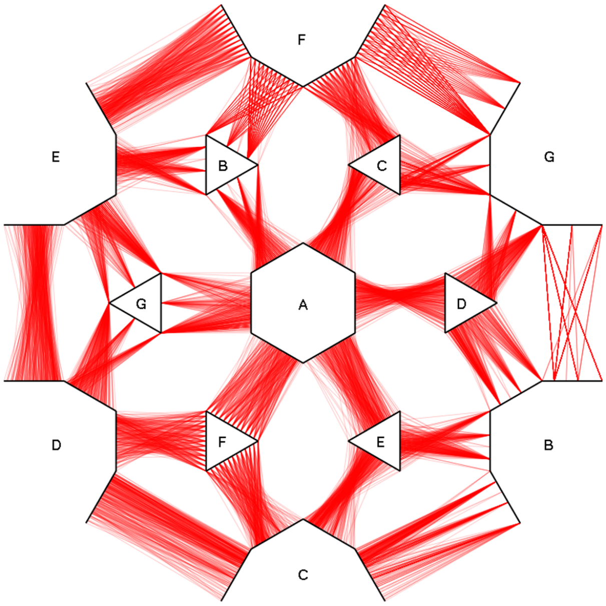

Figure 3: The seven variable arrangement found by tessellating the basic four variable pattern. This results in three relations being duplicated. The data used for the illustration is again from the cars data set with variable assignments as in figure 2. The added three variables refer to: E acceleration, F year and G origin.

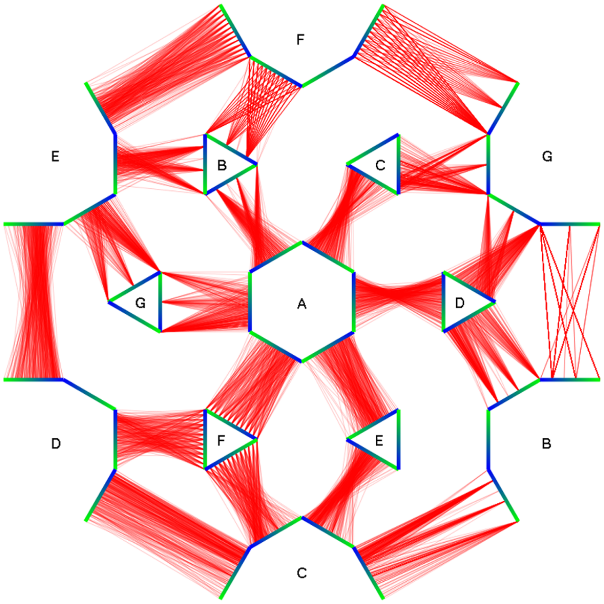

Figure 4: The same relations as in figure 3 but here with redundant relationships removed and the axes colour coded using blue to green to indicate the direction of increasing values on each axis.

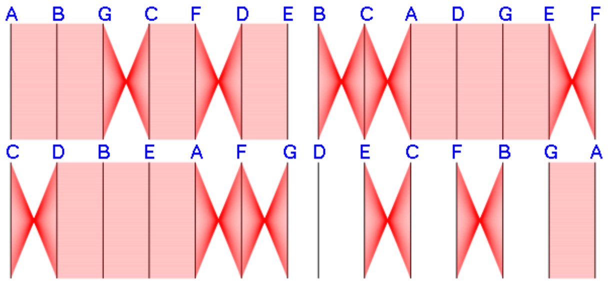

Figure 5: The set of four standard parallel coordinates displays needed to depict all possible interrelations between seven variables. Redundant pairings have been omitted.

4 A First Usability Evaluation

The most interesting question from a usability point of view is whether the reduced number of repetitions of the variables in our displays will allow users to more efficiently find interesting relationships between variables compared with the use of multiple standard parallel coordinates displays. In order to start an investigation of this we devised a simple experiment.

4.1 Method

Since a possible difference in detectability between different complex relationships was not the main question, we used only two relations: a positive linear relation and a negative linear relation (see figure 1). We devised our data set such that only two of the seven variables had negative relations with five other variables. These variables will be referred to as the target variables. The subjects' task was to identify these target variables as correctly and quickly as possible. Half of the subjects were first shown 21 instances of our proposed seven variable display where each combination of two variables were the target variables once. The other half of the subjects were first shown 21 instances of multiple standard parallel coordinates displays depicting the exact same set of relations as when our proposed type of displays were used. After these 21 trials the two groups of subjects switched conditions and performed the task with the other type of display. A total of twelve subjects performed the task and the design was a simple one factor within subject design. The order of presentation between the two conditions was counter balanced and the order between the 21 stimuli was randomly selected for each subject.

