Requirements for Reproducibility of Research in Situational and Spatio-Temporal Visualization

Position Paper

André Calero Valdez*
Human-Computer Interaction Center
RWTH Aachen University

Anne Kathrin Schaar†
Human-Computer Interaction Center
RWTH Aachen University

Julian Romeo Hildebrandt‡
Human-Computer Interaction Center
RWTH Aachen University

Martina Ziefle §
Human-Computer Interaction Center
RWTH Aachen University

Abstract

Research on spatio-temporal visualization is driven by the development of novel visualization and data aggregation techniques. Yet, only little research is conducted on the systematic evaluation of such visualizations. Evaluation of such technology is often conducted in real-life settings and thus lacks fundamental requirements for laboratory-based replication. Replication requires other researchers to independently conduct their own experiments to verify your results. In this position paper, we discuss the requirements for replication studies of spatio-temporal visualization systems. These requirements are often impossible to achieve for highly contextual visualizations such as spatio-temporal visualizations. We argue that reproducibility—allowing other researchers to validate your findings from your data—is a better aim for highly contextual visualizations. We provide a sample workflow to ensure reproducibility for spatio-temporal visualization and discuss its implications.

Index Terms: Human-centered computing—Visualization—Empirical Studies in Visualization; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

Geo-spatial and spatio-temporal visualizations visualize data that is related to locations on maps. Such visualizations play a large role in supporting decision making in a wide range of applications. Typically, these visualization are integrated into dashboards and merged with various other information. The range of applications goes from business intelligence, Internet of Things [13], to agriculture, disaster support and crime reduction. Visualizations, especially spatial ones, are often used as they allow to quickly draw inferences from data in relatively high quality [32]. Users are familiar with mapping concepts and have an intuitive understanding of layered visualizations on such maps.

However, sometimes the visualization does not do the data justice. Often highly complex and uncertain data is aggregated to form a highly simplified overview that then ultimately guides decision making. The widespread use of such systems potentially has drastic economic, societal, and governmental implications [11]. For instance, looking at the example of predictive policing [22], where police forces are assigned to areas of interest based on geo-spatial and temporal data, consequences of misinterpreted data could lead to increases in criminal activity, overlooked criminal offenses, and also unjustified stigmatization of innocent "predicted perpetrators". To reduce the risk of such errors thorough evaluation of such visualizations is critical.

In this position paper we look at the peculiarities of such spatio-temporal visualizations. In particular, we look at how to evaluate such visualizations with respect to replicability and reproducibility. Foremost, we argue that true *replication* in such complex settings is often not feasibly and thus *reproduction* is a target best aimed for. To strengthen this argument, we show where the complexity in spatio-temporal visualizations originates from (i.e., complex data and temporal dependencies, see Sect. 3), what requirements can be derived from these complexity drivers (see Sect. 4), and how to best address them using a formalized process (see Sect. 5).

The overarching questions is: How can we ensure that visualizations are understood correctly and chosen optimally when the conditions of using the visualization are highly context-dependent?

2 Reproducability vs. Replicability

The challenges of replication have been discussed in research for quite some time [20]. However, only after the replication of 100 studies from leading journals in the field of social psychology failed to find the same effects as the initial studies, a large amount of core findings from social psychology research have been questioned [7]. The replication crisis has plunged science into a crisis of credibility.

Has the pressure to "publish or perish" systematically biased researchers to produce a large quantity of subpar research? In quantitative research fields, criticism focuses on the one hand on incorrect usage of statistical methods, over-reliance on null-hypothesis significance testing, and p-values [8], as well as on the other hand on the publication-bias—only significant findings have a chance to get published. Negative findings are harder to get published, thus increasing the likelihood of false positive findings to get published.

To address these problems, the open-science foundation has formulated a set of procedures to improve reproducibility in research in general [7], such as pre-registration, publication of data, or publication of analysis code. But, how do these recommendations translate to visualization research?

In order to determine requirements for reproducibility in spatio-temporal visualizations, we first must differentiate what reproducibility means, especially in contrast to the related term replicability. Often these terms are used interchangeably, although the refer to different concepts.

While the term *replication* refers to the process of repeating an experiment or study with exactly the same conditions (e.g., same lab-equipment, same data, etc.), *reproduction* has weaker requirements [23]. The purpose of *replication* is to test whether the effects that were found in initial studies are stable across the population. The questions are: Are effects generalizeable? Do we have the same findings, if different people partake in our experiments? Here, the aim is that measurements and context are kept as similar as possible to the original study. Replication is thus a very strong retest of previous findings.

