Evaluation of Serial Periodic, Multi-Variable Data Visualizations

Alexander Mosolov
13705 Valley Oak Circle
Rockville, MD 20850
(301) 340-0613
AVMosolov@aol.com

Benjamin B. Bederson [i]
Computer Science Department
Human-Computer Interaction Lab
3171 A.V. Williams Building
University of Maryland
College Park, MD 20742
bederson@cs.umd.edu

ABSTRACT

In this paper, I present the results of an evaluation of the effectiveness of a new technique for the visualization and exploration of serial periodic data. At this time, the only other visualization to support this task is the "Spiral" by Carlis and Konstan [1], which an issue with space usage that I attempt to address– namely, the data points on the fringes of the spiral are sparse and the data points towards the middle are crowded. My solution to this is to use a grid-like structure, where space is used is uniformly throughout, and no space is wasted. I have conducted a study to compare the effectiveness of several variations of the grid approach of looking at multiple variables simultaneously, and the findings of this study are discussed.

Keywords

Information visualization, DataGrid, Grid, Serial Periodic Data, Multi-Variable Data, Data Exploration, Evaluation, User Study.

INTRODUCTION

Serial periodic data is data that has both serial and periodic properties – the most obvious example is time-based data, where time continually moves forward (the serial aspect), and there are cycles of days, weeks, months, etc. (the periodic aspect). The DataGrid is an attempt to enable the user to find periodicity in their data, as well as see other pertinent information once the period has been found. In the DataGrid visualization, this data is displayed in rows and columns, similar to the way the days are arranged on a calendar. The exploration of the data is done through interactively varying the number of data points displayed in each row, thus varying the period. When the period displayed gets close to a period present in the data displayed, we see a telltale diagonal pattern (see Figures 1-3). When the period that we're displaying the data with matches a periodicity inherent in the data we're exploring, we see a vertical pattern emerge. See Figure 4 for an example of what the results look like when a month's worth of daily light data, taken at 15-minute intervals, is displayed with a period of 24 hours.

Note that the periodicity of the data is not the only thing revealed – it's also clear that the light intensity is going up each day (the red is brighter towards the bottom), and the day is getting longer (the red column is getting slightly wider towards the bottom). These additional observations make sense in light of the fact that the data displayed is for January, in the Northern hemisphere, when this is exactly what's supposed to be happening.

EXPLORING MULTIPLE VARIABLES

When only one variable is displayed, each of the small rectangles seen in Figures 1-4 corresponds to one reading of a single variable for a point in time, with the intensity of the color reflecting the value of the variable. DataGrid also allows the user to look at up to 3 variables simultaneously. Each additional variable is displayed using a different color (red and green for 2, red green and blue for 3 variables). There are 5 different ways of combining the different variables on the screen.

- Diagonal – each rectangle is split diagonally, and the portion allocated for each variable is colored accordingly.
- Horizontal – each rectangle is split horizontally, and each section colored accordingly.
- Vertical – same as horizontal, but the rectangle is split vertically.
- Color Blend– each rectangle's color is the blending of the red, green, (and possibly blue, for 3 variables) color components of each variable.
- Multiple Views (MV) – displays, one below another, 2 to 3 single variable views, all of which are controlled simultaneously.

The effectiveness of these different methods is evaluated by the study that is outlined below. See Figures 5, 6, 7, 8, and 9 for examples of what each method looks like.

STUDY DESCRIPTION

The study conducted was relatively small (10 subjects total). Each subject was asked to perform a variety of tasks on each of 10 different datasets. Five datasets were 2-variable, and 5 were 3 variable. For each dataset, each

subject had to use only one of the 5 possible visualization methods. The methods were staggered across datasets and users in such a manner as to ensure that the ease or difficulty of performing the tasks on a particular dataset did not affect the outcome of the study. The "Multiple Views" method was used as a baseline to compare other methods against, as it does not attempt to combine multiple variables in the same space.

Tasks:
The subject will have two tasks to perform for each data set:
- correctly identify the period, if any, of each variable that is being displayed
- identify the relationship, if any, between the variables that are being displayed

Datasets:
Pseudo-randomly generated data, suited for the study.

