Evaluating an InfoVis Technique Using Insight Reports

Markus Rester[1], Margit Pohl[1], Sylvia Wiltner[1],
Klaus Hinum[2], Silvia Miksch[3],
Christian Popow[4], Susanne Ohmann[4]
[1] Institute of Design and Assessment of Technology, Vienna University of Technology
[2] Institute of Software Technology & Interactive Systems, Vienna University of Technology
[3] Department of Information & Knowledge Engineering, Danube University Krems
[4] Department of Child and Adolescent Neuropsychiatry, Medical University of Vienna
{markus, margit}@igw.tuwien.ac.at

Abstract

The evaluation of Information Visualization (InfoVis) techniques can help to identify specific strengths and weaknesses of these methods. The following article describes the results of an empirical study assessing the contribution of an interactive InfoVis method based on a spring metaphor (GRAVI), Exploratory Data Analysis (EDA) and Machine Learning (ML) to ease understanding. The application domain is the psychotherapeutic treatment of anorectic young women. The three methods are supposed to support the therapists in finding the variables which influence success or failure of the therapy. To conduct the evaluation we developed a report system which helped subjects to formulate and document in a self-directed manner the insights they gained when using the three methods. The results indicate that the three methods are complementary and should be used in conjunction.

Keywords—**Explorative Information Visualization, Evaluation, Insight Reports**

1 Introduction

Several authors have pointed out the importance of evaluation studies of Information Visualization (InfoVis) techniques (see e.g. [1, 2, 8, 17]). In the past few years usability studies concerning visualization methods have become more frequent, and valuable information about the design of such systems has been gathered. Nevertheless, as Spence [15] mentions, there is still too little systematic information about the specific strengths and weaknesses of the features of InfoVis techniques. On the basis of existing evidence it is still difficult to decide which InfoVis technique to use for which purpose. Therefore, evaluation studies are especially important to give the developers some insights into the usefulness of a given InfoVis technique for the intended area of application.

The following study describes an investigation in how best to support psychotherapists in their work. The aim of these therapists is to analyze the development of anorectic young women taking part in a psychotherapy. During this process a large amount of highly complex data is collected. Statistical methods are not suitable to analyze these data because of the small sample size, the high number of variables and the time-dependent character of the data. Only a small number of anorectic young women attend a therapy at one time. The young women and their parents have to fill in numerous questionnaires before, during and after the therapy. In addition, progress in therapy is often not a linear process but a development with ups and downs. All of this indicates that InfoVis techniques might be a better method of analysis of these data. The aim of the therapists is to predict success or failure of the therapy depending on the results of the questionnaires, and, more generally, to analyze the factors influencing anorexia nervosa in more detail.

We tested three possibilities how to support the therapists' work: an InfoVis technique specifically developed for this purpose, Exploratory Data Analysis (EDA) and Machine Learning (ML). The InfoVis technique which is called Gravi++ (GRAVI, for legibility) was developed in cooperation with the psychotherapists and reflects their requirements. EDA seems to be an interesting alternative because of its exploratory nature. Machine Learning might also yield interesting results because of its computational power. These three methods were compared, and the results of this comparison are described in the following text. Our original assumption was not that one of these techniques might be the best but rather to find the specific strengths and weaknesses of these methods and how these methods could best

be combined.

The investigation of the InfoVis technique GRAVI was conducted in two stages. We distinguished between the usability study (which was conducted first) and a study of the technique as such. It is well-known that a good InfoVis technique might be rejected because of usability problems of the concrete implementation. Therefore, we solved the usability problems of the software first (see [11]). A similar approach was adopted by North [4]. In the second phase we assessed the three techniques mentioned above. Main results gained in the second phase of the investigation will be reported below.