Reproduction in contrast refers to the process of testing whether

*e-mail: calero-valdez@comm.rwth-aachen.de
†e-mail: schaar@comm.rwth-aachen.de
‡e-mail: hildebrandt@comm.rwth-aachen.de
§e-mail: ziefle@comm.rwth-aachen.de

other researchers come to the same conclusions, given that they have access to the same data [23]. One might think, that this is a natural assumption, however different instrumentation or even software versions of evaluation software can lead to differences in results.

Often visualization itself is used as a tools for reproduction [21]. By visualizing resulting data other researchers can verify if they come to the same conclusions when presented with the same visualization.

3 Spatio-Temporal Visualizations and Challenges

Visual representations have been used to understand spatial data for a very long time. Maps, e.g., are selective representations of space that display aspects of interest, such as roads, ocean, hills, and cities to the viewer. With the rise of computerized mapping technology, data is used to generate maps. Algorithms transform the data to different map-projections and data source such as open-street map provide access to a global map of the world. However, even such projections warp reality by including some features—think roads, buildings, and national borders—and excluding other features—think flora, fauna, or animal territories. In many spatio-temporal visualizations maps are used as the backdrop for additional data visualization. The data at interest is not the map itself, but additional data layered over this map. The user then interprets this data. Besides the map being warped the data itself maybe warped, too.

To simplify understanding of the these aspects we focus on the example of situational visualization in a joint-operation center (JOC) for large-scale disaster recovery, e.g. after an maximum credible accident (MCA) in an atomic power plant. This example is may be fictitious, but many of the properties of this scenario can be mapped to similar other examples. The *users* are analysts coordinating the deployment of support personnel, supplies, and managing evacuation routes. The *task* is to identify changes in the situational context. Have new places for action occurred, e.g., changes in wind direction or newly damaged infrastructure. Additionally, it is necessary to understand the natural evacuation routes of the population in real-time to identify the possibility of congestion en-route.

In this example, the public interest in the success of these operations is very strong. People's lives are at stake and decisions derived from under designed visualizations would be catastrophic. Therefore, the need to evaluate visualizations in such contexts is critical, although the "real" use-case—the MCA—may never actually occur. Ideally, the evaluation studies of such systems would be fully replicable. Why this is hard, is explained in the following sections.

3.1 Data Sources for Spatio-Temporal Visualizations

There are several ways of attaining data for spatio-temporal visualizations. Each of them has different implications with respect to replicability or reproducibility. In particular, replicating experiments with such data is often not feasible as the complexity and the intricate connections of the data are hard to discover in a very similar fashion. Either data has to be generated artificially or data is drastically different between different experiments, making a replication attempt futile. The following subsections discuss different sources of data and why these sources make replication hard.

3.1.1 Using Data from Mobile Phone Apps

One way of attaining data for visualization is by gathering data from end-users or the "crowd" using a specific mobile phone app. Such scenarios are often used to get local information in disaster settings, where no remote sensing is available. Users may provide their own data, gathered, e.g., by mobile phones and data is integrated in a server and then a spatial-visualization is generated to distribute the information to all "crowd-members" or a disaster joint-operation center (JOC). Some implementations may even utilize Augmented Reality (AR)-to overlay the real-world with information [16].

This type of data often contains very specific data relating to the individual spatial context and the event that occurs. Users decide whether or not to share information correspondingly to their individual relatedness to event at hand. Such data is hard to artificially *replicate* at scale and specificity.

3.1.2 Using Indirect Data

Another approach is to automatically measure data from a crowd and to detect regularities and irregularities to inform the joint-operation center (JOC) about points of interest [14]. For example, visualizing cell phone tower logins can help understand where evacuating people are gathering. The JOC can now investigate whether irregularities are of interest or whether they are caused by external reasons irrelevant to the supervision process. Often so called hotspot-visualizations (spatial-heatmaps, see also Fig. 1) are used to inform users where events of interest are to be expected [18].

"Hot" areas are not always of top interest, but instead it could be more relevant to identify areas where something irregular happens. For this case anomaly detection patterns can be used to integrate temporal data in the visualization [5].

What makes such data hard for *replication* is that the spatial context is often very specific, e.g. the structure of the environment, typical routs for locals, etc. Users at the JOC will utilize their knowledge of the location to their advantage, making replications quite hard to achieve. Similar results, can only occur if, e.g. very specific landmarks also exist at the replication site. Further, only if the users have similar knowledge about the specifics of local landmarks, similar results should appear. This increases the complexity of studying such visualization as an additional dimension is added to the layers of data in the users mind, which is not recorded in the experimental data.