The following 2-variable datasets were generated:
1) same period for both variables, variables directly related.
2) same period for both variables, variables inversely related
3) same period for both variables, variables are not related
4) different periods for each variable, not related
5) one variable is periodic, one isn't

Three-variable datasets were the same as 2-variable sets (although randomly generated again, but with the same patterns), with an extra, unrelated variable added in. The goal was to try to measure the effect of this extra "clutter" introduced by adding another variable to the display.

Measured Variables:
- The time the subjects take to complete each task.
- The correctness of the answer (binary, either correct or incorrect).

STUDY RESULTS
The Mann-Whitney Test was used to analyze the gathered data for statistical significance. See Tables 1-7 for detailed results of the test. The following comparisons were made, with the following results:

Test I
The time taken to find the period of the first variable was compared, for every method, against the time taken by the Multiple Views method. For 2-variable data sets, the MV method performed significantly better than all the other methods, except for Horizontal, where the difference was not significant. For 3-variable data sets, the MV method performed worse than all the other methods except for Horizontal – so Horizontal actually did relatively worse

than for 2 variables, but these differences were not statistically significant.

Test II
The total time taken to perform all the tasks was compared, for every method, against the time taken by the MV method. For both 2 and 3-variable datasets, the MV method outperformed its counterparts, however the difference was only significant in one case, when it was compared vs. the Diagonal method on 3-variable datasets.

Test III
The time taken to identify any relationship between the displayed variables was compared, for every method, against the time taken by the MV method. In all cases except one, the other methods outperformed MV, but the difference was not statistically significant. The exception was with the Vertical method for 3-variable datasets, where MV outperformed it, but also not significantly.

Test IV
The time taken to find the period of the first variable in a 2-variable dataset was compared, for every method, against the time taken by the same method to do the same task for a 3-variable dataset. All the differences were statistically insignificant, however, notably, the 2 largest ones were for the Horizontal and MV methods.

The correctness of the subjects' answers was not analyzed for significance, as the fraction of incorrect answers turned out to be extremely small.

ANALYSIS OF STUDY RESULTS
Test I shows that while, for 2 variables, MV clearly outperforms the other methods, for 3 variables, the other methods actually slightly outperform it. This is probably due to the fact that as the number of variables goes up, the space allotted for each variable in the MV method goes down. This indicates sharing the given space between multiple variables becomes more efficient than simply splitting the space up, as the number of variables displayed goes from 2 to 3. It seems likely that this trend would continue, and become more pronounced, as the number of variables is increased. Test I also indicates that the Horizontal is adversely affected by the increase in the number of variables, compared to the other methods except MV.

Test II shows MV outperforms, though mostly insignificantly, all the other methods on the total time taken to complete all the tasks. Looking at the data, I think this is due to the fact that the first variable was always periodic, and the other 2 weren't always so. A lot of time was generally taken by subjects to identify that something wasn't periodic, and this was much more clear

in the MV view. The extra time was usually spent "making sure" that there really isn't a pattern there, whereas for MV it was very clear. However, since the subjects generally felt that there wasn't a pattern, and were just trying to make sure that was the case, it's reasonable to suppose that with more experience with using the other methods, they would be more comfortable identifying something as non-periodic.

Test III shows MV slightly outperformed by all methods except for Vertical at the task of identifying relationships between variables. I think this is because the users were able to glean extra information the variable relationships while trying to identify individual variable periods in the methods where the space was shared – in fact, many times the subjects identified the variable relationships immediately. With MV, the users gained no extra information from identifying the periods, and looking for relationships was a whole new task to them. The Vertical method tended to introduce a lot of confusion, because the vertical splitting of the rectangles inadvertently introduced a lot of vertical patterns that made vertical patterns due to periodicity harder to find. It also made variables appear to be inversely related, as the different colored vertical lines appeared side by side (See Figure 10).

Test IV, though it did not produce statistically significant results, seems to indicate that the Horizontal and MV methods suffered most from the clutter introduced by adding a third variable. For MV, this is consistent with the results from Test I – that MV is more affected by the reduction of available space for each variable than other methods are by being forced to fill the shared space with more variables. For Horizontal, I think this is due to the fact that adding more horizontal lines per rectangle increases the vertical separation between values of the same variable, making vertical patterns harder to spot. This is also consistent with Test I's results.