2 Related Work

GRAVI is based on a spring metaphor. Icons for the questions from the questionnaires are positioned on a circle. Other icons for the anorectic young women are arranged within this circle depending on the strength of attraction of the single questions. The questions function, to a certain extent, like magnets or springs. The final position of an icon for the anorectic young women is a combination of the forces of all questions (see Figure 1). Similar InfoVis techniques have already been developed (see e.g. [6]). There is also some recent literature describing empirical investigations of similar techniques (see [20, 7]). Although there is some similarity of GRAVI to these techniques, there are also noticeable differences. GRAVI has very specific features for interaction with the system. It is possible to visualize dynamic, time-dependent data. It also combines a visualization based on a spring metaphor with other visualization methods (e.g. a star glyph).

It is mentioned by Yi, et al. [20] that the spring metaphor makes it easy to understand their visualization technique. They describe occlusion as one of their biggest problems. This is also an issue with GRAVI, although we found a possible solution for this problem. Yi, et al. [20] also point out that such visualization techniques might not be appropriate for people who cannot formulate questions. This is no problem for GRAVI because the target population (psychotherapists) are professionals in their field. Pillat, et al. [7] posit that each visualization technique has different advantages. They found out that the identification of clusters and the visualization of general features of the dataset are the advantages of visualization techniques based on the spring metaphor.

2.1 Evaluation in the InfoVis Area

In his position paper for the Beliv'06 workshop, Stasko [16] points out that the evaluation of information visualizations is a challenging task, especially if the goals of these techniques are not straightforward. Many evaluations of InfoVis techniques use task completion times and error rates as the only variables tested. In an ill-structured domain with no clear-cut results like psychotherapy other approaches are necessary. In such a domain it is often difficult to decide whether a result is "true" or "false". The definition of mental health in psychotherapy, e.g., is highly controversial. Furthermore, getting valid insights will take up a lot of time and efficiency in a traditional sense is not an issue. Therefore Saraiya, et al. [14] suggest "insight" as an outcome variable. We found this approach also very valuable. So far, there are no general frameworks for categorizing insights. So we developed our own classification system which is highly dependent on the tasks our subjects had to solve. Nevertheless, we think that it should be possible to develop a more generic framework for insight classification because it seems to be plausible that insights like clustering or finding detailed, factual information will be necessary for many exploratory InfoVis techniques. This is certainly an area for future research.

3 GRAVI

Users can interact with GRAVI [3] in several ways. The most basic form of interactivity is positioning icons on the screen. These represent the patients and questionnaires they answered. According to the answer a patient gave to a question, the patient's icon is attracted by the question's icon. This leads to the formation of clusters of patients who gave similar answers (see Figure 1). The therapists are especially interested in those variables which predict the outcome of the therapy (successful or not successful). By analyzing clusters of "positive" and "negative" cases they can identify those variables.

GRAVI can also represent dynamic, time dependent data. It uses animation and traces to show the paths of the patients' icons over all time steps. The position of the patients' icons change over time. This allows analyzing and comparing the changing values. The therapists need this

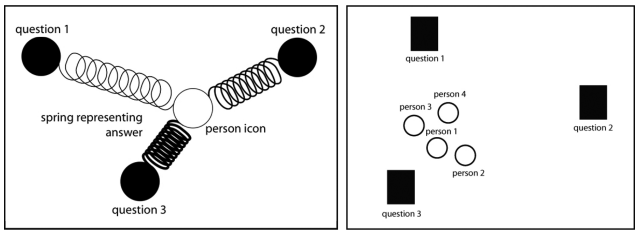

Figure 1. GRAVI Concept of Spring-Based Positioning (Left), Leading to Formation of Clusters (Right)