3.1.3 Using Data From Social Media

Another way of acquiring data for visualization is to scrape data from social media. While this type of content is not intended for this specific use-case, it may be utilized for it. Here, user-generated content is gathered and related to the geo-spatial markers embedded in the content. For example, by utilizing the self-disclosed location of twitter users and the content of tweets, locations of users can be inferred and utilized in geo-spatial visualization [17]. Additionally, geographic data can be obtained by analyzing data that volunteers have submitted, often by submitting photos that contain geo-tags [30]. By integrating photos and location-data a high-resolution image of locations can be obtained.

However, one must be aware that data from crowd or social media sources has the potential for high levels of uncertainty. The users may contribute incorrect location data accidentally or on purpose. Nevertheless, by integrating content from twitter posts, e.g., location data could be extracted relatively well [6]. Similar approaches have been implemented for other blogging services on the web as well [3]. By visualizing multi-panel data from different data sources (e.g., twitter posts, GPS-locations in photo exif-data[1], user locations, or even remote sensors), higher accuracy of data can be achieved [9].

For the purpose of *replication*, it is hard to artificially generate large amounts of such data that is realistic. This means to have locations data embedded in them, in part on purpose, in part accidentally. Further, it requires thinking about strategies of re-posting/sharing and to generate such data. This alone can be a full research project.

3.2 The Temporal Aspect of Data

In many cases it is not only interesting where data is located, but also where and when certain events happen. Temporal data requires different approaches for visualization. Temporal glyphs can be used to indicate when events or changes occur in relation the current

[1] Exif data is the meta data that is stored with image files that often contains geo-locations.

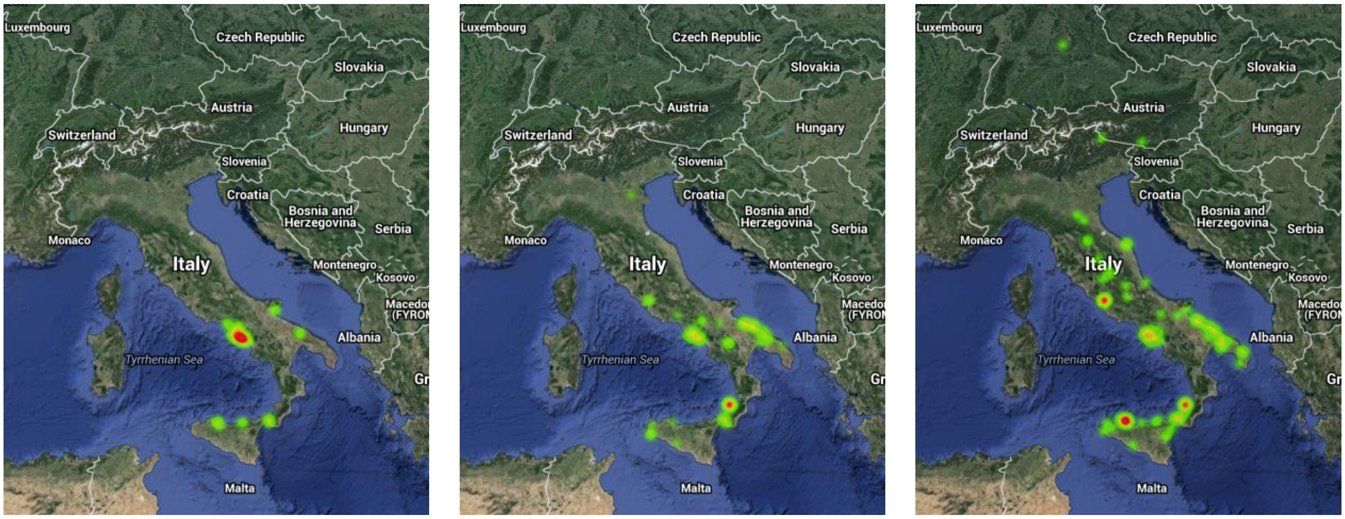

Figure 1: Heatmap spatio-temporal visualization that shows the frequency of social media posts related to a disaster event shown as an animation (frames from left to right). Taken from [16] used under CC-BY 4.0

situation [1]. Often animations are used to naturally map temporal data in space. Typical examples are weather forecasts or rain-maps, where users can see where clouds have moved, rain was expected, and whether rain will occur in their vicinity.