USER FEEDBACK
The subjects seemed to be excited about using the visualization tool, and largely enjoyed the process of completing the tasks, in particular when they were able to quickly spot patterns in the data they were working with. The Vertical method seemed to cause a lot of confusion, and the study results bear that out somewhat. Many subjects commented that the Color Blend method was hard to use, as they weren't quite sure which colors combine to create which. However, despite that, the Color Blend method did quite well. I think that's because, even if mentally someone isn't quite sure which color combinations form which, they just have an intuitive sense for it – for example, someone looking for red would be more likely to look at yellow (red + green) instead of cyan (blue + green). In fact, one user

commented, "This is pretty cool. I just think 'red', and I see the pattern… I didn't even notice the other colors".

CONCLUSION
So, which of the methods is better? What are any of these methods good for? Obviously, the data under consideration needs to either be known to be periodic, or needs to be evaluated for periodicity. If only 2 variables need to be displayed, then Multiple Views is probably the best choice. The Horizontal method is a close second. For 3 variables, Color Blend and Diagonal seem to be the best choices – they maintain a sense of vertical continuity, like Vertical, but, unlike Vertical, don't introduce false patterns. For 4 or more variables, Color Blend isn't an option, which leaves Diagonal. Its effectiveness for that many variables would need to be explored more, but it shows some promise.

REFERENCES
1) Carlis and Konstan. Interactive Visualization of Serial Periodic Data. *ACM Symposium on User Interface Software and Technology* (1998), 29-38.
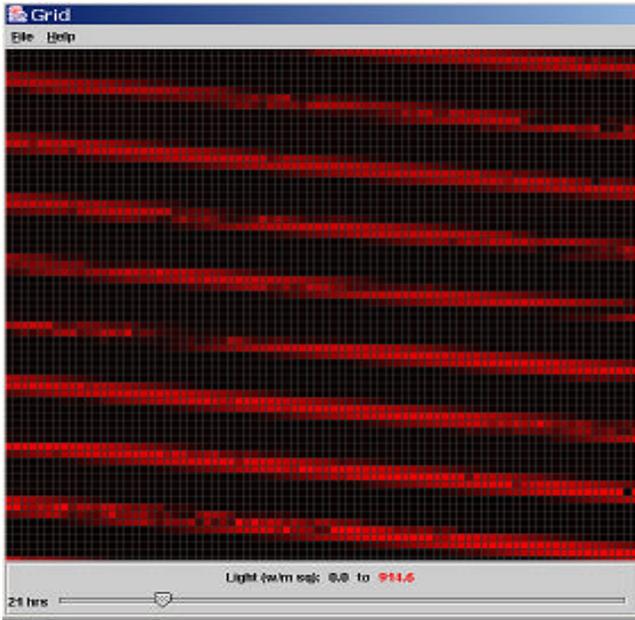
APPENDIX A: FIGURES

Figure 1: Light information for January 2000, displayed with a 21-hour period. We see a very slanted diagonal pattern.
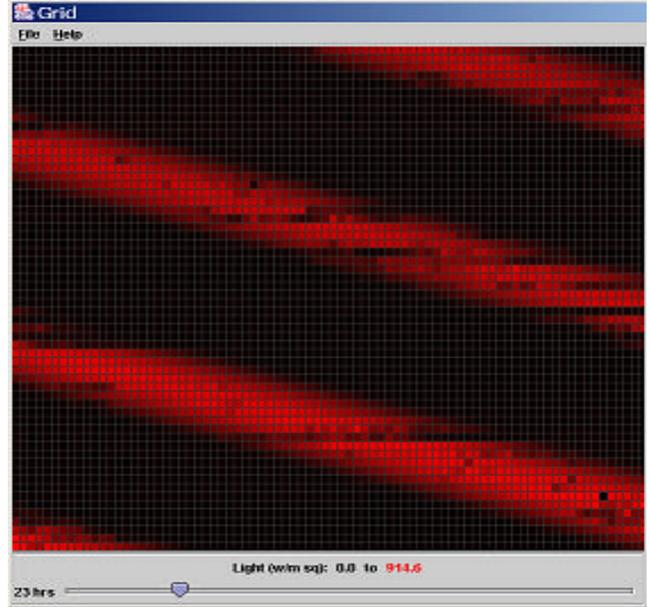