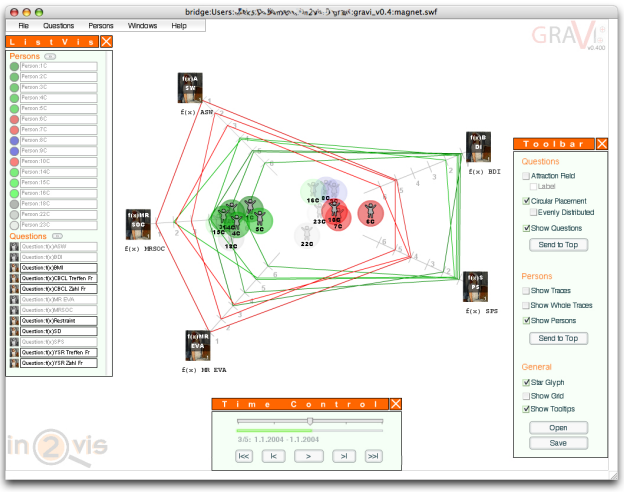

Figure 2. Typical Screenshot of GRAVI

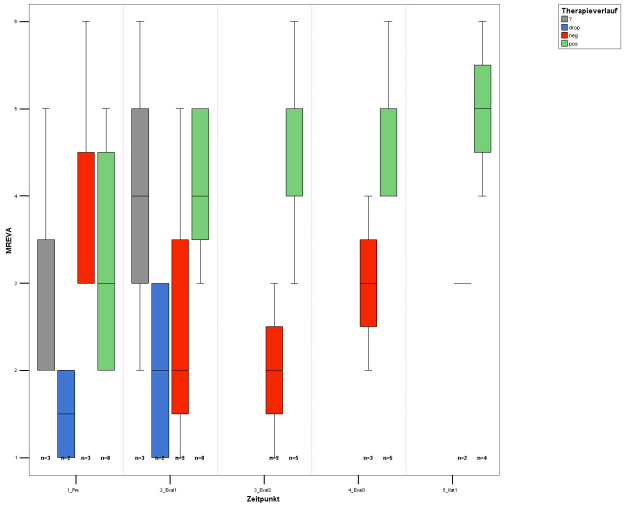

Figure 3. Sample of EDA: Boxplots (Printout Material)

feature to visualize information recorded at different points in time. The development in time is a very important aspect of the analysis of the progress of the therapy. In addition to the spring based visualization GRAVI also offers other methods, e.g. Star Glyphs to communicate the exact values of each answer (see Figure 2).

GRAVI provides various interaction possibilities to explore the data and generate new insights. The icons and visual elements can be moved, deleted, highlighted and emphasized by the user. Each change leads to an instant update of the visualization. For more details on visualization options, user interactions, and implementation see [3].

4 Other Techniques (EDA, ML)

We investigated the InfoVis technique and other methods used so far like EDA (in this case boxplots, histograms, scatterplots, and statistical measures). Machine Learning algorithms was the other choice as it might be able to reveal structures in the complex data. The ML algorithms are: a C4.5 decision tree and a Support Vector Machine (SVM) trained by Sequential Minimal Optimization (SMO).

Exploratory Data Analysis (EDA) was developed by Tukey [18] and is based on statistics. It helps users to review and analyze data on a descriptive level. Tukey thought that the emphasis on statistical testing might be too narrow an approach. He, therefore, suggested EDA as a possibility to formulate hypotheses and assess assumptions. Subjects in our tests were given printouts of these methods (e.g., Fig. 3).

Machine Learning is an area of AI concerned with the development of algorithms that enable computers to 'learn'. A Machine Learning method learns from observed examples or data. In general, there are two types of machine learning algorithms: supervised and unsupervised. In case of supervised learning, a priori knowledge about the data is used and in case of unsupervised learning, no prior information is given regarding the data or the output. We utilized two supervised schemes using WEKA [19]: a Support Vector Machine with Sequential Minimal Optimization algorithm [9] and a pruned C4.5 decision tree [10]. The output of these two methods were available to the subjects as handouts on paper (e.g., Fig. 4).

Such algorithms allow users to detect significant patterns in complex data sets. Decision trees, e.g., can be used as a decision support tool to identify possible outcomes of various strategies.

5 General Study Design

As mentioned above, usability issues in a narrow sense were not addressed in the study described here. The aim of the study was rather to find out whether meaningful insights can be found by using various analytical methodologies (GRAVI, EDA, ML), with the emphasis being on the analysis of GRAVI, and whether these methodologies can be used meaningfully for solving the psychotherapists' problems. For the importance of different evaluation stages and appropriate evaluation methods see [12].