As soon as predictive models are used, spatio-temporal visualization should reflect uncertainty that is contained in the model [2]. Visualizations for uncertainty in temporal data have been studied [12], however, mostly not in combination with spatial data. Some approaches by Shrestha et al. [27] allow to visualize spatial and temporal uncertainty, however they lose their natural mapping for real-time interpretation. Such visualizations can be used in post-hoc analyses approaches. Only few types of animations lend themselves as fitting uncertainty visualizations. One approach is showing different versions of the future, while other approaches directly map the uncertainty to the spatial representation of the data. The latter approach is often used, when the uncertainty relates to the path of an event, but not necessarily to the overall likelihood of the event.

3.3 Influence of User Diversity

Besides data- or visualization-related factors, human factors and user-diversity could also have a considerable influence on data interpretation and therefore influence replicability and reproducibility. This is the case as subjects' interpretation of visual data is a complex and an iterative process involving sensory processing and perception of environmental stimuli, working & long term memory, cognition and attentional resources [31]. This entire process involves several Human Factors that impact objective performance and subjective user experience.

The problem for *replication* is that the specific individual characteristics of the analyst may have a strong influence on the evaluation of the visualization. Some characteristics may be highly specific to the space that is visualized (e.g., local route knowledge, familiarity with landmarks, etc.). Naturally, evaluation studies aim to record not only the independent and dependent variables, but many factors can have confounding influences unbeknown to researcher. Even the influence of these factors could be subconscious and therefore the user could be incognizant of their importance.

3.4 Summary

Given that the evaluation of spatio-temporal visualization should include these factors, how can evaluation processes be designed in a way that replication of results can be achieved? What is necessary for replication? What are the benefits of replication over reproduction? The next sections will explain, why reproducible research should be favored over reproduction in the case of complex spatio-temporal visualizations.

4 Barriers for Replication

As we have seen, the variety of spatio-temporal visualizations and their data sources is quite complex. At all levels of data transformation error and uncertainty could exist increasing the compounding error in the data to action pipeline. In the case of real-time spatio-temporal visualizations these effects can be dramatic. Still, such visualizations need to be reliably evaluated.

As mentioned before, we argue that for real-time spatio-temporal visualization *replication* is a task that is almost out of scope. Given that the evaluation of a spatio-temporal visualization requires the inclusion of end-users and data that somewhat resembles real-life conditions, circumstances such as these render replication of such studies nearly impossible.

Think of our example and the evaluation of a real-time disaster warning system. The potential users are specialists and are well aware of potential hazardous locations in their immediate neighborhood. One way of evaluating such a visualization would be to create fictional data or use historic data. Fictional data has the drawback of creating scenarios that are highly unlikely and thus in conflict with the domain expertise of the users. Historic data is easily recognized as such and leads to fast high quality judgments that might overestimate the effectiveness of a visualization. Testing the application in the real-world using real live data, could either lead to no relevant events occurring during evaluation or to highly relevant events that induce strong affects (i.e., emergency) in the participants. The results are thus either irrelevant for the purpose of the visualization or not replicable on purpose. The emotional state of "there is immediate danger" is hard to replicate in a laboratory setting.

The inherent locality of spatio-temporal visualization is also hard to replicate. Evaluating the same visualization somewhere else could lead to different results simply due to the differences in local terrain and landmarks. Asking remote experts to evaluate the same situation, could also lead to different outcomes, as judgment is also influenced by the often localized domain expertise.

Real-time spatio-temporal visualization depend on spatial and temporal context, which are both part of the users' domain expertise.

Forging or faking any of this context, will lead to different outcomes in evaluation. Yet, how can the evaluation of such visualization be attained?

We believe that the pursuit for *reproducibility* by following standardized research protocols is one way of ensuring the validity of evaluation studies on spatio-temporal visualizations.

4.1 Requirements for Reproducibility

But what is necessary for reproducibility in such studies? Some of the recommendations by the open science foundation are not always applicable to real-world research. Public pre-registration of trials could lead to differences in behavior of the "data". For example, if a study aims at visualizing GPS location data from images posted on twitter, going public with this aim, could cause users to remove GPS location data or post fake exif data in their photos, affecting the outcome of the study.