Figure 3: Light information for January 2000, displayed with a 23-hour period. The diagonal pattern becomes less and less slanted as we get closer to the period of the variable.
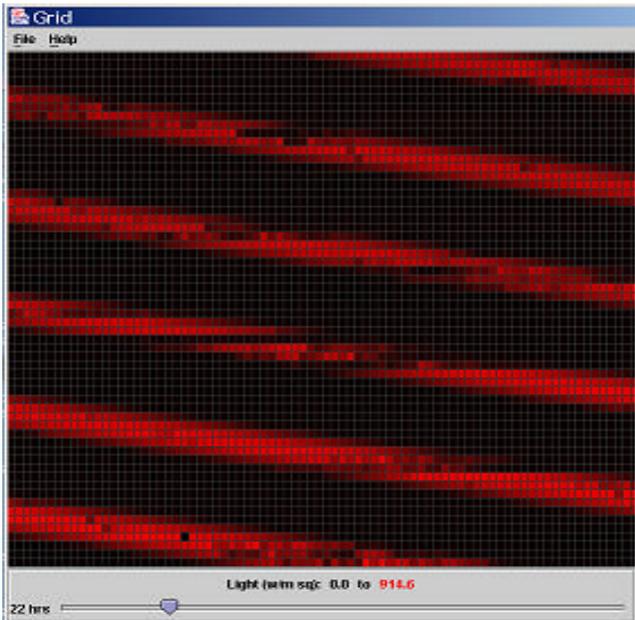

Figure 2: Light information for January 2000, displayed with a 22-hour period. The diagonal pattern is a little less slanted.
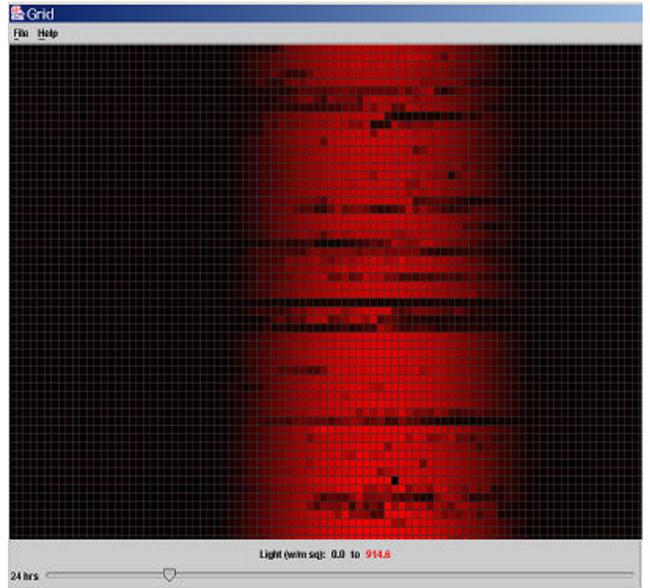