It is well known that real users should be used for the evaluation of InfoVis techniques (see e.g. [8]). Nevertheless, there are situations when this is not possible. We cooperate with two psychotherapists with marked time constraints. Extensive testing is, therefore, not possible with our project partners. So, we decided to use computer science students as subjects. The sample size was 32. The

J48 t=all

```
Instances:    80
Attributes:   13 (BMI, ASW, BDI, SPS, SD, Restraint, MREVA, MRSOC, YSR:ZahlFr,
YSR:TreffenFr, CBCL:ZahlFr, CBCL:TreffenFr, Therapieerfolg)

J48 pruned tree
------------------

ASW <= 3
|   CBCL:TreffenFr <= 1
|   |   BMI <= 3: neg (9.66/2.34)
|   |   BMI > 3: pos (3.69/1.37)
|   CBCL:TreffenFr > 1
|   |   ASW <= 2: neg (6.77/3.24)
|   |   ASW > 2: pos (11.03/5.62)
ASW > 3: pos (33.85/3.65)

Number of Leaves  :       5
Size of the tree :  9

=== Summary ===

Correctly Classified Instances          52               80      %
Incorrectly Classified Instances        13               20      %
Total Number of Instances               65
Ignored Class Unknown Instances         15

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.975     0.4        0.796     0.975     0.876      0.887    pos
  0.867     0.06       0.813     0.867     0.839      0.918    neg
  0         0          0         0         0          0.694    drop

=== Confusion Matrix ===

  a  b  c   <-- classified as
 39  1  0 |  a = pos
  2 13  0 |  b = neg
  8  2  0 |  c = drop
```

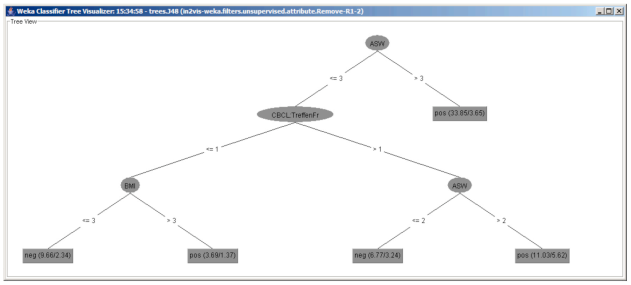

Figure 4. Sample of ML: Printout Material of C4.5 Decision Tree

students got a one-hour introduction into the subject area and another similar one into the methods used. The actual testing took place in a laboratory at our university and lasted 3 hours (approx. one hour for each method). The tasks the subjects had to solve were exploratory in nature. The psychotherapists are especially interested in the variables influencing success or failure of the therapy (= predictors). Such factors might be e.g. social phobia, depression or lack of social contacts (derived from the questionnaires). The tasks were formulated in the form of scenarios: two specifying meaningful subsets of data to explore (questions, patients, time steps) and two more stating concrete questions additionally. The intended types of investigations can be characterized as follows: (1) realize the change over time of 16 patients in 5 dimensions and identify positive and negative predictors (e.g. all patients whose depression scores do not decrease in the middle of the therapy are likely to have a negative therapy outcome), (2) recognize the consistent/inconsistent answers of parents and patients in the first time step and their role as predictor, (3) analyze the effect of the therapy on one specific dimension over time, and (4) predict a positive or negative therapy outcome of a so far unclassified patient with the available data of the first two time steps.

The major goals of this study is to answer the following questions: What are the types of insights gained with the different tools? Can the tools be used together to maximize the comprehension of the data? What are the specific strengths and weaknesses of the three methods?

6 Report System

Qualitative information about exploratory processes of InfoVis techniques is often gained by using thinking aloud techniques. This is very time-consuming and allows only a restricted sample size. We, therefore, developed a report system allowing subjects to document their insights. The subjects used this system to document their findings during the exploration process. It is accessible with any conventional web browser and is implemented in Perl and MySQL. The following data is collected: used material, description of finding, and confidence rating.