First, one must discern between exploratory research and confirmatory research. For exploratory studies, where no direct hypotheses are tested, we believe that following a standardized iterative research protocol as provided by the constraints in the design study-methodology [26] should yield reproducible research, if certain additional criteria are respected. This means e.g. studying visualization prototypes with real domain experts in an iterative approach, along with keeping track of data-quality criteria, implicit and tacit knowledge of experts—especially location knowledge—and precise descriptions of visualization tasks. While some of these seem trivial, especially trying to keep track of source of uncertainty in data and implicit analyst knowledge is crucial for confirmatory research. The question is, what additional information must be recorded? From our experience, the strongest concern for judging evaluation of visualizations stems from uncertainty in data, unclear tasks, and differences between evaluations caused by different test users.

4.1.1 Requirements for Data

As we have seen in Sect. 3.1.1 to Sect. 3.1.3 data quality may differ dramatically between different data sources, but also between different times for the same data source. GPS signals vary in location accuracy, coverage, and reliability [34] depending on weather conditions, local geographic conditions, and in the case of *assisted GPS* depending on available WIFI networks or cellular radio poles. When visualizing data such information is often lost in data storage. The level of uncertainty may be available to the user from their domain knowledge and knowledge of, e.g., the current weather conditions, however such information is not often recorded for reproducibility. By this we mean that the uncertainty must not necessarily be visualized in the visualization, as in uncertainty visualization. Often this can lead to overwhelming visualizations. But the data that was used for evaluating the visualization must be stored with the respective knowledge about uncertainty of data.

Therefore, as one requirement for reproducibility we believe that it is crucial to enrich raw data with information about the *uncertainty of data*, be it either additional information such as weather information or expert judgments recorded during the experiments. Uncertainty can refer to different aspects of uncertainty:

Spatial uncertainty refers to the uncertainty in the accuracy of location data, either from sensor uncertainty or from storage uncertainty. Social media posts, e.g., have very coarse location data, often located at points of interest nearby the actual photography. Storing uncertainty information along the data could be one approach to ensure reproducibility in this regard.

Temporal uncertainty refers to the uncertainty that is produced from imperfect or incorrect mapping of events to the timeline. A mobile phone could have incorrect time settings, possibly the home time of the owner during holidays, which gets stored in exif data. Data that is shared on social media could have been produced earlier than stored on the server. Ensuring that this uncertainty is recorded should increase reproducibility of experiments.

Lastly, *missing data* is an aspect often under-reported. Naturally, when a certain column in a table is missing, `NA` values are reported, but there are different types of missing data. The best case is, when we receive a missing data-signal (such as `NA`), in other cases certain values might refer to missing data, and in worse cases we just lack certain data. Here, someone who just sees the final data, might be uninformed where data might be missing. Adding meta-data for possible missing data—beyond the `NA`—is crucial for reproducibility.

4.1.2 Requirements for Visualization and Analysis Tasks

Besides ensuring that the quality of data is tracked across evaluation studies, one should also ensure that the tasks that a user conducts is tracked and stored. This encompasses two processes. The first process is how the visualization is generated. A simple description of the visualization can lead to inconsistencies in understanding what was visualized. For example, the ordering of location data classes (e.g., different event types) in a database could lead to different drawing orders, thus changing the resulting visualization when trying to replicate such a study. As a best-case scenario evaluation studies should either record the stimuli that were presented (as a video), or publish the code for the visualization. The latter would also include keeping track of library or package versions as these could have different effects on the visualization. Naturally, this is not always possible. Proprietary code or code that reveals sensitive information may not be shared in public repositories. In such cases it may be helpful for reproducibility to explain what can and what cannot be shared and allow access to those parts that can be shared. Another drawback of screen recordings is that often visualizations in JOCs cover several screens with very high-resolution screens, and sharing of such large files is still costly.

The second process is the *analysis task* the user has to complete in the evaluation of the visualization. Here it is relevant to note that judgments that occur during experiments can depend on perception, tasks, and social context of the experiment. It could be that different task orders cause differences in interpretation and therefore the selection and ordering of tasks must be tracked and recorded. If different tasks influence each other during an experiment this must reflected in the data. Randomization can be a way to minimize such effects, but when real-life data is used, randomization is not always possible. Keeping track of these *decisions* is helpful for reproducibility as it allows other researchers to understand how the evaluation data was generated.

Additionally, it is helpful to keep track of the visualization and analysis task in combination with detailed information about the participants in an evaluation study. Factors such as expertise, expectations, and preconceptions influence decision making in an analysis task and should therefore be recorded. The next section provides details on this aspect.