Figure 4: Light information for January 2000, displayed with a 24-hour period. The vertical pattern we see indicates that we have found the period of the variable (which, in this case, was obviously 24 to begin with).
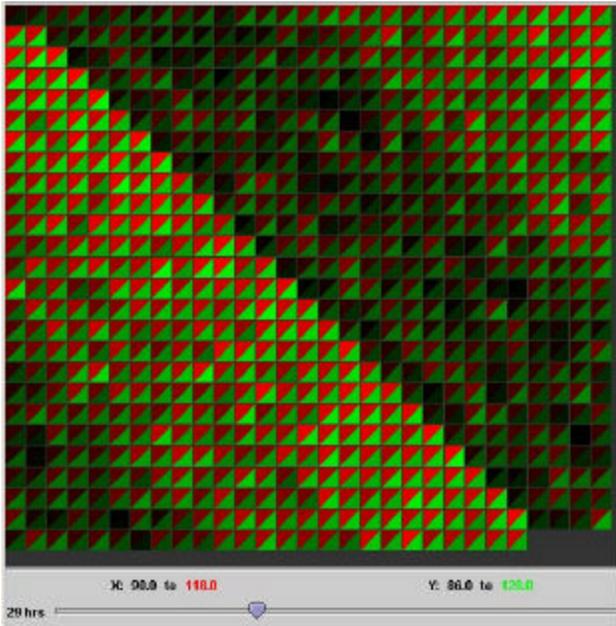
Figure 5: Diagonal method, 2 variables. A vertical pattern is about to emerge for both red and green, both of which have the same period.
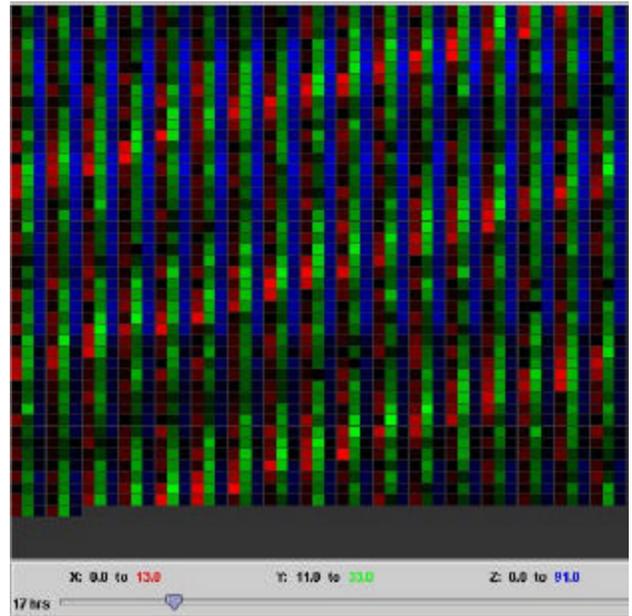
Figure 7: Vertical method, 3 variables. A vertical pattern is about to emerge for both red and green, which share the same period. Note that although blue is non-periodic, we can see definite vertical strips of it.
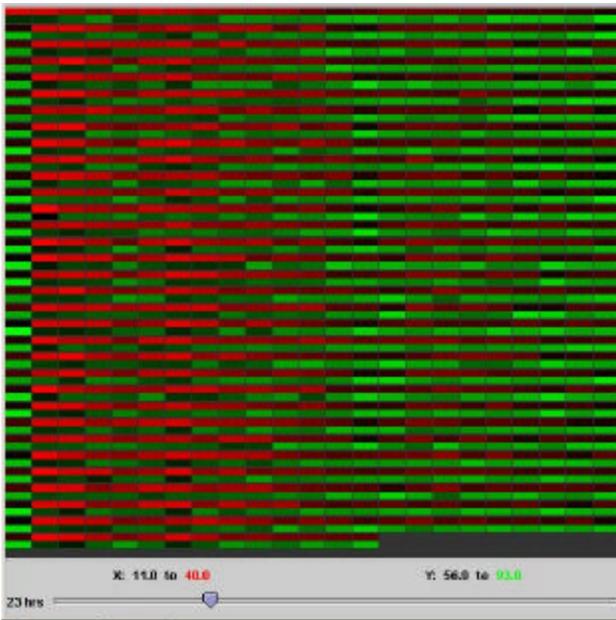
Figure 6: Horizontal method, 2 variables. A vertical pattern has emerged for both red and green, which are inversely related. Note that there is some vertical discontinuity for both colors. This becomes worse in the 3 variable case.