6.1 Used Material

For the three different methods used by the subjects three different report forms were developed. In the case of EDA subjects can specify via checkboxes whether scatterplots, boxplots, histograms, or statistical measures were used. For ML the options are SMO or J48.

If GRAVI is used, screenshots are possible to document the state of the visualization at the time an insight occurs. Of course, an exploration process sometimes last several minutes before a meaningful documentation is possible. Also, there often is not the one and only screenshot if the documentation deals with changes over time. In all these cases the subjects were told to upload a representative screenshot.

6.2 Documented Findings

An input field is provided for documentation of findings in natural language. There is no limitation in length. The subjects were requested to fill out a new report for every insight they may have. Here is an example of such an insight (taken from study; translated into English):

> "Patients with a negative therapy outcome have little self-confidence, which even continues to worsen [over time in therapy] and which is accompanied generally with an increasing depression."

6.3 Confidence Rating

With each documentation there has to be a confidence rating on a three-step scale: low, medium, and high. Without this rating the report cannot be saved.

7 Classification System

To compare the three methods it is necessary to define several relevant variables. These variables form a classification system. Some of the variables are more generic (complexity, plausibility, argument) and similar to evaluation characteristics suggested by North [5]. Others are more task specific as for example the identification of a predictor, an activity which was defined by the therapists as essential for the analysis. The classification was performed by three persons to ensure the reliability of the results.

7.1 Assigned Insights

We decided to split long reports into several in basic insights instead of dealing with them as unique occurrences of one complex insights. Long Reports are sometimes simply a cumulative documentation from a subject who did not adhere to the test procedure of reporting insights immediately. By doing so we can ensure comparability.

The classification system was developed bottom up but with the relevant literature in mind. Of course, we designed the scenarios which defined meaningful subsets of the data in close cooperation with our domain experts and also asked them for an extensive list of possible insights. But it is impossible to anticipate all valid documentations. So we compiled a preliminary list of insight categories after a first review of a few tens of reports. This list was then adjusted repeatedly to cover the documented findings. Every extension or consolidation was discussed thoroughly among researchers. This procedure resulted in thirteen main insight categories (see Table 1).

In terms of a hierarchical naming scheme every insight was uniquely identified. For instance, to the main prefix "data" we added more detailed specifications as needed in the process of insight classification. This resulted in 10 sub categories of "data" insights (e.g., reading off scores for individual patients, comparing identical scores, reading off scores for patient groups). A carefully designed hierarchy allows for summarizing for instance all data observations where the cognitive performance is the comparison of groups of patients.

7.2 Complexity

The complexity of each insight was rated on a three-step scale: low, medium, and high. This rating was influenced by the following factors: the domain value (e.g. the identification of a predictor is much more interesting than the recognition of missing data of a patient in various time steps); whether the observation deals with only one time step or with the change over time; the number of patients the insight deals with. And of course, we tried to ensure the relation within the different complexity ratings.

Main Insight Category	Meaning
abstract	General Insights on a Very Abstract Level
cluster	Identification of Visual Cluster (GRAVI)
data	Reading Off/Comparing Scores
error	Obvious Incomprehension of Used Method
eval	Reading Off Evaluation Scores (ML)
class	Classification of Patient(s)
coeff	Reading Off/Comparing Coefficients (ML)
meta	Remarks on Used Method
missing data	Recognition of Missing Data
no PRED	No Predictor
outlier	Outlier Identification
PRED	Predictor Identification
pseudo	Unclassifiable Insight

Table 1. Thirteen Main Categories of Assigned Insights.

7.3 Plausibility

It is sometimes a difficult task to rate the correctness of an insight assigned to a report in a range of true and false. Instead it is often more a question of plausibility. Especially, by using three notably different methods to explore the data, the subjects sometimes even documented contradictory insights all of which seem plausible with the used method. This classification takes place on a three-step scale: not plausible, moderately plausible, very plausible.