4.1.3 Impact of Human Factors and User-diversity

Due to the fact that little is known about the importance of individual factors and their interaction to date, we propose a systematic recording and evaluation of factors that have already been identified as relevant in HCI research. In addition to factors that can be derived from the experimental setting, we suggest to record general demographic information, such as age and gender, which are often identified as carrier variables for other characteristics [29], cognitive characteristics, personality traits, as well as job-related aspects.

From the field of the cognitive characteristics it can be assumed that the sense of orientation and spatial imagination are of importance. In addition to that, personality related factors like coping with uncertainty [15, 24] or the need for cognition [4] are hypothesized to have a bearing on the handling and evaluation of visualizations. The third area of relevant factors is related to job-specific aspects.

In this context the experience in dealing with visualizations, knowledge about local conditions, or other background information in the context of the information to be visualized, are relevant. For exmaple: How often do events in question occur? Do contexts change regularly for the users?

It is important to note that even if such factors were possibly irrelevant to hypotheses investigated, recording such data can be helpful when trying to reproduce the results and understand the context of the conducted study. One has to pick a trade-off between experimental economy and long-term reproducibility. Whatever this trade-off is, recording it is necessary.

5 Recommendations for the Analysis Workflow

One means of enhancing reproducibility is the sharing of data and analysis code (see Fig. 2). This naturally requires that data does not contain sensitive information which could reveal individuals identities. This is particularly hard, when high-resolution temporal GPS location data is used [25]. Such information is highly individual and identifying individual users can be achieved from small data sets. Ideally, researchers would follow the following steps:

1. Store raw data, with meta information about data uncertainty and missing data. Timestamp the data and share the raw data in a format that is stable across time (i.e., non binary, but in a format that supports interoperability such as CSV, TSV, JSON or similar).
2. Share the data cleaning code and pre-processing code.
3. Share the cleaned data with meta-information and time-stamps.
4. Share the analysis-code (e.g., using GitHuB).
5. Share the results in formats beyond archival publications, including interactive visualizations (e.g., using libraries such as plot.ly).

5.1 Anonymization

In the case of sensitive data a *step zero* should be included that conducts *anonymization* of data. Several methods for anonymization exists with different benefits and drawbacks which need to be evaluated. K-anonymity is a method where individual data cases are changed, e.g., by changing column data to a common level, that at any given time at least *k* individuals share a the same data and are thus indifferentiable. Using *k*-anonymity [28] on GPS-data could yield the outcome of a visualization useless, depending on the task. Further, if spatio-temporal data is stored, how do you select individuals to anonymize? It is not only locations that must be changed, but also timing data. Do you select nearest neighbors to equalize? Are these people socially connected (e.g. passengers in the same car)? Does this provide enough privacy? Some approaches use unlinking in mixing-zones [33], while other approaches use rule-based systems [35]. Still, data shared by k-anonymity has potential for privacy leaks when combined with other data.

To prevent such privacy leaks, more advanced methods, such as differential privacy [10], could be used. Such approaches would require setting up a database and sharing API-access to a database instead of raw-data [19]. As on example the chorus project can be used to create differentially private databases. Still, with every reproduction effort data would degrade, forcing researchers to manage reproduction claims. In many cases, where anonymization is impossible, sharing of raw data will be out of the question. For such cases, it is necessary to document why different approaches of data sharing are impossible, unethical, or possibly illegal.

5.2 Sharing Code and Data

Sharing of code and data should utilize open services such as the open-science framework[2], GitHub or similar local repositories. This allows data and code to be cited and reused and fosters the idea of science as a cumulative effort. Many of the open-data frameworks provide embargo periods, allowing researchers to submit data for publication but restricting access to selected peers until a certain amount of time has passed (e.g., a project has expired). For larger amounts of data, Harvard's Dataverse or Google's BigQuery Database could be utilized, with the respective ethical and legal considerations.

It is crucial to share code and data from between the different workflow steps as errors in every single step can yield the later results flawed. Sharing the code and results of e.g., data cleaning and data pre-processing can help find errors in data before the analysis. Ideally, the analysis should automatically update if such errors are found in earlier steps of the workflow.

This way of improving reproducibility of analyses can be achieved by utilizing tools such as Jupyter notebooks for Python or RMarkdown scripts for R that contain both the interpretation of data and the analysis code. The analysis contains the textual outcome for possible outcomes regarding hypotheses in confirmatory research and automatically generated text for exploratory research. Several packages exist for both Python and R that simplify reproducible research workflows [21].

To provide only a few examples for the language R, `packrat` allows fixing library versions in an analysis to ensure that later improvements in a library do not alter results, `vcr` allows storing http requests and responses to ensure the web behaves the same in later points in time. Several packages can be found at the rOpenSci repository[3] including a package for artificial landscape generation (NLMR).