4.2 Subjects

The twelve subjects were all students or staff at the University of Uppsala and none of them had any prior knowledge of parallel coordinates displays, either in its standard form or in our version. They were aged from 22 to 35 years with a median age of 27 years and had all normal or corrected to normal vision. Seven of the subjects were male and five female.

4.3 Apparatus and Viewing Conditions

All displays were shown on a 17" TFT monitor connected to a PC running Microsoft Windows with an OpenGL capable graphics card. Each subject was seated about 60 cm in front of the screen. The physical size of the two visualizations was kept the same in terms of area although they had different outer shapes; the multiple standard parallel coordinate displays were more rectangular than our visualizations as can be inferred from figures 3–5. The experiment program was constructed using the Psychophysics Toolbox [9] and Matlab.

4.4 Procedure

Each subject was first greeted and shown to their seat. The test leader, who had not taken any part in the work leading up to the new proposed visualizations, then gave the subject a one page written and illustrated set of instructions for the task. There were two such pages, one for each of the condition and the texts were as identical on both pages as possible. After the subject had read the instructions they were asked to run a four trial demo version of the program for the condition at hand. This was done before both conditions. The program was self paced. To start a trial the subject clicked any mouse button and, as soon as an answer had been reached, hit the space bar on the keyboard. The time for one trial was thus considered to be the time between the mouse click and the space bar hit. As the subject hit the space bar the visualization disappeared and a neutral white screen was shown with a text telling the subject to please give their answer.

The instructions specifically told the subjects that the most important aspect of their session was that they should look at the displays until they were certain of the answer, that is to produce as few errors as possible. Given that, they were also told that they should work as fast as possible.

4.5 Results

First the error data were analysed. Errors were scarce and evenly distributed between the two conditions. The median number of error per subject was 0 in our new visualization condition (this condition will in the future be referred to as the 'many-to-many' condition) and the total number of errors (over all $12 \times 21 = 252$ trials) was 9. In the standard parallel coordinates condition, in the future referred to as the 'standard' condition, the median number of errors was 1 and the total number of errors 11. In that the errors were few and evenly distributed between the conditions, the times will be indicative of the relative efficiency of these two types of visualizations for the designated task. Since reaction time data are typically non-normally distributed, the recorded times were first transformed by determining the base-ten logarithm of each of

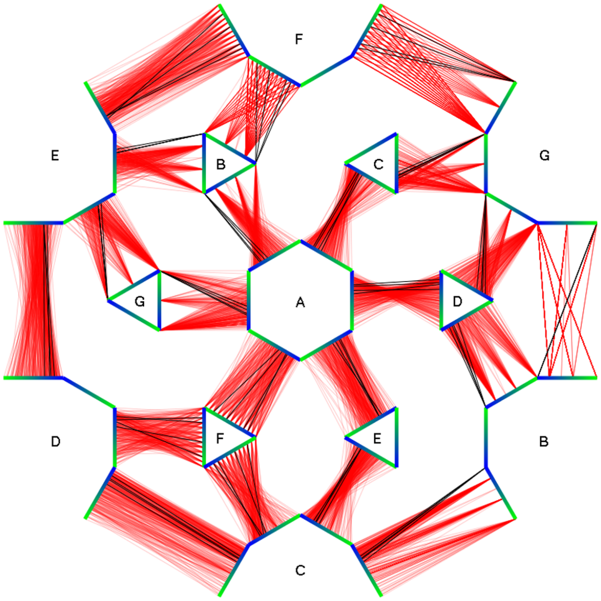

Figure 6: The seven variable cars data set. Here with all 3-cylinder cars highlighted by colouring their respective lines black.

them. All subsequent analyses were then performed on these transformed data.

For each subject the mean time over all 21 stimuli in a condition was first calculated. This was taken as an indicative value of the subject's performance. Thus 24 values were obtained, twelve in each condition. The mean time in the many-to-many condition, transformed back to seconds, was 9.8 seconds. In the standard condition, the mean time, here too transformed back to seconds, was 12.3 seconds. Thus, the subjects performed about 20% faster in the many-to-many condition. This difference was tested by means of a t-test using a decision criterion of 0.05. The difference was significant ($t = -2.613$,N=$11, p < 0.05$).

Interviews with some of the subjects after the experiment indicated that this difference might be an underestimation of the difference for more general types of stimuli. As can be seen in figure 5, two negative relations mapped onto two adjacent axes in a standard parallel coordinates display happen to form a very distinct 'meta' pattern. In our extremely simplified stimuli, a subject could look for the occurrences of these meta patterns and in that way solve the task without having to look at all axes. Still, however, performance in the standard condition was worse than in the many-to-many condition.