7.4 Argument

This category reflects the depth of the description of insights. Here we take into account the fact that the different methods may facilitate or hinder elaboration of findings. Once more, a three-step scale is used: absurd argument, no argument, meaningful argument.

7.5 Auxiliary Variables

We used different checkboxes to keep track of the classification status of every report which had to be proofread

by at least a 'second set of eyes'. A third pass was required for final classified status. There exist also various to-discuss flags (e.g. between investigators, with domain experts).

8 Results and Discussion

32 subjects documented an overall of 876 reports in sessions of 155 minutes. The number of reports and assigned insights for the three methods is shown in Table 2. An average of 2.47 insights were assigned to a report.

	Reports	Insights
EDA	375	846
GRAVI	235	711
ML	266	609
	876	2166

Table 2. Number of Documented Reports Using Three Different Methods and Assigned Insights.

Pearson's χ^2 test shows significant difference in the number of reports in column one of Table 2 ($\chi^2 = 37.034$, $df = 2$, $p = 9.08e{-}09$). Also the number of insights in column two of Table 2 differs significantly ($\chi^2 = 39.149$, $df = 2$, $p = 3.153e{-}09$).

Once again, by analyzing the whole of Table 2, we find significant divergency regarding reports and insights. It happens that there are fewer reports generated while using GRAVI and at the same time more insights gained with GRAVI than expected ($\chi^2 = 10.5003$, $df = 2$, $p = 0.005247$).

Main Insight Categories	E	G	M	
abstract	14	18	25	57
cluster	0	28	0	28
data	290	171	27	488
error	20	11	42	73
eval	0	0	116	116
class	29	19	23	71
coeff	0	0	72	72
meta	11	0	35	46
missing data	3	19	0	22
no PRED	57	37	57	151
outlier	17	12	0	29
PRED	395	390	210	995
pseudo	10	6	2	18
	846	711	609	2166

Table 3. Number of Insights of 13 Main Categories of EDA, GRAVI, and ML

Obviously, the numbers of insights of the thirteen main categories in Table 3 differ significantly for the three used methods ($\chi^2 = 857.7601$, $df = 24$, $p < 2.2e{-}16$).

It is obvious that both decision tree J48/C4.5 and SVM/SMO formula of ML cannot communicate individual scores of patients in most cases ('data') (see Tab. 3). Finding individual data also seems to be slightly difficult with GRAVI. Clusters of patients could only be found by GRAVI, although this should be, in principle, also possible with EDA. Most errors were made using ML (this method was found to be the most difficult by subjects). 'Eval' and 'coeff' were categories specific for ML. 'Meta' describes the formulation of hypotheses going beyond the data given. We have no explanation for the fact why this category does not appear with GRAVI. A very interesting result is that subjects found significantly less predictors with ML than with EDA and GRAVI. This probably also has something to do with the fact that ML seems to be the most difficult methodology.

Reports	E	G	M	
Scenario A	186	109	111	406
Scenario B	128	98	110	336
Question 1	36	15	23	74
Question 2	25	13	22	60
	375	235	266	876

Table 4. Number of Reports of EDA, GRAVI, and ML by Tasks

Returning to numbers of reports, Table 4 shows how many documentations were made for the four different tasks. Surprisingly, we cannot discard the null hypothesis ($\chi^2 = 8.3186$, $df = 6$, $p = 0.2157$). There is no systematic relationship between scenarios and methods used.

Insights	E	G	M	
Scenario A	443	354	236	1033
Scenario B	279	289	273	841
Question 1	71	38	39	148
Question 2	53	30	61	144
	846	711	609	2166

Table 5. Number of Insights of EDA, GRAVI, and ML by Tasks

But if we analyze the number of assigned insights (see Table 5) instead of the number of reports, we face once more significant difference ($\chi^2 = 50.8084$, $df = 6$, $p = 3.236e{-}09$). The underrepresentation of insights for scenario A (first task) with the use of ML and at the same time an overrepresentation of insights for scenario B / question

2 (second and fourth task) may be due to the need for an extensive familiarization phase with this method. The same precaution is advisable in interpreting possible reasons for underrepresentation of insights gained with the use of EDA in scenario B, because scatterplots should have provided good material for exploring the given problem area.