6 Conclusion

In this position paper, we argue that full replication of studies of real-life spatio-temporal visualization applications is not suitable. We instead advocate for fostering reproducibility and the development of standards in evaluation studies for spatio-temporal visualization—in particular for real-time visualizations. This should be done regarding the assessment of data quality, the description of analysis tasks, and the operationalization of participant diversity. We further propose a sample workflow that would enable reproducibility in the evaluation from a data-analysis perspective by sharing data and analysis code in public repositories.

The workflow is a result of our shift from "one-time analysis of data in SPSS" to reproducible research using the "R" programming language. The workflow was discussed in our research project, with domain experts, visualization experts, and experts in statistical analyses. Additionally, we included feedback from experts in digital ethics and law in particular with regard to anonymization. We are currently applying this workflow in a large research project with governmental users and hope to share our experience using this workflow.

As a final note we want to stress that the efforts to enhance reproducibility should not come from a vantage point of placing blame on researchers who are unable to share data or code. Researchers should instead be empowered by the improvements of reproducible research workflows, which include the reuse of data analysis code in later projects and the improved availability of existing empirical data in meta-analytic approaches of science. We look forward to presenting these ideas in the BELIV workshop and hope to improve our ideas in the discussions with the other participants.

Acknowledgments

This research has been funded by the Federal Ministry of Education and Research of Germany in the framework of "Research for civil security 2012-2017" (Project number 13N14764).

[2] http://www.osf.io

[3] https://github.com/ropensci

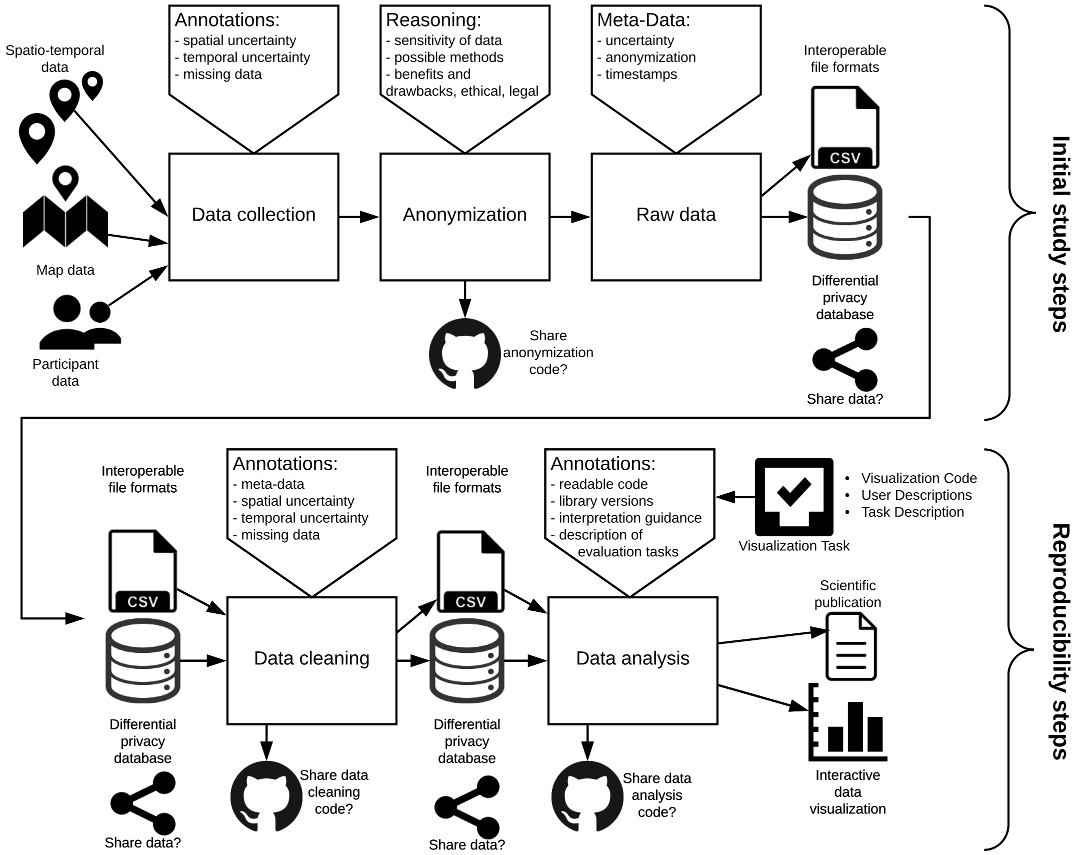

Figure 2: Sample workflow for reproducible research in spatio-temporal visualization evaluation

References

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE transactions on visualization and computer graphics*, 14(1):47–60, 2008.

