EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2015)
W. Aigner, P. Rosenthal, and C. Scheidegger (Editors)

Should we Dream the Impossible Dream of Reproducibility in Visual Analytics Evaluation?

Michael Smuc[1], Günther Schreder[1], Eva Mayr[1] and Florian Windhager[1]

[1]Department for Knowledge & Communication Management, Danube University Krems, Austria

Abstract

Especially in the field of Visual Analytics, where a lot of design decisions have to be taken, researchers strive for reproducible results. We present two different evaluation approaches aiming for more general design knowledge: the isolation of features and the abstraction of results. Both approaches have potentials, but also problems with respect to generating reproducible results. We discuss whether reproducibility is possible or even the right aim in the evaluation of Visual Analytics methods.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—Standards

1. Introduction

In the Visual Analytics (VA) domain, the development of a new VA technique or method is usually characterized by an extensive amount of design decisions which have to be taken. From the authors' experience, such developments often lead to very complex but also very specific prototypes or products which are designed to be used by a highly specialized target group, which is more or less known. To make things even more complicated, VA methods often make heavy use of interaction methods and also the analytics part of VA adds up to the complexity of developments. This not only holds true for the design space, but also for the evaluation space.

When evaluating VA techniques, a huge amount of empirical methods has been developed in the recent decade (see [BPS07] [LIIS14]). During an evaluation study, many strategic and operational decisions have to be taken. In this paper we do not focus on the question how to document and share all the decisions taken during the development and evaluation to make it replicable. Instead we discuss two different approaches how to deal with this multitude of decisions on a theoretical basis: the isolated evaluation of single features and the abstraction of evaluation results.

2. The feature isolation approach

The main aim of the feature isolation approach, often seen in experimental settings or laboratory experiments similar to graph comprehension research [CS98] [KOS89], is to isolate specific features (e.g. the coloring, the legend, the basic layout...) of a VA technique and test this single feature in a highly standardized setting. In other words, in this approach the visualization or material gets abstracted or simplified as far as possible, by getting rid of many peculiarities and complexities which a visualization "in the real world" might consist of. Such a standardisation requires careful preparation: Not only the idea behind a specific feature has to be worked out very clearly to prepare the material for testing and comparison, also the intended effect of the feature has to be predefined to measure it appropriately. But as soon as you define everything and know the theories, you should be able measure the direct effect of a feature, for example by A/B comparison, measuring task completion times or by counting errors. In a perfect world this approach would lead to highly generalizable results that could be replicated exactly or reproduced in different contexts.

But the applicability of this approach has been criticized by the VA evaluation community for various reasons: First of all, as soon as you isolate a feature and test it, you cannot predict how this single feature will influence the whole VA technique. How different features in combination will work could be something different than the effects of a feature under isolated observation. Further, visual material is often developed for specific data, which makes this material hard to standardize and no widely acknowledged standards for material preparation exist until now. Finally, there are too many

DOI: 10.2312/eurorv3.20151148

possible usage scenarios for visual material to allow for an undisputed standardization. When we also consider interaction methods to be an essential part of VA, these components are sometimes very hard to isolate. To sum up, the isolation of features can easily lead to artificial situations, since the environment and most of the specifics are cut out.

But how else should we evaluate such (combinations of) techniques? Are VA tools in general too complex to be evaluated?

3. The abstraction method approach

Another approach is based on the abstraction of the results of an evaluation: Taking the results to a higher level and reflecting on their applicability in different situations.

Let us take a look at one of the methods with the most unreproducible outcome, the *think-aloud method*, where we hardly ever find exactly the same comments in the same order. One prominent way to bring some structure into its outcome in the field of VA is to extract *insights*, that is *"an individual observation about the data by the participant, a unit of discovery"* [SND05]. A first step to abstract these individual observations could be their categorization, e.g. to identify different kinds of insights into the data and into the functionality of the tool which can be counted and compared across multiple participants [SML*09]. In a next step, we proposed to relate these insights to one another to understand how multiple participants use visualization features and their prior knowledge in combination to make sense of the data ("Relational Insight Organizer", RIO, [SML*09]).

If the think-aloud method is applied during task completion, further analyses of the applied problem solving processes can help to understand, which processing strategies are activated by a VA method and where they fail [MSR10]. If we even go one step further, we can try to understand, which simple skills and more complex rules are activated by a VA method and where users have to move to more demanding cognitive *levels of processing* to develop new skills and rules [Smu14].

Higher levels of granularity may help to derive more general insights that could be transferred to other contexts or domains. But this approach does not come without costs: The abstraction could be biased - since every abstractions leaves room for interpretation. Furthermore this abstraction process could lead to results that are not truly valid for other settings, since essential aspects could be lost during abstraction. Consequently, data and reflections have to be published to ensure credibility and transparency of results.