Reports	E	G	M	
Confidence +	185	128	111	424
Confidence $\sim$	143	90	107	340
Confidence –	47	17	48	112
	375	235	266	876

Table 6. Number of Reports of EDA, GRAVI, and ML by Confidence Rating

The subject had to state with every report what their confidence was in the documented finding (see Table 6). Unmistakably the reason for significant difference ($\chi^2 = 15.9368,\ df = 4,\ p = 0.003105$) between the three methods are many low confidence ratings with ML and few low confidence ratings with GRAVI. This fact has to be studied in more detail, because it might indicate the danger of being too sure of findings with GRAVI.

Insights	E	G	M	
Complexity C+	453	427	270	1150
Complexity C$\sim$	130	130	48	308
Complexity C–	263	154	291	708
	846	711	609	2166

Table 7. Number of Insights of EDA, GRAVI, and ML by Complexity

Table 7 shows the numbers of assigned insights and their complexity classification. We get a significant difference of the three methods ($\chi^2 = 111.1428$, $df = 4$, $p < 2.2e{-}16$): ML lacks high and medium complex insights and shows overly low complex insights. The opposite holds true for GRAVI.

Insights	E	G	M	
Plausibility P+	671	561	459	1691
Plausibility P$\sim$	97	67	69	233
Plausibility P–	78	83	81	242
	846	711	609	2166

Table 8. Number of Insights of EDA, GRAVI, and ML by Plausibility

As mentioned above the plausibility plays an important role. The numbers of insights and their plausibility classifications in Table 8 show no significant difference ($\chi^2 = 8.0717,\ df = 4,\ p = 0.08899$). Furthermore, all of the three methods have their highest ratios in the 'very plausible' category.

Insights	E	G	M	
Argument A+	244	195	137	576
Argument A$\sim$	585	474	433	1492
Argument A–	17	42	39	98
	846	711	609	2166

Table 9. Number of Insights of EDA, GRAVI, and ML by Argument

The last of the primary classification levels is whether an assigned insight has been elaborated in more detail (see Table 9). The significant difference ($\chi^2 = 26.175,\ df = 4,\ p = 2.917e{-}05$) is primarily caused by fewer wrong arguments with EDA and more wrong arguments with ML than expected.

9 Outlook and Future Work

As outlined above, the presented work is part of a extensive evaluation of GRAVI on multiple levels. On the level of InfoVis technique evaluation there are further questions to be investigated: Is there a relevance of the order in which the three groups of subjects used the three methods (i.e. MEG, EGM, GME)? Will a cluster analysis of the subjects reveal different interaction strategies with Gravi++ and strengths and weaknesses in having particular types of insights? Is there a significant amount of unbalanced insights like those with high confidence ratings but low plausibility and vice versa?

On a methodological level interesting questions arise, for instance is there a correlation between length of documentations, number of assigned insights to this report, and argument classifications?

Log Files: All sessions with GRAVI of the 32 subjects (65 minutes each) have been recorded and resulted in log files with more than 50000 entries. A simple parser to be written in Perl will do the job enumerating the different interactions, etc. Log file chunks between later insights will probably not reflect much of the exploratory interactions leading to these insights, because an existing learning curve would not be accounted for. But the analysis of the log files on a subject-level will probably help to identify different interaction strategies.

Case Study: After participatory design, development, and extensive usage of GRAVI qualitative interviews with the psychotherapists will reveal to what extent the visualization has proven itself useful for their daily work.