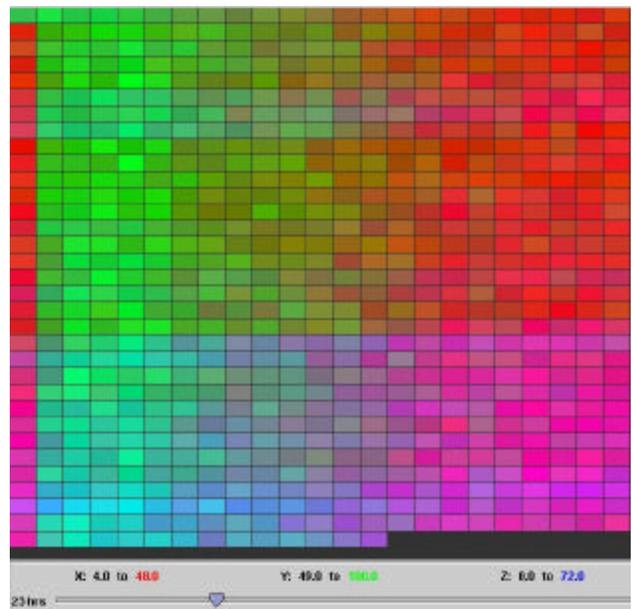
Figure 8: Color Blend method, 3 variables. We see red and green have the same period (which is currently displayed), and are inverses. Blue, which is non-periodic, doesn't appear to introduce much clutter.
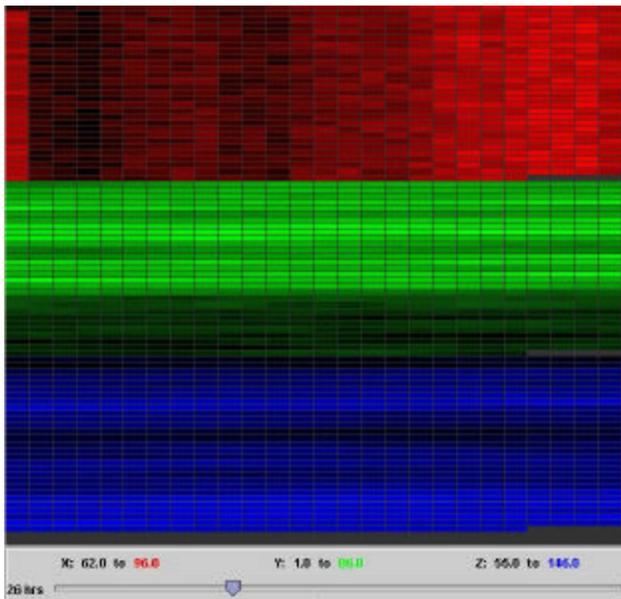
Figure 9: Multiple Views method, 3 variables. We are at the correct period for red. Green and blue are non-periodic – note the telltale absence of diagonal lines in either.
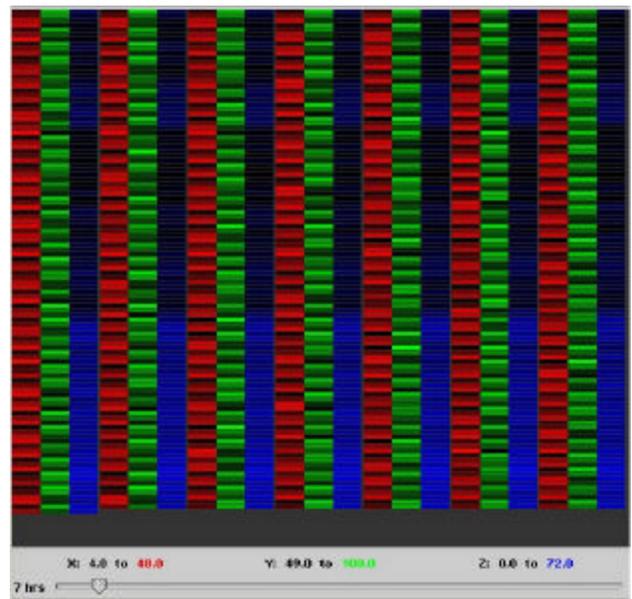
Figure 10: Vertical method, 3 variables. We are not at the correct period for any variable, but we see strong vertical patterns. Also, the variables (falsely) appear to alternate, creating the impression that there is some inverse relationship.