[2] W. Aigner, S. Miksch, B. Thurnher, and S. Biffl. Planninglines: novel glyphs for representing temporal uncertainties and their evaluation. In *Information Visualisation, 2005. Proceedings. Ninth International Conference on*, pp. 457–463. IEEE, 2005.

[3] H. Bosch, D. Thom, M. Wörner, S. Koch, E. Püttmann, D. Jäckle, and T. Ertl. Scatterblogs: Geo-spatial document analysis. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 309–310. IEEE, 2011.

[4] J. T. Cacioppo and R. E. Petty. The need for cognition. *Journal of personality and social psychology*, 42(1):116, 1982.

[5] J. Chae, D. Thom, H. Bosch, Y. Jang, R. Maciejewski, D. S. Ebert, and T. Ertl. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pp. 143–152. IEEE, 2012.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768. ACM, 2010.

[7] O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[8] G. Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2012.

[9] M. Dörk, S. Carpendale, C. Collins, and C. Williamson. Visgets: Coordinated visualizations for web-based information exploration and discovery. *IEEE Transactions on Visualization and Computer Graphics*, 14(6), 2008.

[10] C. Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pp. 338–340. Springer, 2011.

[11] S. Elwood. Geographic information science: Emerging research on the societal implications of the geospatial web. *Progress in human geography*, 34(3):349–357, 2010.

[12] T. Gschwandtner, M. Bögl, P. Federico, and S. Miksch. Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):539–548, 2016.

[13] M.-S. Kim. Research issues and challenges related to geo-iot platform. *Spatial Information Research*, pp. 1–14, 2017.

[14] R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pp. 1–10. ACM, 2010.

[15] R. Lipshitz and O. Strauss. Coping with uncertainty: A naturalistic decision-making analysis. *Organizational behavior and human decision processes*, 69(2):149–163, 1997.

[16] G. Luchetti, A. Mancini, M. Sturari, E. Frontoni, and P. Zingaretti. Whistland: An augmented reality crowd-mapping system for civil protection and emergency management. *ISPRS International Journal of Geo-Information*, 6(2):41, 2017.

[17] A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 181–190. IEEE, 2011.

[18] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, M. Yakout, M. Ouzzani, W. S. Cleveland, S. J. Grannis, and D. S. Ebert. A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):205–220,

2010.

[19] J. Near. Differential privacy at scale: Uber and berkeley collaboration. In *Enigma 2018 (Enigma 2018)*. USENIX Association, Santa Clara, CA, 2018.

[20] H. Pashler and E.-J. Wagenmakers. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530, 2012.

[21] J. M. Perkel. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature*, 554(7690):133–134, 2018.

[22] W. L. Perry. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.

[23] H. E. Plesser. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76, 2018.

[24] O. Renn. *Risk governance: coping with uncertainty in a complex world*. Routledge, 2017.

[25] L. Rossi, J. Walker, and M. Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):11, 2015.

[26] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[27] A. Shrestha, Y. Zhu, and B. Miller. Visualizing uncertainty in spatio-temporal data. In *ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA)*, pp. 117–126, 2014.

[28] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[29] H. M. Trautner. *Lehrbuch der Entwicklungspsychologie: Theorien und Befunde*, vol. 2. Hogrefe Verlag, 1997.

[30] J. J. D. White and R. E. Roth. Twitterhitter: Geovisual analytics for harvesting insight from volunteered geographic information. In *In Proceedings of GIScience, 2010*, 2010.

[31] C. D. Wickens, J. G. Hollands, S. Banbury, and R. Parasuraman. *Engineering psychology & human performance*. Psychology Press, 2015.

[32] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):973–979, 2010.

[33] P. Zacharouli, A. Gkoulalas-Divanis, and V. S. Verykios. A k-anonymity model for spatio-temporal data. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pp. 555–564. IEEE, 2007.

[34] P. A. Zandbergen. Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning. *Transactions in GIS*, 13(s1):5–25, 2009.

[35] H. Zhang, C. Wu, Z. Chen, Z. Liu, and Y. Zhu. A novel on-line spatial-temporal k-anonymity method for location privacy protection from sequence rules-based inference attacks. *PloS one*, 12(8):e0182232, 2017.