Conclusion

The comparison of the three analytical methods (GRAVI, EDA, ML) indicates that these methods have different strength and weaknesses. From the results reported in this study one might conclude that ML is not a recommendable methodology. The subjects' confidence ratings were low, the complexity of the gained insights was low and few predictors were found. Nevertheless, we know from a focus group study described elsewhere [13] that subjects thought ML to be a trustworthy and interesting method for experts. GRAVI and EDA seem to complement each other very well. EDA methods are well know and easy to understand. Therefore, they support exploration well (they generated most reports and insights). They are especially suited to analyze single values. GRAVI, on the other hand, works well for the generation of complex insights. Both GRAVI and EDA enabled subjects to find many predictors.

All in all a combined usage of the three methods is indicated because of their different strengths and weaknesses. Any obvious (or maybe superficial) benefits and limitations of the different methods should not lead to an exclusive use of one technique. Used in conjunction all three methods will very likely contribute to a deeper comprehension of the data to explore.

Acknowledgments

The project "Interactive Information Visualization: Exploring and Supporting Human Reasoning Processes" is financed by the Vienna Science and Technology Fund (WWTF) [Grant WWTF CI038]. Thanks to Bernhard Meyer for the collaboration in the classification process.

References

[1] C. Chen. Empirical evaluation of information visualizations: an introduction. *Int. J. Human-Computer Studies*, 53(5):631–635, 2000.

[2] C. Chen. *Information Visualization. Beyond the Horizon.* Springer, London, Berlin, Heidelberg, 2004.

[3] K. Hinum, S. Miksch, W. Aigner, S. Ohmann, C. Popow, M. Pohl, and M. Rester. Gravi++: Interactive information visualization to explore highly structured temporal data. *Journal of Universal Comp. Science*, 11(11):1792–1805, 2005.

[4] C. North. Snap-together visualizations: can users construct and operate coordinated visualizations? *Int. J. Human-Computer Studies*, 53(5):715–739, 2000.

[5] C. North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, 2006.

[6] K. Olsen, R. Korfhage, M. Spring, K. Sochats, and J. Williams. Visualization of a document collection: the vibe system. *Information Processing and Management*, 29:69–81, 1992.

[7] R. Pillat, E. Valiati, and C. Freitas. Experimental study on evaluation of multidimensional information visualization techniques. In *CLIHC '05: Proceedings of the 2005 Latin American conference on Human-computer interaction*, pages 20–30, New York, NY, USA, 2005. ACM Press.

[8] C. Plaisant. The challenge of information visualization evaluation. In *Proc. of Advanced Visual Interfaces AVI'04*, pages 109–116. ACM Press, 2004.

[9] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schoelkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 12, pages 185–210. MIT Press, 1998.

[10] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA., 1993.

[11] M. Rester, M. Pohl, K. Hinum, S. Miksch, S. Ohmann, C. Popow, and S. Banovic. Assessing the usability of an interactive information visualization method as the first step of a sustainable evaluation. In *Proc. Empowering Software Quality*, pages 31–44. Austrian Computer Society, 2005.

[12] M. Rester, M. Pohl, K. Hinum, S. Miksch, C. Popow, S. Ohmann, and S. Banovic. Methods for the evaluation of an interactive infovis tool supporting exploratory reasoning processes. In *Proc. BELIV '06*, pages 32–37. ACM Press, 2006.

[13] M. Rester, M. Pohl, S. Wiltner, K. Hinum, S. Miksch, C. Popow, and S. Ohmann. Mixing evaluation methods for assessing the utility of an interactive infovis technique. In *Proc. 12th Intl. Conf. on HCI*, LNCS. Springer, 2007. Forthcoming.

[14] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 11(4):443–456, 2005.

[15] R. Spence. *Information Visualization*. ACM Press, 2001.

[16] J. Stasko. Evaluating information visualizations: Issues and opportunities (Position Statement). In *Proc. BELIV'06*, pages 5–7. ACM Press, 2006.

[17] M. Tory and T. Möller. Human factors in visualization research. *Visualization and Computer Graphics, IEEE Transactions*, 10(1):72–84, 2004.

[18] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass., 1998.

[19] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA., 2nd edition, 2005.

[20] J. S. Yi, R. Melton, J. Stasko, and J. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.