APPENDIX B: TABLES

The Z value is the confidence interval. Z >=1.96 means that a finding is significant with a confidence level of 95%.

Tables for Test I

MV vs…	Z value	Significance Level
Diagonal	-2.00	.05
Horizontal	-.72	Not significant
Vertical	-2.31	.05
Color Blend	-2.00	.05

Table 1: MV vs all other methods, time to identify period of 1st variable, for 2 variable datasets. Negative Z values indicate MV performing better.

MV vs…	Z value	Significance Level
Diagonal	0.11	Not significant
Horizontal	-0.42	Not significant
Vertical	0.04	Not significant
Color Blend	1.25	Not significant

Table 2: MV vs all other methods, time to identify period of 1st variable, for 3 variable datasets. Negative Z values indicate MV performing better.

Tables for Test II

MV vs…	Z value	Significance Level
Diagonal	-1.93	Not significant
Horizontal	-.94	Not significant
Vertical	-1.40	Not significant
Color Blend	-1.47	Not significant

Table 3: MV vs all other methods, total time to perform all tasks for 2 variable datasets. Negative Z values indicate MV performing better.

MV vs…	Z value	Significance Level
Diagonal	-2.08	.05
Horizontal	-1.40	Not significant
Vertical	-1.25	Not significant
Color Blend	-1.25	Not significant

Table 4: MV vs all other methods, total time to perform all tasks for 3 variable datasets. Negative Z values indicate MV performing better.

Tables for Test III

MV vs…	Z value	Significance Level
Diagonal	0.86	Not significant
Horizontal	0.79	Not significant
Vertical	0.56	Not significant
Color Blend	0.23	Not significant

Table 5: MV vs all other methods, time to identify relationships between variables in 2 variable datasets. Positive Z values indicate MV being outperformed.

MV vs…	Z value	Significance Level
Diagonal	.56	Not significant
Horizontal	1.09	Not significant
Vertical	-.34	Not significant
Color Blend	.18	Not significant

Table 6: MV vs all other methods, time to identify relationships between variables in 3 variable datasets. Positive Z values indicate MV being outperformed.

Tables For Test IV

2 vs 3 var	Z value	Significance Level
Diagonal	.94	Not significant
Horizontal	-1.47	Not significant
Vertical	.26	Not significant
Color Blend	.49	Not significant
Mult.Views	-1.47	Not significant

Table 7: 2 variable vs. 3 variable times to identify period of 1st variable, for each method vs. itself. Negative Z values indicate that the method performed better on 2 variable datasets.

APPENDIX C: RAW TIME DATA

Below is the raw data gathered during the study. Data on the correctness of the subjects' answers is not included, as the vast majority of the answers were correct, and there didn't appear to be any correlation between the time it took to answer a question and the answers' correctness.

Times to identify period of 1st variable, for 2 variable datasets.

Method:	Times:									
Diagonal	34	120	7	224	82	100	4	23	18	12
Horizontal	24	90	8	9	79	11	8	8	3	2
Vertical	35	60	7	9	34	21	18	73	12	43
ColorBlend	80	123	16	14	26	10	12	12	12	22
Mult.Views	34	21	5	2	5	8	5	5	10	39

Times to identify period of 1st variable, for 3 variable datasets.

Method:	Times:									
Diagonal	10	75	10	50	10	29	13	71	17	10
Horizontal	14	50	8	30	27	20	31	13	53	10
Vertical	10	10	6	45	8	21	180	105	12	27
ColorBlend	12	13	15	41	13	13	14	50	12	5
Mult.Views	30	27	17	67	16	11	21	19	4	15

Times to finish analyzing 2 variable datasets:

Method:	Times:									
Diagonal	57	235	15	345	207	120	9	34	20	65
Horizontal	36	123	10	16	118	34	17	15	5	57
Vertical	51	120	9	20	84	44	21	75	33	70
Color Blend	125	190	22	24	40	21	25	18	14	51
Mult. Views	454	54	7	9	9	22	7	7	12	109

Times to finish analyzing 2 variable datasets:

Method:	Times:									
Diagonal	11	240	11	87	73	105	185	131	141	84
Horizontal	15	175	31	140	49	71	92	16	147	93
Vertical	12	130	23	107	22	113	317	146	101	34
Color Blend	22	370	32	116	60	88	133	86	14	46
Mult. Views	35	80	34	89	34	51	28	37	10	44

Times to find relationship between variables, 2 variables datasets:

Method:	Times:									
Diagonal	11	1	1	1	1	44	20	1	3	1
Horizontal	11	1	1	1	1	10	1	1	1	5
Vertical	17	1	1	5	1	26	1	1	1	4
Color Blend	6	1	1	10	1	23	1	9	1	2
Mult. Views	9	1	1	34	1	68	1	5	1	2

Times to find relationship between variables, 3 variable datasets:

Method:	Times:									
Diagonal	35	1	5	1	49	15	3	2	1	4
Horizontal	13	5	1	1	5	9	1	11	2	3
Vertical	11	105	6	13	7	5	1	14	5	1
Color Blend	49	69	5	5	1	30	1	6	3	2
Mult. Views	31	4	15	4	1	5	2	10	10	7

---

[i] This work was originally started as a class project for an Information Visualization course taught by Prof. Bederson at the University of Maryland.