4. Discussion

In this paper we presented two possible evaluation approaches which are often applied to generate more general, reproducible results. As outlined above, both strategies, the isolation of features and the abstraction of results, show specific advantages, but also major problems. Do we dream an impossible dream, when we strive for reproducible results? Or is it even the wrong target? Especially in the context of applied science, it might not always be necessary to create reproducible results, but rather develop good solutions for a well-defined application scenario. This is achieved by qualitative research in the best possible way, as it generates rich data and suggestions for improvements. But the aim of qualitative research is ***not*** to generate replicable results. The results should be credible and understood by other researchers, but not necessarily have to be replicated.

So we have to ask ourselves what do we mean by reproducibility? That one can replicate exactly the same study? That one can reproduce the same results of a specific VA technique or design idea, maybe with another tool or in another application context? Or that one can transfer the results to another topic? For sure, we wish for all of that, but it might be difficult to reach.

In our research we steer a middle course: For applied research the major goal has to be the achievement of results how to (effectively) improve a VA method, which is reliably reached with qualitative evaluation methods. To make the design decisions usable for others, we clearly describe our design decisions and evaluation results and make our methods transparent [SFW*14]. To gain more generalizable results, we move individual insights to a higher level of abstraction and take interrelations into account.

But when reflecting on the possibility of reaching further levels of methodological soundness, we came to ask ourselves whether it is just an illusion to get replicable results for a specific visualization. Against this background we want to contend that plain replicability is the wrong target for the abstraction method approach and should not be confused with the endeavor to achieve reproducible results. It is the development of more standardized methods and evaluation procedures, which could lay the groundwork for subsequent meta studies, and thus lead to more general, reproducible results. And it is the development of theories based on the higher level of abstraction and interrelations that lead to more general, reproducible results. We can still productively and confidently enhance our theories and methods beyond time and errors to effectively approach more coherent and comparable collections of results in the long run. Finally, we should not forget that in qualitative research reproducibility is not the holy grail - transparency and credibility are at least as important.

References

[BPS07] BERTINI E., PLAISANT C., SANTUCCI G.: Beliv'06: Beyond time and errors; novel evaluation methods for information visualization. *interactions 14*, 3 (May 2007), 59–60. URL: http://doi.acm.org/10.1145/1242421.1242460, doi:10.1145/1242421.1242460. 1

[CS98] CARPENTER P. A., SHAH P.: A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied 4*, 2 (1998), 75–100. 1

[KOS89] KOSSLYN S.: The Psychology of Visual Displays. *Investigative Radiology 24*, 5 (1989), 417. 1

[LIIS14] LAM H., ISENBERG P., ISENBERG T., SEDLMAIR M. (Eds.):. *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV 2014, Paris, France, November 10, 2014* (2014), ACM. URL: http://dl.acm.org/citation.cfm?id=2669557. 1

[MSR10] MAYR E., SMUC M., RISKU H.: Many roads lead to rome: Mapping users' problem solving strategies. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization* (New York, NY, USA, 2010), BELIV '10, ACM, pp. 8–15. URL: http://doi.acm.org/10.1145/2110192.2110194, doi:10.1145/2110192.2110194. 2

[SFW*14] SMUC M., FEDERICO P., WINDHAGER F., AIGNER W., ZENK L., MIKSCH S.: How do you connect moving dots? insights from user studies on dynamic network visualizations. In *Handbook of Human Centric Visualization*, Huang W., (Ed.). Springer New York, 2014, pp. 623–650. URL: http://dx.doi.org/10.1007/978-1-4614-7485-2_25, doi:10.1007/978-1-4614-7485-2_25. 2

[SML*09] SMUC M., MAYR E., LAMMARSCH T., AIGNER W., MIKSCH S., GÄRTNER J.: To score or not to score? tripling insights for participatory design. *IEEE Comput. Graph. Appl. 29*, 3 (May 2009), 29–38. URL: http://dx.doi.org/10.1109/MCG.2009.53, doi:10.1109/MCG.2009.53. 2

[Smu14] SMUC M.: Just the other side of the coin?: From error- to insight-analysis. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (New York, NY, USA, 2014), BELIV '14, ACM, pp. 35–40. URL: http://doi.acm.org/10.1145/2669557.2669570, doi:10.1145/2669557.2669570. 2

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *Visualization and Computer Graphics, IEEE Transactions on 11*, 4 (July 2005), 443–456. doi:10.1109/TVCG.2005.53. 2