5 Discussion

A number of issues need to be addressed before we can claim that our proposed visualization is the most effective way of handling multivariate data in certain visual exploration tasks. For instance they need to be compared to standard scatter plots in a matrix layout, a technique that has been around since the 1980's (see for instance [14, 12]). Such a comparison needs to address the issue of how discriminable different types of relations are in scatter plots in comparison to standard parallel coordinates displays in general as well as the problem studied in the experiment reported here; how well different types of visualizations help a user to compare relations defined over several variables. However, the saliency of the patterns formed by different types of relationships between variables provide a great advantage for parallel coordinates displays over scatter plots in general. Evaluating the strength of a relationship, given that the type is known, is probably, at least in many types of application areas, best done by analytical methods rather than visual. As can be seen in figure 1 different types of relationships, especially for non-linear and non-functional relationships, produce quite different patterns. This reasoning is also valid in light of recent results [8] comparing parallel coordinates displays to scatter plots indicating that scatter plots are visually more efficient. Our main concern is to aid users discover types of relationships between variables, not to visually evaluate the strength of the relationships. Here it should be noted that our evaluation had a slightly different focus. The two types of relationships used were trivial and would easily be discovered using scatter plots. Our main concern was instead to see whether our proposed visualization better supported overall visual navigation compared with multiple instances of regular parallel coordinates displays and whether the non-horizontal parallel coordinates display would be too difficult to visually analyse.

Although our many-to-many relational parallel coordinates display was primarily intended to aid in the visual exploration of relationships between variables, they can also be used to investigate objects or groups of objects. Figure 6 illustrates this.

One of the main limitations of the proposed visualization method is obviously that it, at least in its present form, does not effortlessly scale to an arbitrary number of variables. Furthermore, the variable placed in the centre of the plot plays a special role in that it is the only one that is not duplicated, at least in the seven variable version. These circumstances point directly to that the main value of our proposed visualization would be in the context of a larger visualization system where a user can choose from a set of visualization options, each targeted to different data exploration needs. The role of the presently proposed visualization would then be to allow a user to scrutinize in more detail a set of four or seven variables that by other means have been judged by a user to be particularly interesting.

Perhaps even a specific variable has been diagnosed by a user as the main center of concern and it would then be a candidate for the central position in our seven variable plot. The nature of a usable design for such interplay between our proposed visualization and other visualization methods is, however, unknown at present and is therefore left for further research.

Acknowledgements

The authors wish to thank Camilla Forsell for assistance with the experiment.

References

[1] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *IEEE Symposium on Information Visualization 1998*, pages 52–60,153, 1998.

[2] A. Asuncion and D. J. Newman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[3] G. Falkman. The use of a uniform declarative model in 3D visualisation for case-based reasoning. In *6th European Conference on Advances in Case-Based Reasoning*, pages 103–117, 2002.

[4] C. Forsell and J. Johansson. Task-based evaluation of multi-relational 3D and standard 2D parallel coordinates. In *SPIE-IS&T Electronic Imaging*, volume 6495, pages 64950C–1–12, 2007.

[5] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91, 1985.

[6] J. Johansson, M. Cooper, and M. Jern. 3-dimensional display for clustered multi-relational parallel coordinates. In *Proceedings 9th IEEE International Conference on Information Visualization*, pages 188–193, 2005.

[7] J. Johansson, C. Forsell, M. Lind, and M. Cooper. Perceiving patterns in parallel coordinates: Determining thresholds for identification of relationships. *Information Visualization*, 7(2):152–162, 2008.

[8] J. Li, J.-B. Martens, and J. J. van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization advance online publication, May 1, 2008; doi:10.1057/palgrave.ivs.9500179*.

[9] D. G. Pelli and L. Zhang. Accurate control of contrast on microcomputer displays. *Vision Research*, 31(7–8):1337–1350, 1991.

[10] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings IEEE Symposium on Information Visualization 2004*, pages 89–96, 2004.

[11] H. Theisel. Higher order parallel coordinates. In *Proceedings of the 5th Fall Workshop on Vision, Modeling, and Visualization*, pages 415–420, 2000.

[12] S. H. C. Du Toit, A.G.W. Steyn, and R.H. Stumpf. *Graphical Exploratory Data Analysis*. Springer-Verlag, 1986.

[13] C. Tominski, J. Abello, and H. Schumann. Axes-based visualizations with radial layouts. In *Proceedings ACM Symposium on Applied Computing*, pages 1242–1247, 2004.

[14] P.A. Tukey and J.W. Tukey. Graphical display of data sets in three or more dimensions. In V. Barnett, editor, *Interpreting Multivariate Data*, pages 189–213, 1981.

[15] R. Wegenkittl, H. Löffelmann, and E. Gröller. Visualizing the behavior of higher dimensional dynamical systems. In *Proceedings IEEE Visualization 1997*, pages 119–125, 533, 1997.

[16] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85(411):664–675, 1990.

[17] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *IEEE Symposium on Information Visualization 2003*, pages 105–112, 2003.